Supervised–Component versus PLS regression The case of GLMMs with autoregressive random effect

Jocelyn CHAUVET
joint work with **Catherine TROTTIER** and **Xavier BRY**

CASI 2018, Galway,
May 18

A simple Gaussian model

- $\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma} = \mathsf{Id}_n\right)$
- $\boldsymbol{X} = \Bigg[\; \underbrace{\boldsymbol{x}^{\mathbf{1}} \ldots\ldots\ldots\ldots \boldsymbol{x}^{\mathbf{10}}}_{\substack{\text{large bundle} \\ \hookrightarrow \text{noise}}} \;\; \underbrace{\boldsymbol{x}^{\mathbf{11}} \ldots\ldots \boldsymbol{x}^{\mathbf{15}}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{predicts y}}} \;\; \underbrace{\boldsymbol{x}^{\mathbf{16}} \ldots\ldots \boldsymbol{x}^{\mathbf{20}}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{predicts y}}} \Bigg]$

☛ **PLSR vs "Supervised Component Regression"**

(a more flexible way to build components)

A simple Gaussian model

- $\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma} = \mathsf{Id}_n\right)$
- $\boldsymbol{X} = \Big[\ \underbrace{\boldsymbol{x}^1 \ldots\ldots\ldots\ldots \boldsymbol{x}^{10}}_{\substack{\text{large bundle} \\ \hookrightarrow \text{ noise}}}\ \underbrace{\boldsymbol{x}^{11} \ldots\ldots \boldsymbol{x}^{15}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{ predicts y}}}\ \underbrace{\boldsymbol{x}^{16} \ldots\ldots \boldsymbol{x}^{20}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{ predicts y}}}\Big]$

☛ **PLSR vs "Supervised Component Regression"**
(a more flexible way to build components)

- Supervised Component Regression also available for Bernoulli, binomial and Poisson responses (**R package : SCGLR**)

How to extend it to GLMMs with both individual– and time–specific random effects ?

We consider balanced panel data with:

- N individuals ...
- ...observed at the same R time–points

Notations:

- $\boldsymbol{y}_{NR\times 1}$: response vector
- $\boldsymbol{X}_{NR\times p}$: design matrix of the **many and redundant** explanatory variables

We consider balanced panel data with:

- N individuals ...
- ...observed at the same R time–points

Notations:

- $\boldsymbol{y}_{NR \times 1}$: response vector
- $\boldsymbol{X}_{NR \times p}$: design matrix of the many and redundant explanatory variables

Difficulties

- High level of correlation among the explanatory variables
  ↪ **Regularisation** is needed
- Individual–specific and time–specific effects
  ↪ Need to take into account the induced **complex dependence structure**

Example of real data

$\hookrightarrow$ **Econometrics:** all companies share a common economic climate (latent phenomenon) which tends to persist over time...

Example of real data

↪ **Econometrics:** all companies share a common economic climate (latent phenomenon) which tends to persist over time...

In general: data with

- a **dependence within individuals** on which data is repeatedly collected
- a **serially correlated time–specific effect** shared by all the individuals

The method we propose must

- take into account the dependence structure:
  - ↪ Within–individual dependence modelled by a **random effect with independent levels**
  - ↪ Time dependence modelled by a **random effect with AR(1) levels**
  - ↪ **GLMM** in order to deal with non–Gaussian response (e.g. count or binary response)
- handle the high correlations among the explanatory variables
  - ↪ **Ridge–based regularisation**
  - ↪ **Supervised component–based regularisation**

GLMM framework

$$\boldsymbol{Y}_i \,|\, \boldsymbol{\xi} \overset{\text{iid}}{\sim} F \text{ belonging to the exponential family}$$

$$g(\underbrace{\mathbb{E}\left(\boldsymbol{Y} \,|\, \boldsymbol{\xi}\right)}_{\boldsymbol{\mu}}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2}$$

- $\boldsymbol{\beta}$ fixed effect vector
- $\boldsymbol{\xi^1} = \left(\xi^1_1, \xi^1_2, \ldots, \xi^1_N\right)^\intercal$ the "individual–specific" random effect vector, $\boldsymbol{U_1} = \mathbf{Id}_N \otimes \mathbf{1}_R$ the associated design matrix
- $\boldsymbol{\xi^2} = \left(\xi^2_1, \xi^2_2, \ldots, \xi^2_R\right)^\intercal$ the "time–specific" random effect vector, $\boldsymbol{U_2} = \mathbf{1}_N \otimes \mathbf{Id}_R$ the associated design matrix

Random effects

$$\boldsymbol{y} = \big(y_{\mathbf{11}}, y_{\mathbf{12}}, \dots, y_{\mathbf{1}\boldsymbol{R}}, \\ y_{\mathbf{21}}, y_{\mathbf{22}}, \dots, y_{\mathbf{2}\boldsymbol{R}}, \dots\dots \\ y_{\boldsymbol{N}\mathbf{1}}, y_{\boldsymbol{N}\mathbf{2}}, \dots, y_{\boldsymbol{N}\boldsymbol{R}}\big)^{\mathsf{T}}$$

- $\boldsymbol{\xi}^{\mathbf{1}} = \left(\xi_1^1, \xi_2^1, \dots, \xi_N^1\right)^{\mathsf{T}} \sim \mathcal{N}_N\left(\mathbf{0},\, \boldsymbol{D}_{\mathbf{1}}\right),\ \boldsymbol{D}_{\mathbf{1}} = \sigma_1^2\, \boldsymbol{A}_{\mathbf{1}}$
- $\boldsymbol{\xi}^{\mathbf{2}} = \left(\xi_1^2, \xi_2^2, \dots, \xi_R^2\right)^{\mathsf{T}} \sim \mathcal{N}_R\left(\mathbf{0},\, \boldsymbol{D}_{\mathbf{2}}\right),\ \boldsymbol{D}_{\mathbf{2}} = \sigma_2^2\, \boldsymbol{A}_{\mathbf{2}}(\rho),$

$$\boldsymbol{A}_{\mathbf{2}}(\rho) = \left(\frac{\rho^{|i-j|}}{1-\rho^2}\right)_{1\leqslant i,j\leqslant R}$$

- $\boldsymbol{\xi}^{\mathbf{1}} \perp \boldsymbol{\xi}^{\mathbf{2}}$

Principle of penalised EM algorithm

Green, P.J. (1990) *On use of the EM for penalized likelihood estimation. Journal of the Royal Statistical Society.* Series B (Methodological), 443-452.

$$\mathsf{E} : \mathcal{Q}_{\textsf{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\mathcal{L}_{\textsf{pen}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) \,|\, \boldsymbol{\theta}^{[t]}\right]$$

$$\mathsf{M} : \boldsymbol{\theta}^{[t+1]} \longleftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}_{\textsf{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$$

Principle of penalised EM algorithm

Green, P.J. (1990) *On use of the EM for penalized likelihood estimation. Journal of the Royal Statistical Society.* Series B (Methodological), 443-452.

$$\textsf{E} : \mathcal{Q}_{\textsf{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\mathcal{L}_{\textsf{pen}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) \,|\, \boldsymbol{\theta}^{[t]}\right]$$

$$\textsf{M} : \boldsymbol{\theta}^{[t+1]} \longleftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}_{\textsf{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$$

Usual penalised complete log-likelihood

$$\mathcal{L}_{\textsf{pen}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) - \lambda\, \textsf{pen}(\boldsymbol{\beta})$$

$$\textsf{pen}(\boldsymbol{\beta}) = \begin{cases} \|\boldsymbol{\beta}\|_1 \\ \|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} \\ \alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1, \quad 0 \leqslant \alpha \leqslant 1 \end{cases}$$

Ridge-based regularisation

$\hookrightarrow$ **EM algorithm,** $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \rho\right)$ **and** $\boldsymbol{\xi} = \left(\boldsymbol{\xi}^{\mathbf{1}}, \boldsymbol{\xi}^{\mathbf{2}}\right)$

$$\mathsf{E} : \mathcal{Q}_{\textbf{ridge}}\left(\boldsymbol{\theta}, \lambda \,|\, \boldsymbol{\theta}^{[t]}\right) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) - \lambda\, \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{\beta} \,|\, \boldsymbol{\theta}^{[t]}\right]$$

$$\mathsf{M} : \begin{cases} \lambda^{[t+1]} \longleftarrow \mathsf{GCV}^{[t+1]}(\lambda) \\ \boldsymbol{\theta}^{[t+1]} \longleftarrow \underset{\boldsymbol{\theta}}{\arg\max}\; \mathcal{Q}_{\textbf{ridge}}\left(\boldsymbol{\theta}, \lambda^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right) \end{cases}$$

Eliot, M., Ferguson, J., Reilly, M.P. and Foulkes, A.S. (2011) *Ridge Regression for Longitudinal Biomarker Data.* The International Journal of Biostatistics, **7**, 1–11.

Component–based regularisation
$\hookrightarrow$ New linear predictors

$$\boldsymbol{Y}_i \,|\, \boldsymbol{\xi} \overset{\text{iid}}{\sim} F \text{ belongs to the exponential family}$$
$$g(\mathbb{E}\,(\boldsymbol{Y}\,|\,\boldsymbol{\xi})) = \boldsymbol{\eta} = \cancel{\boldsymbol{X}\boldsymbol{\beta}} + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2}$$

replaced with

$$\eta = (\boldsymbol{X}\boldsymbol{u})\gamma + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2} \qquad \text{for a single component}$$

extended to

$$\eta = \sum_{k=1}^{K^\star} (\boldsymbol{X}\boldsymbol{u_k})\gamma_k + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2} \qquad \text{for } K^\star \text{ components}$$

Complete log–likelihood for supervised component regularisation

With $\boldsymbol{\theta} = \left(\boldsymbol{u}, \gamma, \sigma_1^2, \sigma_2^2, \rho\right)$ and a trade-off parameter $s \in [0, 1]$

$$\mathcal{L}_{\text{SC}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = (1 - s) \, \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) + s \, \phi(\boldsymbol{u})$$

- **Log-likelihood :** measures (inter alia) the probability that observations $\boldsymbol{y}$ have been generated from component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$
- **Structural relevance criterion :** measures the closeness of component $\boldsymbol{f}$ to the strongest structures of $\boldsymbol{X}$

A few words about the structural relevance criterion

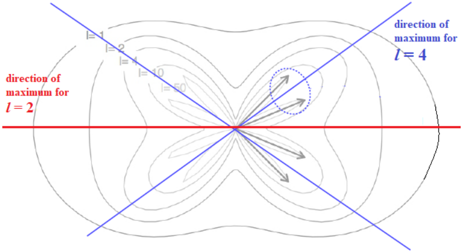
How many bundles do you see ?

$$\phi(\boldsymbol{u}) = \left(\sum_{j=1}^{p} \left[\text{cor}^2(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x^j}) \right]^{\ell} \right)^{\frac{1}{\ell}}$$

A few words about the structural relevance criterion

$$\phi(\boldsymbol{u}) = \left(\sum_{j=1}^{p} \left[\mathrm{cor}^2(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x}^j) \right]^{\ell} \right)^{\frac{1}{\ell}}$$

Iso-value curves of ϕ^{ℓ}



Ridge–based penalisation	Component–based regularisation
$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) - \lambda\, \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{\beta}$	$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) + \frac{s}{1-s}\, \phi(\boldsymbol{u})$

Ridge–based penalisation

- Penalises the "large" coefficients
- Sees the high correlations among the explanatory variables as **pure nuisance**
- η **hard to interpret**

Component–based regularisation

- Gives a bonus to the most interpretable bundles in $\boldsymbol{X}$
- Takes advantage of the high correlations among the explanatory variables
- η easier to interpret through decomposition on components

$\rightsquigarrow$ **Focus on GLMMs**

Starting with the classical Fisher Scoring Algorithm for GLMs, we perform:

LINEARISATION step

- Linearisation of $\boldsymbol{y}_i$ at $\boldsymbol{\mu}_i = \mathbb{E}\left(\boldsymbol{Y}_i \,|\, \boldsymbol{\xi}\right)$:

$$\boldsymbol{y}_i \simeq \boldsymbol{z}_i = g(\boldsymbol{\mu}_i) + (\boldsymbol{y}_i - \boldsymbol{\mu}_i) g'(\boldsymbol{\mu}_i)$$

$$\boldsymbol{z}_i = \boldsymbol{\eta}_i + \boldsymbol{e}_i$$

- Linearised model:

$$\mathcal{M} : \boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2} + \boldsymbol{e}, \quad \text{with } \mathbb{V}(\boldsymbol{e}) = \boldsymbol{\Gamma}$$

ESTIMATION step

Penalised/Regularised EM algorithm on $\mathcal{M}$

Ridge–based penalisation for GLMM–AR(1)

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \rho\right)$$

Linearised model

$$\mathcal{M}^{[t]} : \boldsymbol{z}^{[t]} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi^1} + \boldsymbol{U_2}\boldsymbol{\xi^2} + \boldsymbol{e}^{[t]}, \quad \text{with } \mathbb{V}\left(\boldsymbol{e}^{[t]}\right) = \boldsymbol{\Gamma}^{[t]}$$

Ridge estimation

$$\textbf{E} : \mathcal{Q}_{\textbf{ridge}}\left(\boldsymbol{\theta}, \lambda \,|\, \boldsymbol{\theta}^{[t]}\right) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}\left[\mathcal{L}\left(\boldsymbol{\theta}\,;\, \boldsymbol{z}^{[t]}, \boldsymbol{\xi}\right) - \lambda\, \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\theta}^{[t]}\right]$$

$$\textbf{M} : \begin{cases} \lambda^{[t+1]} \longleftarrow \mathsf{GCV}^{[t+1]}(\lambda) \\ \boldsymbol{\theta}^{[t+1]} \longleftarrow \underset{\boldsymbol{\theta}}{\arg\max}\; \mathcal{Q}_{\textbf{ridge}}\left(\boldsymbol{\theta}, \lambda^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right) \end{cases}$$

Update

Calculate $\boldsymbol{\xi}^{[t+1]}, \boldsymbol{z}^{[t+1]}, \boldsymbol{\Gamma}^{[t+1]}$ with the updated $\boldsymbol{\theta}^{[t+1]}$

Supervised component–based regularisation for GLMM–AR(1)

$$\boldsymbol{\theta} = \left(\boldsymbol{u}, \gamma, \sigma_1^2, \sigma_2^2, \rho\right)$$

Linearised model

$$\boldsymbol{\mathcal{M}}^{[t]} : \boldsymbol{z}^{[t]} = (\boldsymbol{X}\boldsymbol{u})\gamma + \boldsymbol{U_1}\boldsymbol{\xi}^{\boldsymbol{1}} + \boldsymbol{U_2}\boldsymbol{\xi}^{\boldsymbol{2}} + \boldsymbol{e}^{[t]}, \quad \text{with } \mathbb{V}\left(\boldsymbol{e}^{[t]}\right) = \boldsymbol{\Gamma}^{[t]}$$

SC- estimation

$$\mathbf{E} : \mathcal{Q}_{\mathsf{SC}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}\right) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}\left[(1-s)\mathcal{L}\left(\boldsymbol{\theta}\,;\,\boldsymbol{z}^{[t]}, \boldsymbol{\xi}\right) + s\phi(\boldsymbol{u})\,|\,\boldsymbol{\theta}^{[t]}\right]$$

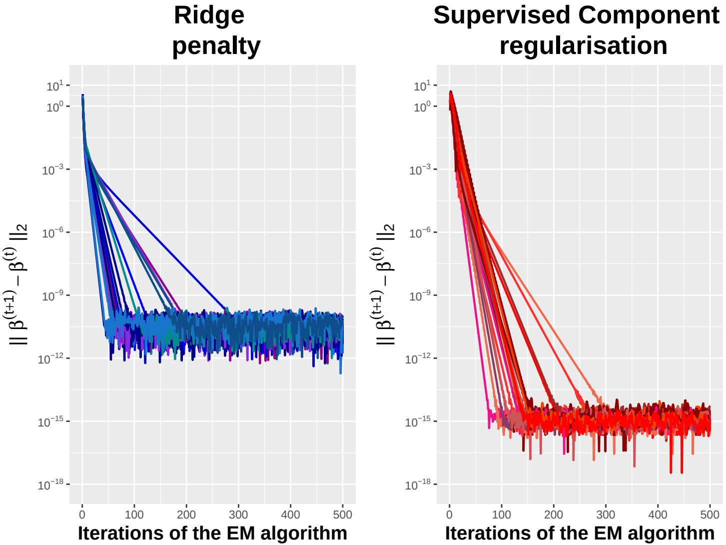
$$\mathbf{M} : \begin{cases} \sigma_1^{2\,[t+1]},\ \sigma_2^{2\,[t+1]},\ \rho^{[t+1]} \text{ computed as previously} \\ \boldsymbol{u}^{[t+1]} \longleftarrow \underset{\boldsymbol{u}:||\boldsymbol{u}||=1}{\arg\max}\ \mathcal{Q}_{\mathsf{SC}}\left(\boldsymbol{u}, \gamma^{[t]}\,|\,\boldsymbol{\theta}^{[t]}\right) \\ \gamma^{[t+1]} \longleftarrow \underset{\gamma}{\arg\max}\ \mathcal{Q}_{\mathsf{SC}}\left(\boldsymbol{u}^{[t+1]}, \gamma\,|\,\boldsymbol{\theta}^{[t]}\right) \end{cases}$$

Update: Calculate $\boldsymbol{\xi}^{[t+1]}$, $\boldsymbol{z}^{[t+1]}$, $\boldsymbol{\Gamma}^{[t+1]}$ with the updated $\boldsymbol{\theta}^{[t+1]}$

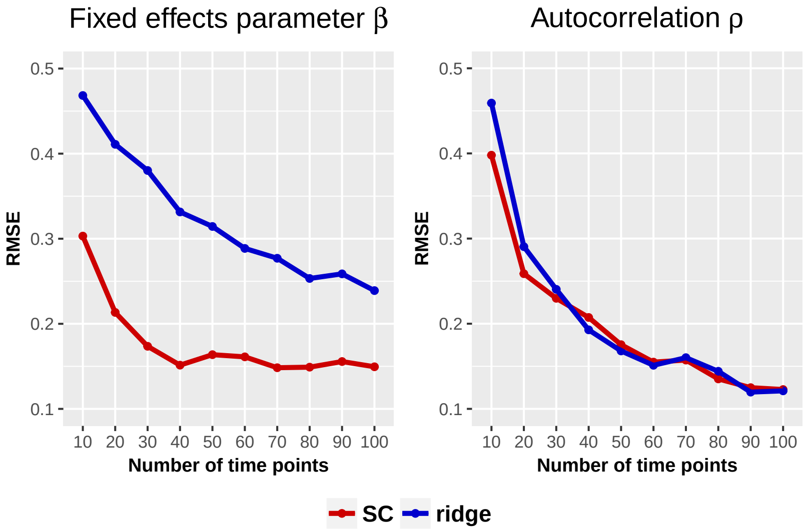
Poisson regression with log link

- $\boldsymbol{y} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi}^{\mathbf{1}} + \boldsymbol{U_2}\boldsymbol{\xi}^{\mathbf{2}}\right)\right)$
- $\boldsymbol{X} = \Big[\ \underbrace{\boldsymbol{x}^{\mathbf{1}} \ldots\ldots\ldots\ldots \boldsymbol{x}^{\mathbf{10}}}_{\substack{\text{large bundle} \\ \hookrightarrow \text{ noise}}}\ \underbrace{\boldsymbol{x}^{\mathbf{11}} \ldots\ldots \boldsymbol{x}^{\mathbf{15}}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{ predicts y}}}\ \underbrace{\boldsymbol{x}^{\mathbf{16}} \ldots\ldots \boldsymbol{x}^{\mathbf{20}}}_{\substack{\text{small bundle} \\ \hookrightarrow \text{ predicts y}}}\ \Big]$

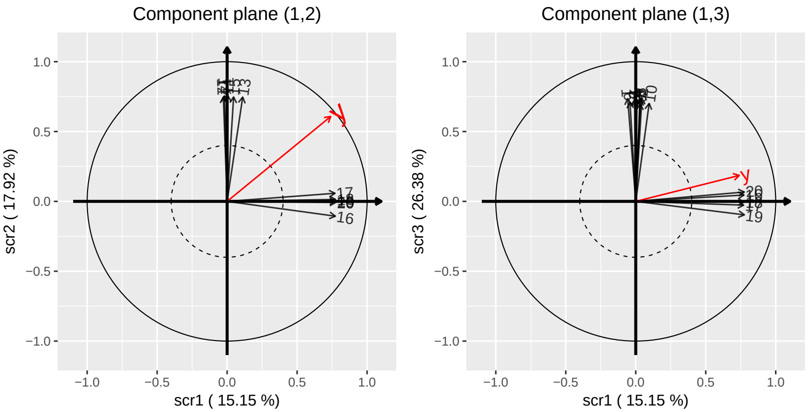
How does convergence go?

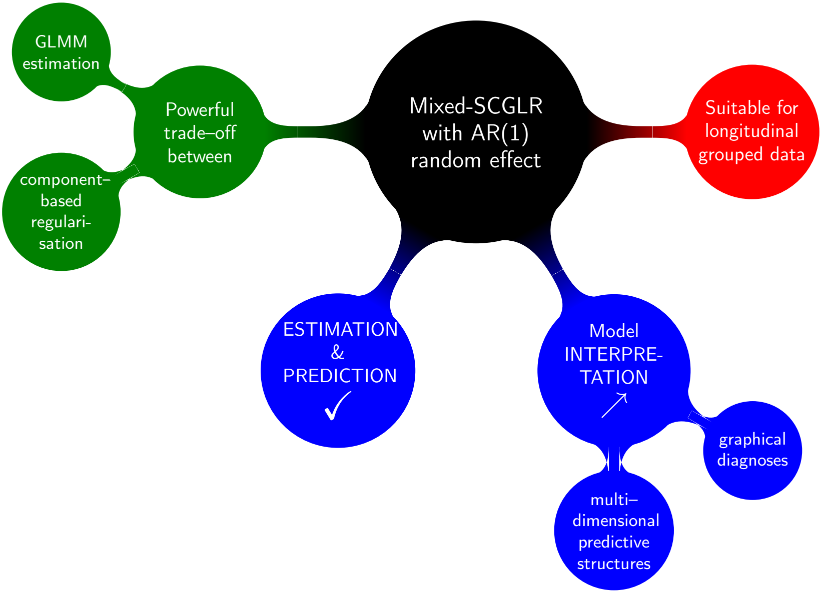


Accuracy of the estimates



Power for model interpretation




GLMM estimation
Powerful trade–off between
component–based regulari-sation
Mixed-SCGLR with AR(1) random effect
Suitable for longitudinal grouped data
ESTIMATION & PREDICTION ✓
Model INTERPRE-TATION ↗
graphical diagnoses
multi–dimensional predictive structures

Bibliography

Bry, X., Trottier, C., Verron, T. and Mortier, F. (2013) *Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm.* Journal of Multivariate Analysis, **119**, 47-–60.

+ Package R : SCGLR

Chauvet, J., Bry, X., Trottier, C. and Mortier, F. (2016) *Extension to mixed models of the Supervised Component-based Generalised Linear Regression.* In COMPSTAT: Proceedings in Computational Statistics.

Eliot, M., Ferguson, J., Reilly, M.P. and Foulkes, A.S. (2011) *Ridge Regression for Longitudinal Biomarker Data.* The International Journal of Biostatistics, **7**, 1–11.

Green, P.J. (1990) *On use of the EM for penalized likelihood estimation.* Journal of the Royal Statistical Society. Series B (Methodological), 443-452.

Marx, B. D. (1996) *Iteratively reweighted partial least squares estimation for generalized linear regression.* Technometrics, **38**, 4, 374–381.