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
Question: What in X may predict what in Y?

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. → constraint of construction: orthogonality.

Multiple $y$:
**same** components,
but each $y$ with its own regression coefficients.

Exponential family distributed
→ **generalized linear regression**.
• **First component** is a *compromise* between the direction of \( X \) that best predicts \( y \) and the first principal component (PC) of \( X \).

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

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

\[
\text{Program to solve: } P_1 : \max_{||u||^2=1} [\langle y, Xu \rangle_W]
\]

• **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:

\[ P_2 : \max_{\|u\|^2,\|v\|^2=1} \left[ \langle Xu, Yv \rangle \right] \]

\[ P_3 : \max_{\|u\|^2=1} \left[ \sum_{k=1}^{q} \langle Xu, y^k \rangle^2 \right] \]

\( P_3 \) is adapted to the case of multiple weighting:

\[ P_4 : \max_{\|u\|^2=1} \left[ \sum_{k=1}^{q} \langle Xu, y^k \rangle^2_{W_k} \right] \]

\[ \implies \text{Solution: eigenvector associated to largest eigenvalue of:} \]

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

• **Further components**: idem, subject to constraint of orthogonality to previous components.
In the GLM, linear predictors are constrained to be collinear to one another:

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

\rightarrow \text{modified Fisher Scoring Algorithm:}

\(u\) and \(\gamma = (\gamma_k)_{k=1,q}\) estimated iterating an alternated least squares two steps sequence:

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

(2) Given \(Xu\), each working vector \(z^k\) is regressed on \([Xu, T]\) with respect to working matrix \(W_k\)
  \(\rightarrow\) updated \(\gamma_k, \delta_k\)
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:  

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

equivalent to:  

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

$= \text{local extended PLS2}$

$\Rightarrow \text{Solution: eigenvector associated to largest eigenvalue of:}$

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

```r
> 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


