Journal of Statistical Software

MMMMMM 2013, Volume VV, Issue II. *http://www.jstatsoft.org/*

SCGLR - An R package for Supervised Component Generalized Linear Regression.

Frédéric Mortier CIRAD
Catherine Trottier University Montpellier III and I3M
Guillaume Cornu CIRAD
Xavier Bry I3M

Abstract

The objective of this paper is to present an R package, **SCGLR**, implementing a new PLS regression approach in the multivariate generalized linear framework. The method allows the joint modeling of random variables from different exponential family distributions, searching for common PLS-type components. We discuss several of the functions in the package focusing in particular on the two main ones: `scglr` and `scglrCrossVal`. The former constructs the components and performs the parameter estimation, while the latter selects the approriate number of components by cross-validation. The package is illustrated on an appropriate ecological dataset through which we aim at predicting the abundance of multiple tree genera given a large number of geo-referenced environmental variables.

Keywords: Multivariate generalized linear model, partial least squares, Fisher Scoring, R.

1. Introduction

The classical generalized linear model (GLM), used for modeling random variables from exponential family distributions, suffer from different limitations: (i) it does not allow modeling of more than one outcome at a time; (ii) for want of regularization, it cannot deal with many correlated regressors - whatever relevant causal factors they may represent - and thus requires some preliminary selection of regressors; (iii) the degree of explanatory realism of the model and the robustness of the prediction may be highly influenced by this selection. We developed the supervised component generalized linear regression (SCGLR) method to overcome these limitations (Bry, Trottier, Verron, and Mortier 2012, 2013). SCGLR is a multivariate extension of partial least squares (PLS) regression to the generalized linear framework. It allows the relevant information contained in the data to be summarized in a few common components that can predict, as best as possible, the multivariate outcomes. The method

was motivated by ecological applications where there is interest in understanding how communities of tree species are structured based on environmental traits. Because species data can be collected through different measurement processes, the outcomes arise from several types of distributions. For example, some species may just be measured through presence/absence and others through count data (e.g. binomial or Poisson distributions). The originality of the SCGLR approach is to allow the simultaneous modeling of distributions from exponential family; Bernoulli, binomial, Gaussian and Poisson distributions can currently be handled in the **SCGLR** package.

SCGLR is based on a multivariate GLM and performs a PLS regression on each step of the GLM estimation algorithm. It uses both the responses (*e.g.* species abundances) and the regressors to calculate common components. Components are constructed sequentially: the first one maximizes some trade-off between its variance and the goodness of fit of the GLM that takes it as sole regressor (cf below for more details), the second one is its complement in the space orthogonal to the first component, etc, until we get a set of K complementary and mutually independent components, just as in principal component analysis (PCA). Ultimately, these components are used in a GLM as covariates, allowing them to have specific effects on each response. The optimal number of components on which to base the linear predictors is the one that allows the best prediction in cross-validation. The quality of prediction is assessed through various well-known criteria.

In this paper, we introduce an R-package that performs SCGLR. We first briefly review the mathematical basis of the method, then describe the program's features and usage. We illustrate **SCGLR** on a dataset built from the CoForChange database (`www.coforchange.eu`). It gives the abundance of 27 common tree genera in the tropical moistforest of the Congo-Basin and measurements on 40 geo-referenced environmental variables for one thousand 8 by 8 km plots (observations). Each plot's data were obtained by aggregating data measured on a variable number of previously sampled 0.5 ha sub-plots. Geo-referenced environmental variables were used to describe the physical factors as well as vegetation characteristics.

2. Description of the SCGLR statistical approach

SCGLR assumes that the q responses, $Y = (Y_1, \ldots, Y_q)$, are dependent on an unknown number of mutually orthogonal components (linear combinations of the covariates X), along with additional covariates A. The components are assumed common to all the responses in that they play some role in the GLM fit of each response. Moreover, the components are designed to stay rather close to the principal directions of the covariates, that is, stray from the noise contained in the group of regressors. Let X be the covariate matrix of size $n \times p$, and u be a p-coefficient vector. Just as in PCA, the structural strength of a component $f = Xu$ is measured through its variance under a unit-norm constraint on u. The components are determined sequentially. The first component $f_1 = Xu_1$ optimizes a trade-off between the goodness of fit of a multivariate GLM using f_1 as common explanatory variable and the variance of f_1. Then, X is deflated on f_1, yielding residual predictor matrix X_1, and the second component f_2 is sought in X_1 ($f_2 = X_1 u_2$) according to the same trade-off optimization, but taking f_1 as covariate. And so forth for higher rank components. Finally, given some integer K, a multivariate GLM of the responses is performed on the set F_K of the first K components, yielding a coefficient vector γ_j for each response y_j, with corresponding linear predictor η_j ($j = 1, \ldots, q$). Now, each component f_k can be expressed as a linear combination of the

original predictors: $f_k = Xv_k$. Hence, in matrix form: $F_K = XV_K$. Thus, we can express each linear predictor as a linear combination of the regressors: $\eta_j = X\beta_j$ with $\beta_j = V_K\gamma_j$. The coefficients β_j can be used in cross-validation to determine the optimal number K of components in order to avoid overfitting. Given the number of components K under trial, the observations are repeatedly partitioned into 2 sub-samples: C (for calibration) and T (for testing). On each partition, C is used to calculate the K components, and hence the β's, which in turn are used to predict the expectation of the responses on T. An appropriate criterion of predictive power is then calculated (depending on the distribution of the responses) and averaged over all (C, T) partitions considered. Eventually, we considered and we select the number K yielding the best performance.

3. Program description and usage

3.1. Main description

SCGLR is developed using R $\geq$ 3.0 version (R Core Team 2013). **SCGLR** is a set of R functions illustrated on a floristic data set, *genus*. `scglr()` and `scglrCrossval()` are the two main high level functions, which are respectively dedicated to fitting the model and selecting the number of components. `print()`, `summary()` and `plot()` methods are also available for the `scglr()` function.

```
> library(SCGLR)
```

The call to `scglr()` has the following structure:

```
> results.scglr <- scglr(formula,data,family,K,size,offset,subset,na.action,crit)
```

The *formula*, *data*, *family* and K arguments are required and *size* must be specified if binomial variables are used. The *formula* object of the *Formula* class (Zeileis and Croissant 2010) is composed of two or three terms. The first term describes the dependent variables whereas the second term describes the regressors used to construct components, and the third term describes additional covariates to be included in the model but not used in the linear combination giving the components. The first two terms should be separated by a $\sim$ symbol as classical R formula objects, whereas the second and third terms, if any, should be separated by a | symbol. All the elements in each term are separated by a + sign. The formula can be written out explicitly or provided using the `multivariateFormula()` function. For example, if $ny = (\text{“}y1\text{”}, \text{“}y2\text{”})$ contains the names of the dependent variables, $nx = (\text{“}x1\text{”}, \ldots, \text{“}x5\text{”})$ the names of the regressors used to construct the component, and $nz = (\text{“}z1\text{”}, \text{“}z2\text{”}, \text{“}z3\text{”})$ the names of the additional regressors

```
> myformula <- multivariateFormula(ny,nx,nz)
> myformula
```

```
y1 + y2 ~ x1 + x2 + x3 + x4 + x5 | z1 + z2 + z3
```

The *data* argument is an object of the *data.frame* class. The *family* is a vector of characters describing the family of each dependent variable. In `SCGLR`, "bernoulli", "binomial", "poisson" or "gaussian" are allowed. For Poisson outcomes, the *offset* argument is either a vector or a matrix of size: number of observations $\times$ number of Poisson dependent variables, allowing a different offset for each dependent variable. If binomial dependent variables are included in the model, *size* must be specified as a matrix describing the number of trials.

The output of the `scglr` function is an object of class **SCGLR** made of:

- u: matrix of size: number of regressors $\times$ number of components, contains the component-loadings, i.e. the coefficients of the regressors in the linear combination giving each component.
- comp: matrix of size : number of statistical units $\times$ number of components, having the components as column vectors.
- compr: matrix of size : number of statistical units $\times$ number of components, having the standardized components as column vectors.
- gamma: matrix of size: number of components $\times$ number of dependent variables, contains the coefficients of the regression on the components.
- beta: matrix of size: number of regressors + 1 (intercept) $\times$ number of dependent variables, contains the coefficients of the regression on the original regressors X.
- lin.pred: data.frame of size: number of statistical units $\times$ number of dependent variables, the fitted linear predictor.
- xFactors: data.frame containing the nominal regressors.
- xNumeric: data.frame containing the quantitative regressors.
- inertia: matrix of size: number of components $\times$ 2, contains the percentage and cumulative percentage of the overall regressors' variance, captured by each component.
- deviance: vector of length: number of dependent variables, gives the deviance of each y_k's GLM on the components.

The `print()` method gives the values of inertia and deviance. `summary()` gives inertia, deviance, and three additional tables. The first one contains the square correlations between X's and the components, along with two columns highlighting the plane on which the regressors are best projected and their associated square correlations. The second table presents the square correlations between fitted linear predictors and components, with two more columns corresponding to the plane on which the regressors are best projected and their associated square correlations. These two tables summarize how well the regressors and the dependent variables, through their linear predictors, are represented on the planes. The third table presents the γ values obtained from the GLM; only γ's with p-values lower than a given cutoff (default 0.05) are printed.

3.2. Plots

Several specialized plot commands are available to show the results of `scglr()`. They are all based on the **ggplot2** package developed by Wickham (2009) and as such can be further customized (i.e. one can add more layers or labels for example).

- `plot()` (see Figure 1): general function to produce various plots from the `scglr()` output by selecting elements to draw. This selection is specified by the *style* argument as a character vector with parameters chosen in Table 1: (nb: *style* elements can be abbreviated)

Style elements	descriptions
covariates	regressors drawn as black arrows.
observations	standardized observations drawn as points.
predictors	linear predictors drawn as red arrows.
circle	correlation circle.
threshold	dashed threshold circle with radius equal to threshold value. Covariates and linear predictors will be filtered accordingly. Default value of threshold (0.8) can be overridden by using the threshold parameter.
factor	centre of observations grouped by factor levels. Factor name must be provided by using factor parameter.

Table 1: Table presenting the *style* parameters available.

- `barplot()` (see Figure 2): takes an **SCGLR** object as input and produces a barplot of the inertia per component.
- `pairs()` (see Figure 3): takes an **SCGLR** object and produces an array plot for pairwise combinations of components (all components or a selected subset).

3.3. Selecting the number of components

Most of the time, the appropriate number of components to best predict dependent variables remains unknown and must be selected. We propose a cross-validation approach using different criteria to determine the number of components. The call to the `scglrCrossVal()` function shares the same arguments as the `scglr()` function with two additional arguments *nfolds* and *type*:

```
> scglrCrossVal(formula,data,family,K,nfolds,types,size,
+                  offset,subset,na.action,crit)
```

nfolds is the number of subsamples to be used in the cross-validation - default is 5. Although *nfolds* can be as large as the sample size (leave-one-out CV), this is not recommended for large datasets.
type is the criterion to use for cross-validation. Currently five options are available in a general setting: “mspe” (Mean Squared Prediction Error), “likelihood”, “aic”, “bic” and “aicc”. When

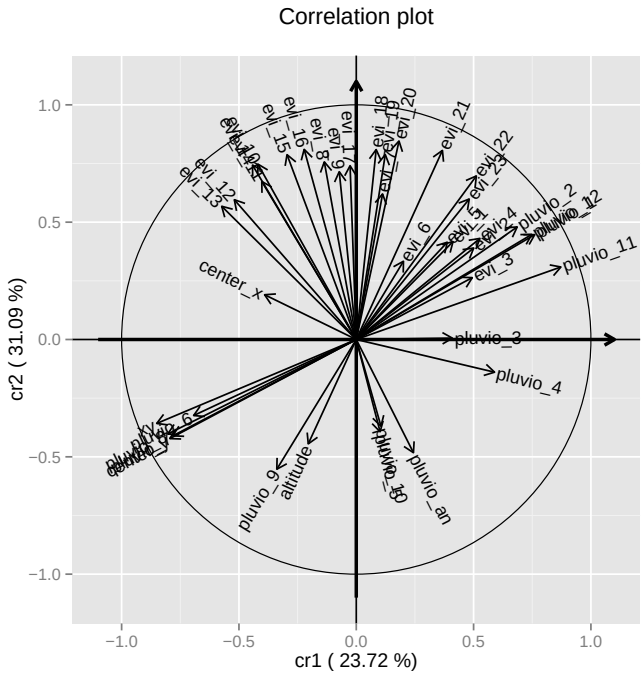


Figure 1: Simple correlation circle on the first plane

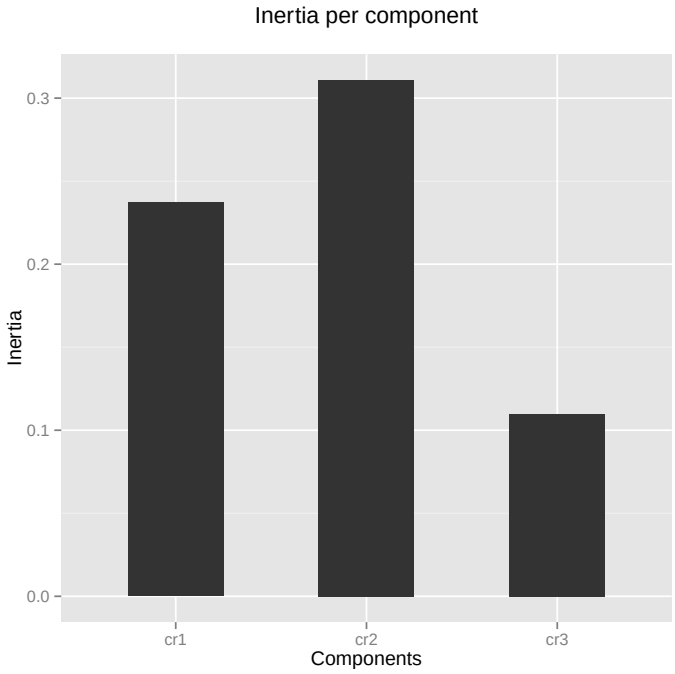


Figure 2: Barplot of the inertia (component's varaince).

all dependent variables are Bernoulli, the option "auc" (area under ROC curve) enables to measure the prediction performance.

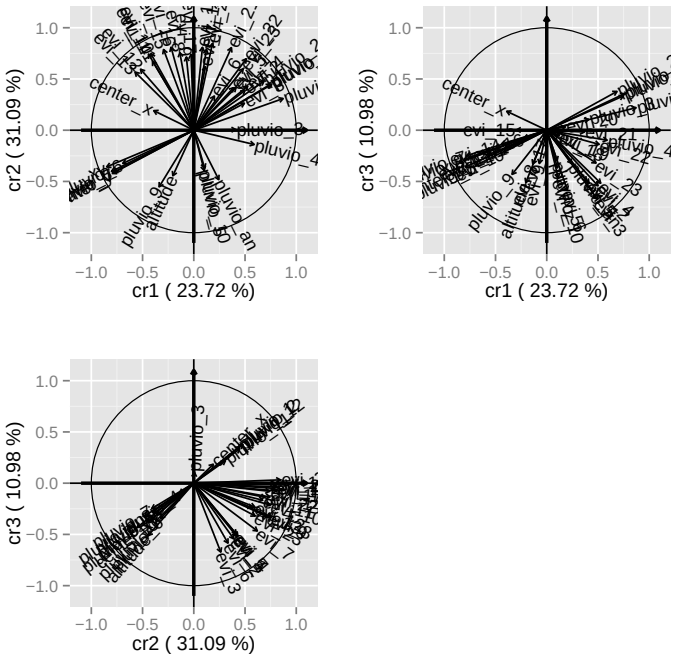


Figure 3: Matrix-plot of correlation circles on all planes.

The output of the procedure is a $(q \times (K+1))$ matrix containing the criterion values for each response variable and each model. The first column corresponds to the model without any component.

4. Examples

4.1. Floristic data set

We illustrate **SCGLR** using the data *genus*. This example highlights the use of the multivariate Poisson count distribution with an offset.

4.2. Count data

genus is a dataset built from the CoForChange database. It gives the abundance of 27 common tree genera in the tropical moistforest of the Congo-Basin and measurments on 40 geo-referenced environmental variables for one thousand 8 by 8 km plots (observations). Data on each plot were obtained by aggregating the data measured on a variable number of previously sampled 0.5 ha sub-plots. The geo-referenced environmental variables were used to describe 16 physical factors pertaining to the description of topography, geology and rainfall and the remaining variables give the vegetation characteristics defined through 16-days enhanced vegetation index (EVI).

```
> data(genus)
```

```
> dim(genus)

[1] 1000   69

> names(genus)

 [1] "gen1"      "gen2"      "gen3"      "gen4"
 [5] "gen5"      "gen6"      "gen7"      "gen8"
 [9] "gen9"      "gen10"     "gen11"     "gen12"
[13] "gen13"     "gen14"     "gen15"     "gen16"
[17] "gen17"     "gen18"     "gen19"     "gen20"
[21] "gen21"     "gen22"     "gen23"     "gen24"
[25] "gen25"     "gen26"     "gen27"     "altitude"
[29] "pluvio_yr" "forest"    "pluvio_1"  "pluvio_2"
[33] "pluvio_3"  "pluvio_4"  "pluvio_5"  "pluvio_6"
[37] "pluvio_7"  "pluvio_8"  "pluvio_9"  "pluvio_10"
[41] "pluvio_11" "pluvio_12" "geology"   "evi_1"
[45] "evi_2"     "evi_3"     "evi_4"     "evi_5"
[49] "evi_6"     "evi_7"     "evi_8"     "evi_9"
[53] "evi_10"    "evi_11"    "evi_12"    "evi_13"
[57] "evi_14"    "evi_15"    "evi_16"    "evi_17"
[61] "evi_18"    "evi_19"    "evi_20"    "evi_21"
[65] "evi_22"    "evi_23"    "lon"       "lat"
[69] "surface"
```

We chose to use the covariate "geology" as an additional factor not directly used in the component construction because of the demonstrated importance of the geological substrates on the spatial distribution of tree species in the Congo Basin (Fayolle, Engelbrecht, Freycon, Mortier, Swaine, Réjou-Méchain, Doucet, Fauvet, Cornu, and Gourlet-Fleury 2012). We also used the covariate "surface" as an offset and we added the product $I(lon * lat)$ as a new covariate.

```
> ny <- names(genus)[1:27]
> sx <- which(names(genus) %in% c("geology","surface"))
> nx <- names(genus)[-c(1:27,sx)]
> family <- rep("poisson",length(ny))
> formula <- multivariateFormula(ny,c(nx,"I(lon*lat)"),"geology")
> formula

gen1 + gen2 + gen3 + gen4 + gen5 + gen6 + gen7 + gen8 + gen9 +
    gen10 + gen11 + gen12 + gen13 + gen14 + gen15 + gen16 + gen17 +
    gen18 + gen19 + gen20 + gen21 + gen22 + gen23 + gen24 + gen25 +
    gen26 + gen27 ~ altitude + pluvio_yr + forest + pluvio_1 +
    pluvio_2 + pluvio_3 + pluvio_4 + pluvio_5 + pluvio_6 + pluvio_7 +
    pluvio_8 + pluvio_9 + pluvio_10 + pluvio_11 + pluvio_12 +
    evi_1 + evi_2 + evi_3 + evi_4 + evi_5 + evi_6 + evi_7 + evi_8 +
```

```
    evi_9 + evi_10 + evi_11 + evi_12 + evi_13 + evi_14 + evi_15 +
    evi_16 + evi_17 + evi_18 + evi_19 + evi_20 + evi_21 + evi_22 +
    evi_23 + lon + lat + I(lon * lat) | geology

> offset <- genus$surface

> genus.cv <- scglrCrossVal(formula=formula,data=genus,family=family,
+                           K=15,nfolds=5,type="mspe",offset=offset,mc.cores=1)
```

mc.cores is an optional argument to launch parallel runs of the cross-validation procedure (default is equal to one).

Concerning the selection procedure, in order to produce comparable values for possibly very different response variables, we used the following heuristic. For each response and each of the $K+1$ models (one model for each number K of components and one for no component), divide the criterion value by its median over all the models. Then calculate for each number of components the mean of the standardized values over the different response variables. Alternatively, the mean can be used to normalize instead of the median.

```
> criterion <- t(apply(genus.cv,1,function(x) x/median(x)))
> criterion <- apply(criterion,2,mean)
> K.cv <- which.min(criterion)-1
```

In the expression of *K.cv*, the minus 1 enables to relable the output such that it matches the actual number of components used. Plotting *criterion* values (see Figure 4) displays the change in the selection criterion as the number of components increases. Here, the *criterion* is minimized for 8 components The number of components that minimizes the criterion 8 . We can therefore call `scglr()` with $K = 8$.

```
> genus.scglr<-scglr(formula=formula,data=genus,family=family,
+                    K=K.cv,size=NULL,offset=offset)
```

Printing *genus.scglr*:

```
> print(genus.scglr)

Call:  scglr(formula = formula, data = genus, family = family, K = K.cv,
    size = NULL, offset = offset)

Inertia:
       cr1         cr2         cr3         cr4         cr5
0.22677183  0.31465078  0.09737181  0.04449536  0.07549732
       cr6         cr7         cr8
0.02552999  0.03509469  0.02029163

Deviance:
     gen1       gen2       gen3       gen4       gen5
 2034.403   2438.178   1316.709   1258.037   1411.318
```

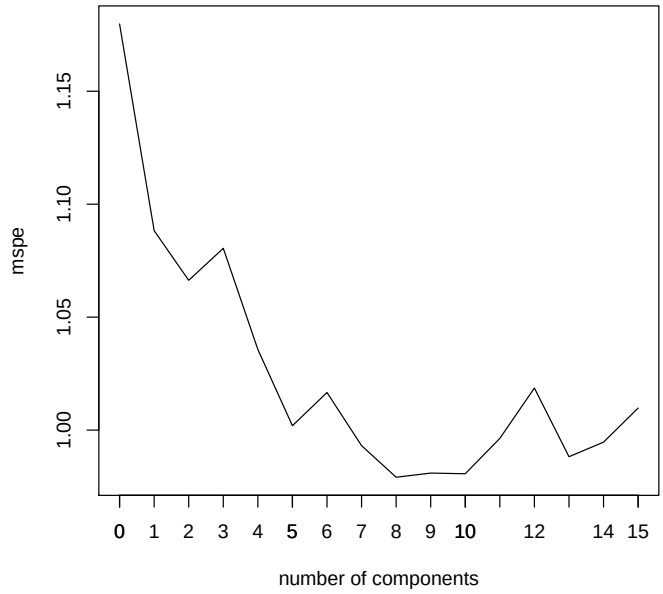


Figure 4: Mean Squared Prediction Error (MSPE) as a function of the number of components.

```
     gen6       gen7       gen8       gen9      gen10
1758.677   2292.380  15159.303   1507.668  14610.577
    gen11      gen12      gen13      gen14      gen15
4726.183   6567.454   7577.443   8234.611   2243.921
    gen16      gen17      gen18      gen19      gen20
2066.180   2639.012   1222.103  25234.683   2061.224
    gen21      gen22      gen23      gen24      gen25
3231.061   1345.524   3258.163   8126.111  10182.539
    gen26      gen27
9333.937   9976.224
```

Inertia of the 8 components (see Figure 5):

```
> barplot(genus.scglr)
```

The following two commands create the plots in Figure 6. The first one gives a simple correlation plot (see Figure 6a)

```
> plot(genus.scglr)
```

The second correlation plot (see Figure 6b) displays only the linear predictors and covariates whose norms in the selected plane exceed the threshold specified by the "thr" *styles* element.

```
> plot(genus.scglr, style=c("simple","predictor","thr"), thr=0.8)
```

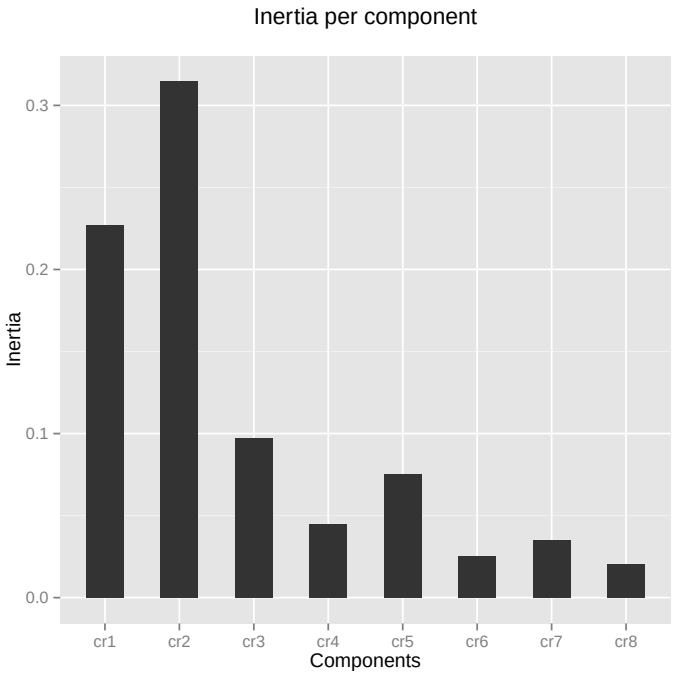
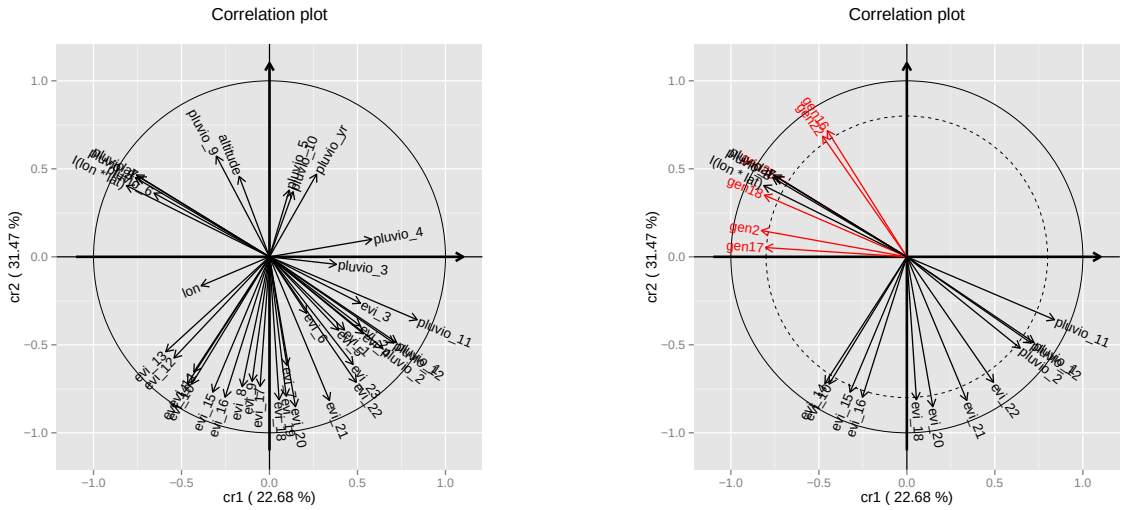


Figure 5: Barplot of inertia per component



(a) Simple correlation plot

(b) Correlation plot with linear predictors and covariates passing a threshold of 0.8.

Figure 6: Two sample plots

Finally, we present the *pairs* plot on the planes spanned by components 1,3,5 and 8 (see Figure 7):

```
> pairs(genus.scglr,plans=c(1,3,5,8),ncol=2,label.size=0.5)
```

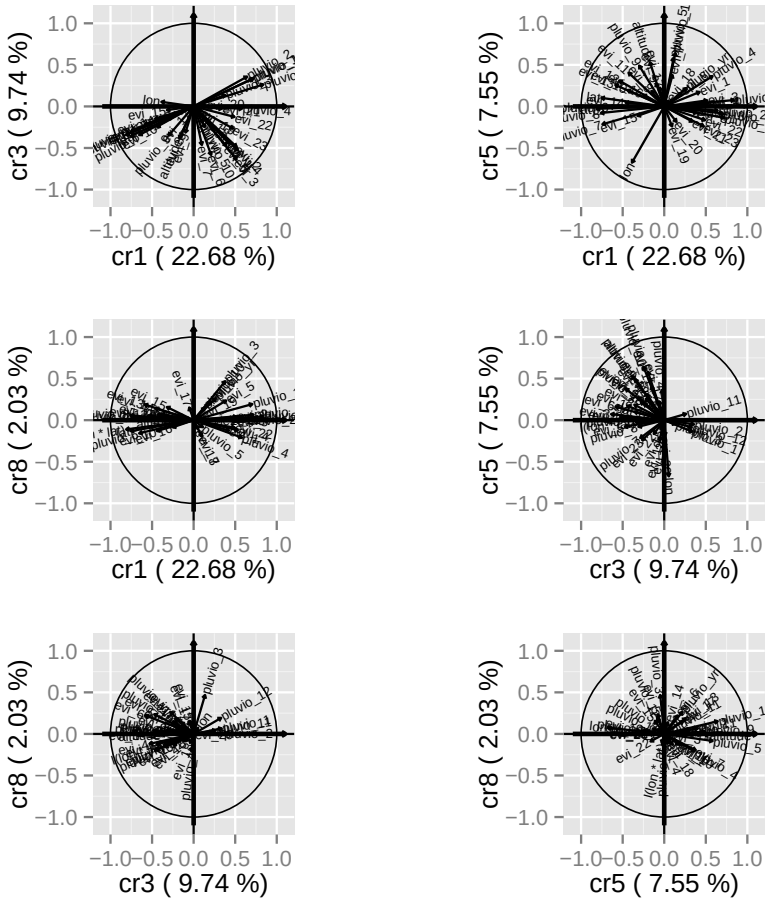


Figure 7: Correlation plots on planes spanned by components 1, 3, 5 and 8

5. Conclusion

The main features of the R package **SCGLR** have been explained and illustrated in this paper using the data set *genus* provided with the package. Contrary to existing PLS-dedicated packages that can only handle Gaussian data, **SCGLR** provides a unified framework to deal with multivariate outcomes arising from any exponential family distribution. The computational time required to run `scglr` depends on the dimension of the problem. Table 2 provides the mean user times required to run 100 simulations of the *scglr()* algorithm using one component with $p = 100$ covariates and a varying number of dependent variables ($q = 10$ and 100) and varying sample sizes ($n = 100, 1000$ and $10,000$). These results highlight the efficiency of the **SCGLR** package.

q \ n	100	1000	10000
10	0.075	1.11	17.10
100	2.942	5.644	31.60

Table 2: Mean user times (in seconds) to calculate one SCGLR component for 10 or 100 dependent variables and 100, 1000 or 10,000 observations

6. Acknowledgments

This research was supported by ITG-SEITA and was part of the CoForChange project (www.coforchange.eu), funded by the ERA-Net BiodivERsA, with the national funders ANR (France) and NERC (UK), part of the 2008 BiodivERsA call for research proposals, involving 16 European, African and international partners including a number of timber companies (see the list on the website, http://www.coforchange.eu/partners), and of the CoForTips project, funded by the ERA-Net BiodivERsA, with the national funders FWF (Austria), BelSPO (Belgium) and ANR (France), part of the 2011-2012 BiodivERsA call for research proposals (http://www.fordev.ethz.ch/research/active/CoForTips).

References

Bry X, Trottier C, Verron T, Mortier F (2012). "Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm." In *COMPSTAT 2012 (Cyprus)*.

Bry X, Trottier C, Verron T, Mortier F (2013). "Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm." *Journal of Multivariate Analysis*, **119**(0), 47 – 60. ISSN 0047-259X. doi:10.1016/j.jmva.2013.03.013. URL http://www.sciencedirect.com/science/article/pii/S0047259X13000407.

Fayolle A, Engelbrecht B, Freycon V, Mortier F, Swaine M, Réjou-Méchain M, Doucet JL, Fauvet N, Cornu G, Gourlet-Fleury S (2012). "Geological substrates shape tree species and trait distributions in African moist forests." *Plos One*. URL http://www.plosone.org/article/info%3Adoi/10.1371/journal.pone.0042381.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. URL http://www.jstatsoft.org/v34/i01/.

Affiliation:

Frédéric Mortier UPR "Biens et Services des Ecosystèmes Forestiers tropicaux (B&SEF)
Département Environnements et Sociétés du CIRAD
Campus International de Baillarguet, TA C-105/D
34398 Montpellier Cedex 5 - France
E-mail: fmortier@cirad.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* 2013-18-07
MMMMMM 2013 *Accepted:* yyyy-mm-dd