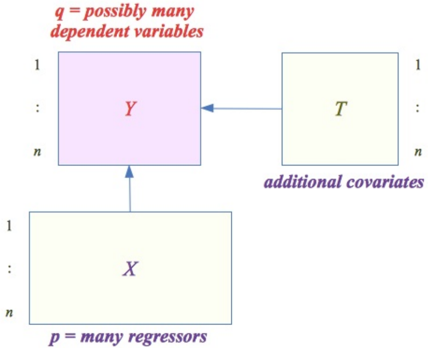
SCGLR: a component-based multivariate regression method to model species distributions.

Frédéric Mortier
with Xavier Bry, Guillaume Cornu & Catherine Trottier

International Statistical Ecology Conference

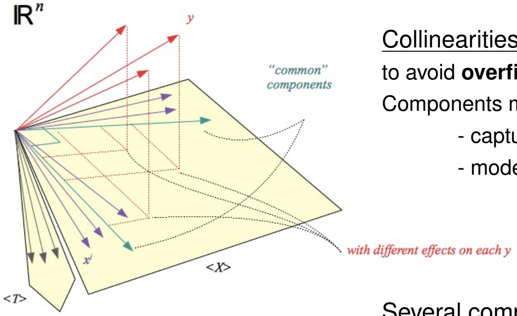
Montpellier
July 2014

Problematic



$\Rightarrow$ Question: What in X may predict what in Y?

$\hookrightarrow$ *Approach: Dimension reduction by construction of components*



Collinearities:

to avoid **overfitting**, search for **components.**

Components must:

- capture enough variance in X,
- model and predict y.

Several components:

to avoid **redundancy**, search for uncorrelation.

$\rightarrow$ constraint of construction: **orthogonality.**

Multiple y:

same components,

but each y with its **own regression coefficients.**

Exponential family distributed

$\rightarrow$ **generalized linear regression.**

• **First component** is a **compromise** between the direction of X that best predicts y and the first principal component (PC) of X.

$\hookrightarrow$ *Criterion:* $\quad max_{||u||^2=1}\,[cov(y, Xu)]$

$$max_{||u||^2=1}\left[\sqrt{var(y)}\,\sqrt{var(Xu)}\,corr(y, Xu)\right]$$

$\hookrightarrow$ *Program to solve:* $\quad P_1\,:\,max_{||u||^2=1}\,[< y, Xu >_W]$

• **Further components**: W-orthogonality of components is ensured using the part of X that is not yet used, i.e. the residuals of X regressed on previous components.

• **First component** can be obtained using several equivalent programs:

$\hookrightarrow \quad P_2 : max_{||u||^2,||v||^2=1}\, [< Xu, Yv >_W]$

$\hookrightarrow \quad P_3 : max_{||u||^2=1}\, \left[\sum_{k=1}^{q} < Xu, y^k >_W^2\right]$

P_3 is adapted to the case of multiple weighting:

$\hookrightarrow \quad P_4 : max_{||u||^2=1}\, \left[\sum_{k=1}^{q} < Xu, y^k >_{W_k}^2\right]$

$\Longrightarrow$ *Solution: eigenvector associated to largest eigenvalue of:*

$$A = X'\Omega X \text{ with } \Omega = \sum_{k=1}^{q} W_k y^k y'^k W_k$$

• **Further components**: idem, subject to constraint of orthogonality to previous components.

Multiple GLM with common predictor

In the GLM, linear predictors are constrained to be collinear to one another:

$$\forall k = 1, q : \quad \eta^k = X\beta_k + T\delta_k = X\gamma_k u + T\delta_k$$

$\longrightarrow$ **modified Fisher Scoring Algorithm**:
u and $\gamma = (\gamma_k)_{k=1,q}$ estimated iterating an alternated least squares two steps sequence:

• Given γ, working data $(z^k)_k$ is regressed on matrix $[\gamma \otimes X, 1_q \otimes T]$ with respect to working matrix $W = diag[W_k]_k$
$\longrightarrow$ coefficient vectors $\hat{u}$, $\hat{\delta} = (\hat{\delta}_k)_k$
$\longrightarrow$ $\hat{u}$ made unit norm $\longrightarrow$ updated u

(2) Given Xu, each working vector z^k is regressed on $[Xu, T]$ with respect to working matrix W_k
$\longrightarrow$ updated γ_k, δ_k

Step *t* of the FSA:

$$min_{\gamma,u:u'u=1}\left[\sum_k ||z^{k[t]} - X\gamma_k u||^2_{W_k^{[t]}}\right]$$
$$\Leftrightarrow min_{u:u'u=1}\left[\sum_k ||z^{k[t]} - \Pi_{Xu} z^{k[t]}||^2_{W_k^{[t]}}\right]$$
$$\Leftrightarrow max_{u:u'u=1}\left[\sum_k ||z^{k[t]}||^2_{W_k^{[t]}} cos^2_{W_k^{[t]}}(z^{k[t]}, Xu)\right]$$

is **replaced** by: $max_{u:u'u=1}\left[\sum_k ||z^{k[t]}||^2_{W_k^{[t]}} cos^2_{W_k^{[t]}}(z^{k[t]}, Xu)\ \ ||Xu||^2_{W_k^{[t]}}\right]$
equivalent to:

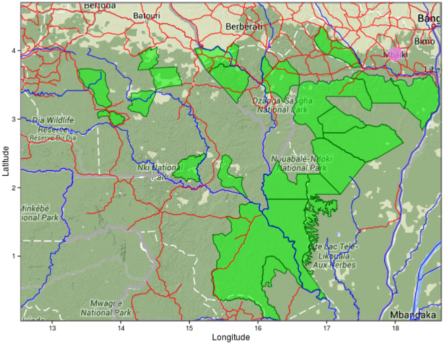
$$max_{u:u'u=1}\left[\sum_k < z^{k[t]}, Xu >^2_{W_k^{[t]}}\right]$$

= local extended PLS2

$\Longrightarrow$ *Solution: eigenvector associated to largest eigenvalue of:*

$$A = X'\Omega^{[t]}X \text{ with } \Omega^{[t]} = \sum_{k=1}^{q} W_k^{[t]} z^{k[t]} z'^{k[t]} W_k^{[t]}$$

Abundance of tropical tree species (CoForChange project)

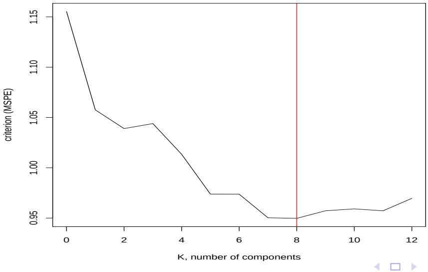


- all trees with diameter higher than 30 cm
- more than 120,000 plots of 0.5 ha
- more than 200 genera
- soil, rainfall, human disturbances, vegetation activity (EVI) maps available

Application II

Select number of component: a cross-validation approach

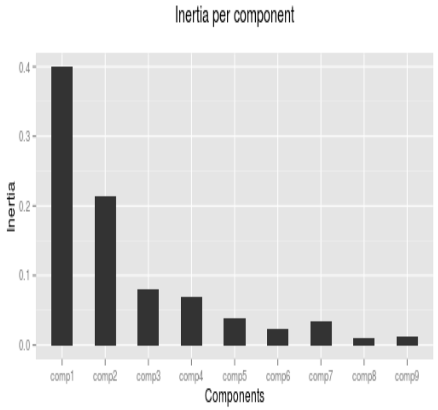
```
> library(SCGLR)
> genus.cv <- scglrCrossVal(formula=form, data=genus, family=fam, K=12,
+  offset=genus$surface)
>
> mean.crit <- t(apply(genus.cv,1,function(x) x/mean(x)))
> mean.crit <- apply(mean.crit,2,mean)
> K.cv <- which.min(mean.crit)-1
> cat("Best number of components: ",K.cv)

Best number of components:  8
```

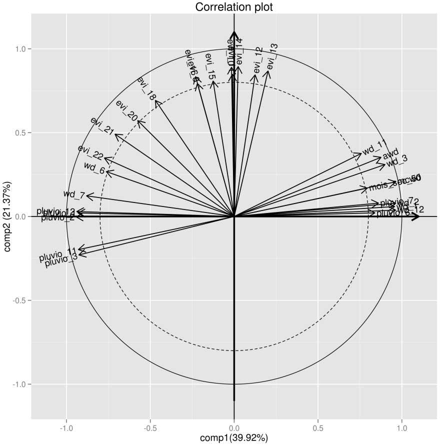


Application III

Fitting and Plots

```
>   genus.scglr <- scglr(formula=form, data=genus, family=fam, K=K.cv,
+  offset=genus$surface)
```

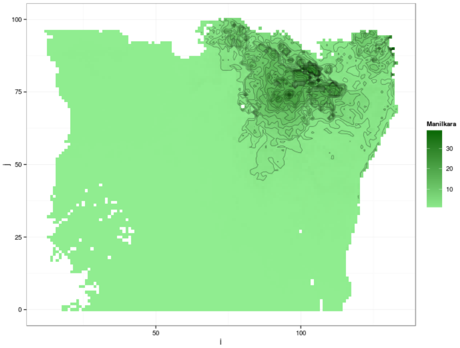




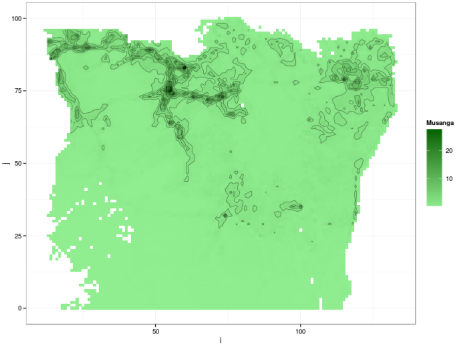
Application IV

Prediction of two genera

Manilkara mabokeensis



Musanga cecropioides



Ongoing works

- New alternate optimization algorithms: Iterative Normalized Gradient
- Multi-table (Theme) support
- SCGLR packages new versions
  - Enhancements for plot customization
  - New distribution families (Negative-Binomial, Exponential, Inverse Gaussian)
  - Multi-theme
  - ...

References

1. X. Bry, C. Trottier, T. Verron and F. Mortier (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119(0), 47.
2. S. Gourlet-Fleury et al. (2009–2014) CoForChange project: `http://www.coforchange.eu/`.
3. F. Mortier, C. Trottier, G. Cornu and X. Bry (2014). SCGLR: Supervised Component Generalized Linear Regression (SCGLR). *R package version 1.2*. `http://CRAN.R-project.org/package=SCGLR`
4. F. Mortier, C. Trottier, G. Cornu and X. Bry (2014). SCGLR - An R package for Supervised Component Generalized Linear Regression. *Journal of Statitistical Software*. submitted