Sparse group latent variables and penalized regression methods: a comparative study

Soufiane Ajana

Niazi Acar, Lionel Bretillon, Boris Hejblum, Hélène Jacqmin-Gadda, Cécile Delcourt

31 January 2018

BORDEAUX POPULATION HEALTH Research Center - U1219

université de BORDEAUX

LEHA / Lifelong exposures, health and aging

Summary

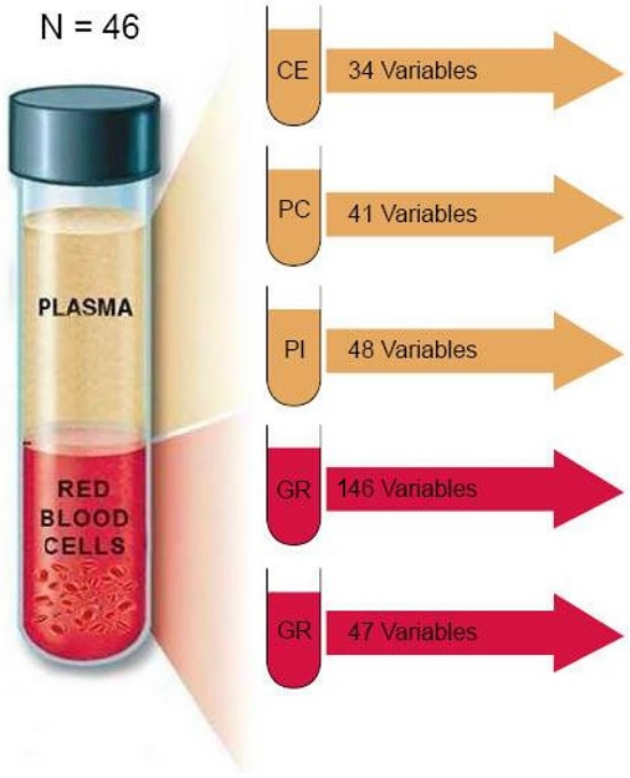
BLISAR study

BLISAR study

Objective : predict retinal n-3 PUFA status from the circulating biomarkers

- Samples of retina, plasma and red blood cells collected from human donors free of retinal diseases
- Lipid profile $\Rightarrow$ gas chromatography (GC)
- Structural analyses of red blood cells $\Rightarrow$ liquid chromatography mass spectrometry (LCMS)
- Circulating biomarkers of retinal n-3 PUFA status $\Rightarrow$ 5 groups :
  1. Cholesteryl Esters (CE)
  2. Phosphatidylcholines (PC)
  3. Total plasma (PI)
  4. Red blood cells (GR) assessed by GC
  5. Red blood cells (GR) assessed by LMCS

BLISAR study



Analytical method	
GC	LCMS
X	
X	
X	
	X
X	

High dimensional issues (p>n)

- Multicollinearity

$$\hat{\theta}^{OLS} = (X^\top X)^{-1} X^\top Y$$

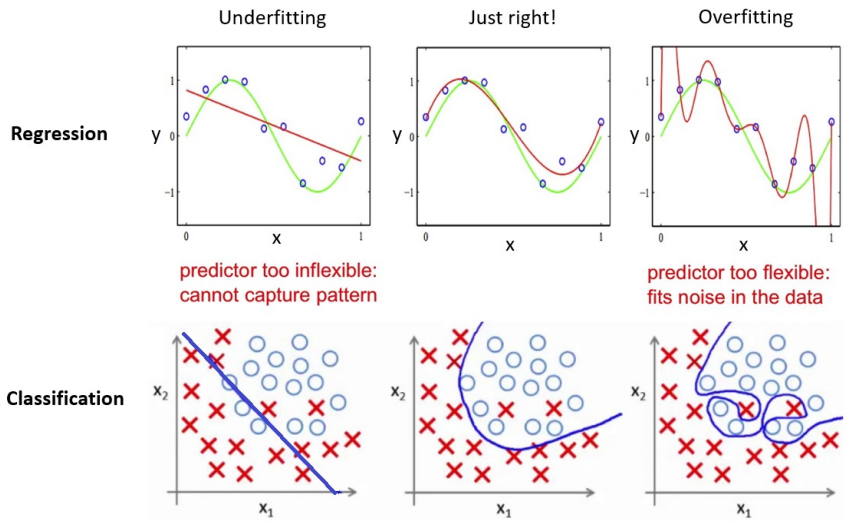
$\Rightarrow X^\top X$ is not invertible
$\Rightarrow$ Variance of $\hat{\theta}^{OLS}$ inflated

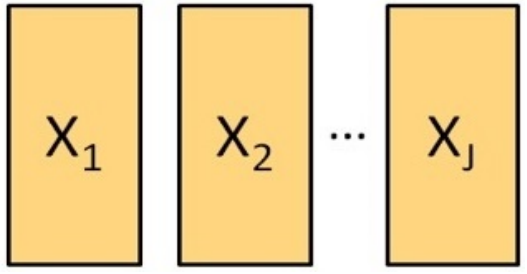
- Overfitting

$\Rightarrow$ Model complexity vs data ressources

- Data structure

$\Rightarrow$ Natural grouping effect
$\Rightarrow$ Select groups of variables





J : number of groups

Solution : Add constraints

Penalized regression methods

With group structure :

$$argmin_{\theta \in \mathbb{R}^p} \left\{ \| y - \sum_{l=1}^{L} X^{(l)} \theta^{(l)} \|_2^2 + \lambda \sum_{l=1}^{L} [(1-\alpha)\sqrt{p_l} \| \theta^{(l)} \|_2 + \alpha \| \theta \|_1] \right\}$$

Without group structure :

$$argmin_{\theta \in \mathbb{R}^p} \left\{ \| y - X\theta \|_2^2 + \lambda [(1-\alpha) \| \theta \|_2^2 + \alpha \| \theta \|_1] \right\}$$

- y : continuous outcome
- $X^{(l)}$: submatrix of X with columns corresponding to the predictors in group l
- $\theta^{(l)}$: coefficient vector for the group l
- $\sqrt{p_l}$: accounts for size of the group l
- λ : penalty parameter (tuning parameter)
- α : controls the combination between L1 and L2 norms (tuning parameter)

Penalized regression methods

With group structure :

$$argmin_{\theta \in \mathbb{R}^p} \left\{ \| y - \sum_{l=1}^{L} X^{(l)} \theta^{(l)} \|_2^2 + \lambda \sum_{l=1}^{L} [(1-\alpha)\sqrt{p_l} \| \theta^{(l)} \|_2 + \alpha \| \theta \|_1] \right\}$$

- $\alpha = 0 \Rightarrow$ Group Lasso (gLasso)
- $\alpha = 1 \Rightarrow$ Lasso
- $0 < \alpha < 1 \Rightarrow$ Sparse Group Lasso (sgLasso)

Penalized regression methods

Without group structure :

$$argmin_{\theta \in \mathbb{R}^p} \left\{ \| y - X\theta \|_2^2 + \lambda[(1-\alpha)\| \theta \|_2^2 + \alpha \| \theta \|_1] \right\}$$

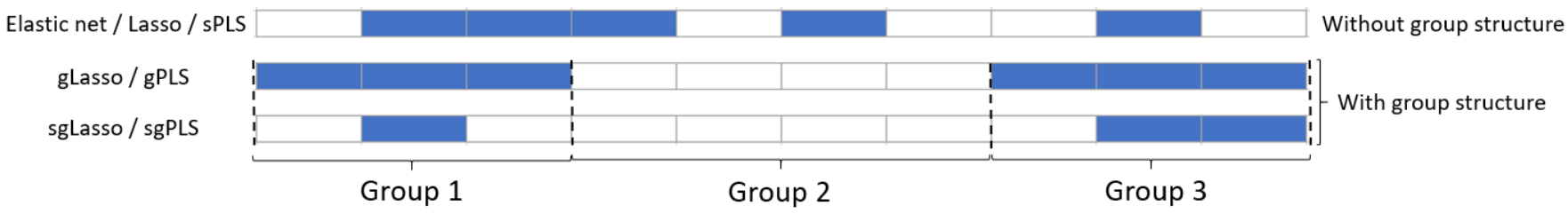
- $\alpha = 0 \Rightarrow$ Ridge
- $\alpha = 1 \Rightarrow$ Lasso
- $0 < \alpha < 1 \Rightarrow$ Elastic net

Latent variable methods

$$argmax_{u \in \mathbb{R}^p, \|u\|_2=1} \left\{ cov(X^{(l)} u^{(l)}, y) - \lambda[(1-\alpha) \sum_{l=1}^{L} \sqrt{p^{(l)}} \| u^{(l)} \|_2 + \alpha \| u \|_1] \right\}$$

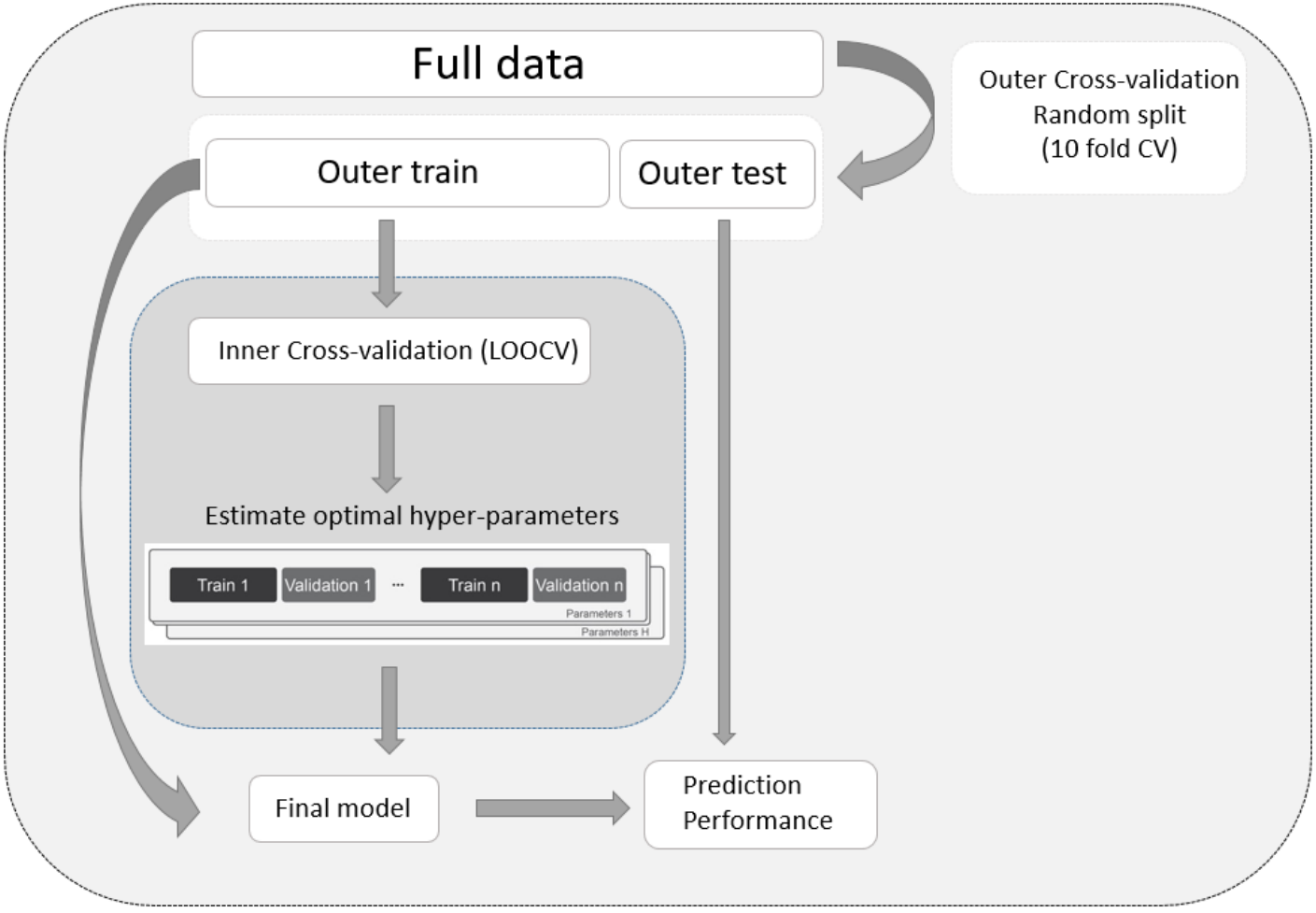
- y : continuous outcome
- $X^{(l)}$: submatrix of X with columns corresponding to the predictors in group l
- $u^{(l)}$: loading vector for the group l
- $\sqrt{p^{(l)}}$: accounts for size of the group l
- λ : penalty parameter (tuning parameter)
- α : controls the combination between L1 and L2 norms (tuning parameter)

Summary



Model	High correlation	Group structure
Lasso		
Group Lasso (gLasso)		X
Sparse Group Lasso (sgLasso)		X
Elastic net	X	
Sparse Partial Least Squares (sPLS)	X	
Groupe Partial Least Squares (gPLS)	X	X
Sparse Groupe Partial Least Squares (sgPLS)	X	X

Repeated double cross-validation



Prediction indices

Mean Squared Error of Prediction (MSEP) :

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

- MSEP $= 0 \Rightarrow$ Perfect prediction
- MSEP $\nearrow \Rightarrow$ Prediction performance $\searrow$

Prediction indices

Goodness of fit (R^2) :

$$R^2 = cor(y_{test}, \hat{y}_{test})^2$$

- $R^2 = 1 \Rightarrow$ Perfect prediction
- $R^2 \searrow \Rightarrow$ Prediction performance $\searrow$

Important note : R^2 calculated via cross-validation on the test data, assesses the quality of predictions on independent (unseen) sets

Results

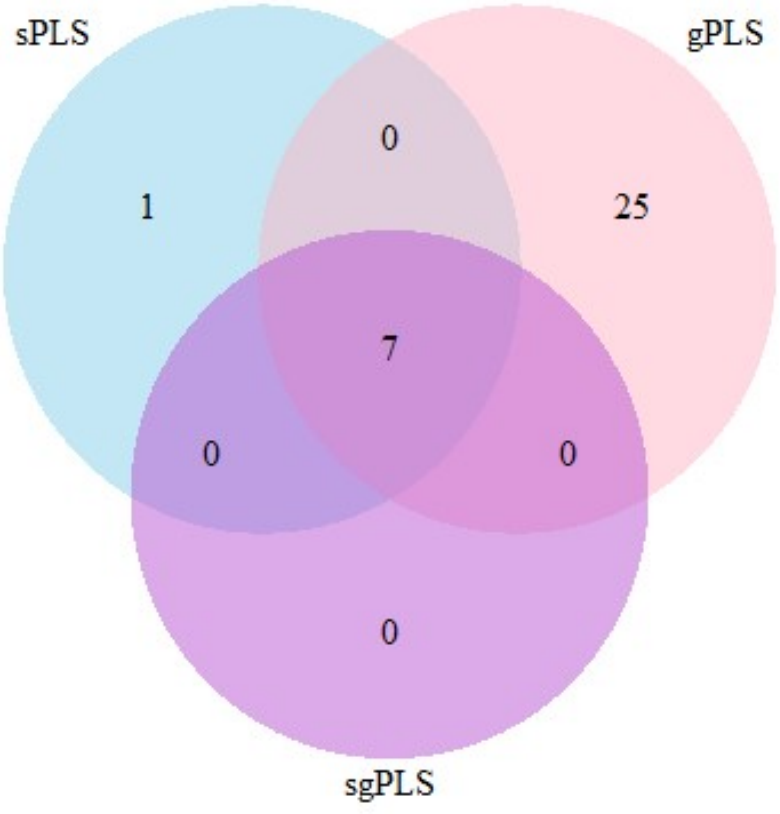
Comparison of the multivariable regression methods for 10 random divisions with 100 runs (N=46, p=332)

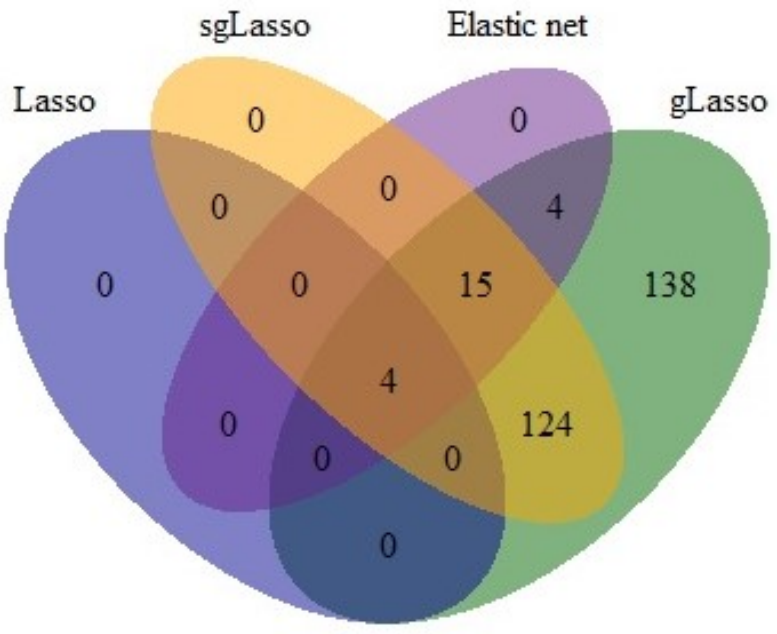
Method	Test data R^2 (SD)	Test data RMSEP (SD)	Number of selected variables*	Selected groups*
Lasso	0.14 (0.05)	2.73 (0.14)	4	CE, PC
sgLasso	0.20 (0.05)	2.72 (0.16)	143	CE, PC, LCMS
gLasso	0.21 (0.05)	2.69 (0.15)	285	CE, PC, PI, LCMS
Elastic net	0.18 (0.05)	2.65 (0.12)	23	CE, PC, PI, LCMS
SPLS	0.36 (0.03)	2.32 (0.05)	8	CE, PI
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

Results

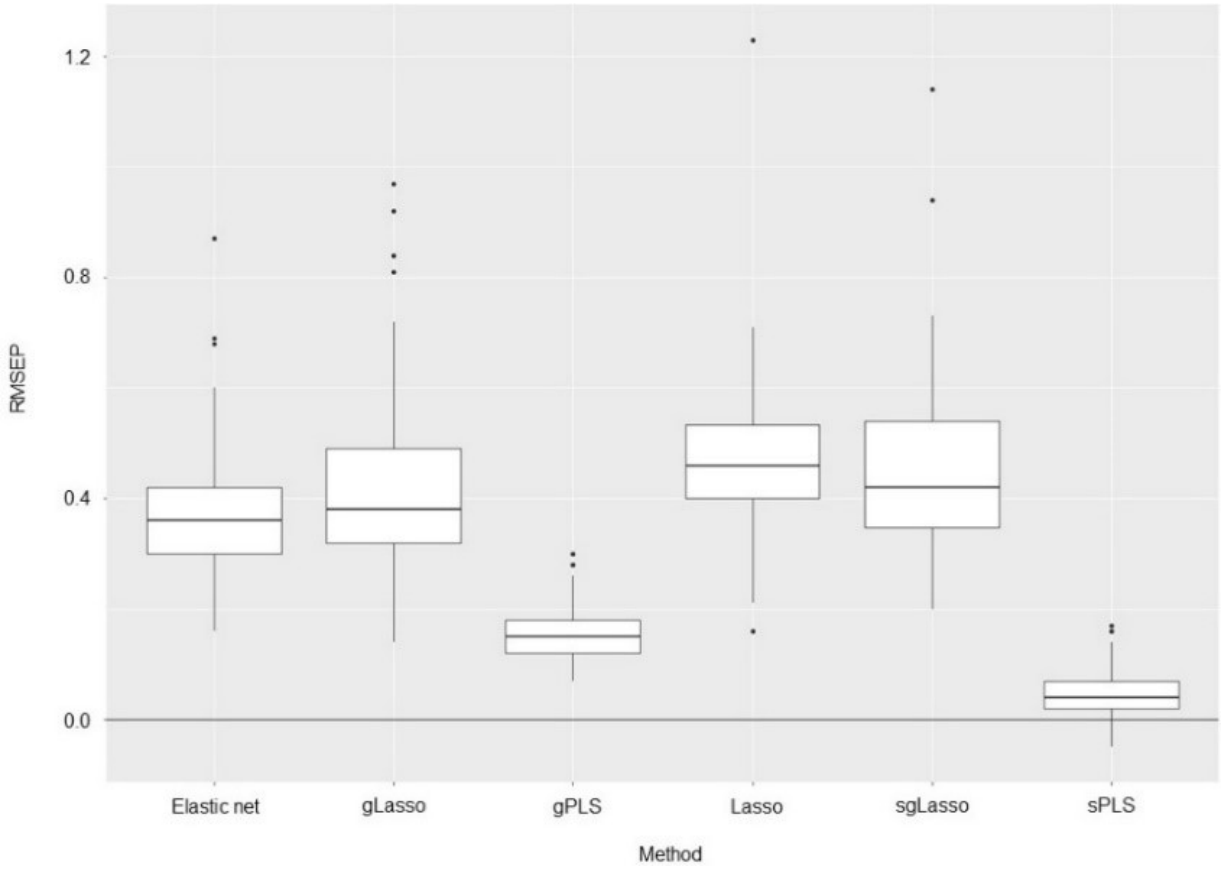
Venn diagrams of the variables selected by latent variable and penalized regression methods





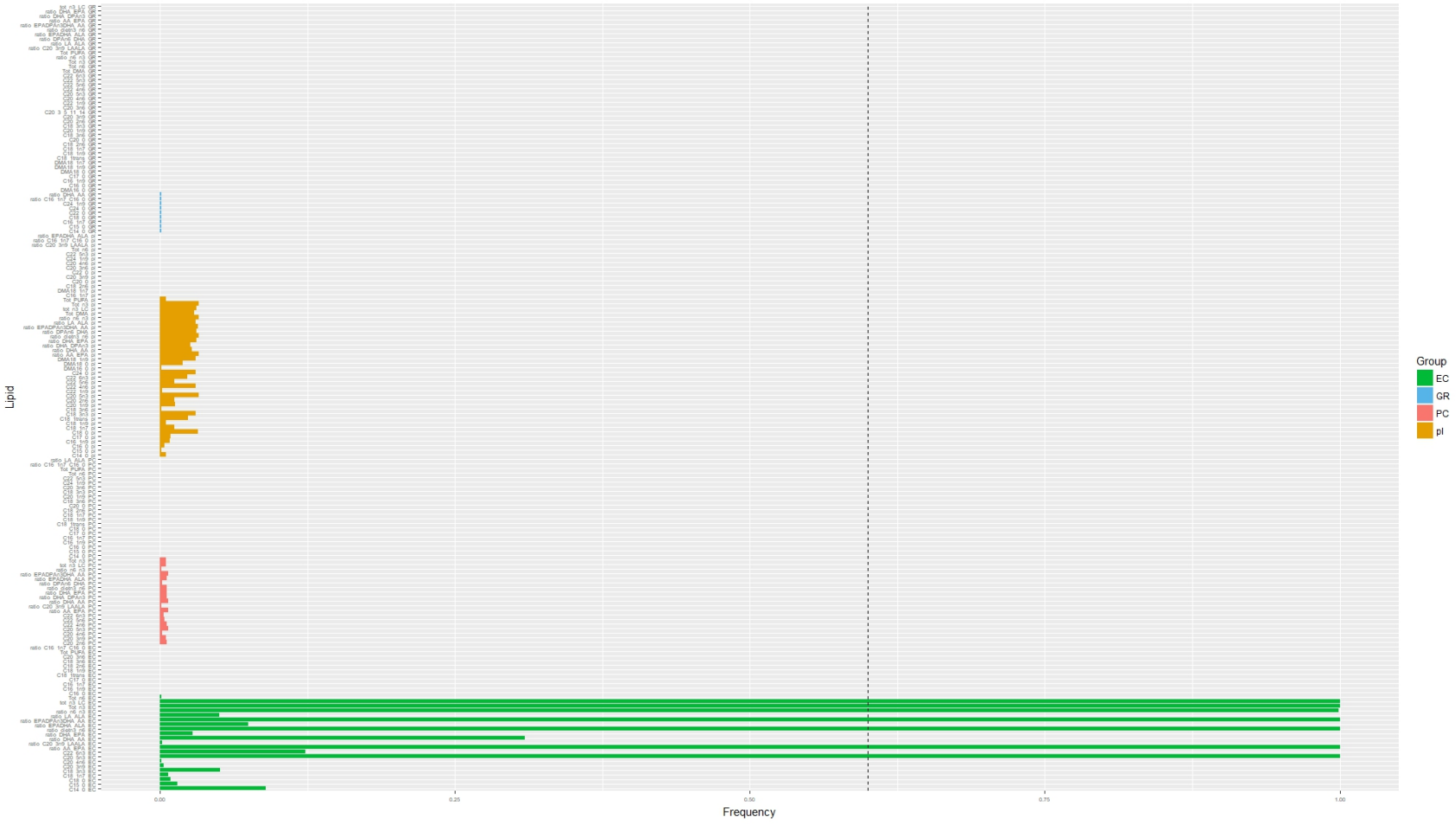
Results

Boxplots of the difference in RMSEP between **sgPLS** and the other six methods over the 100 runs



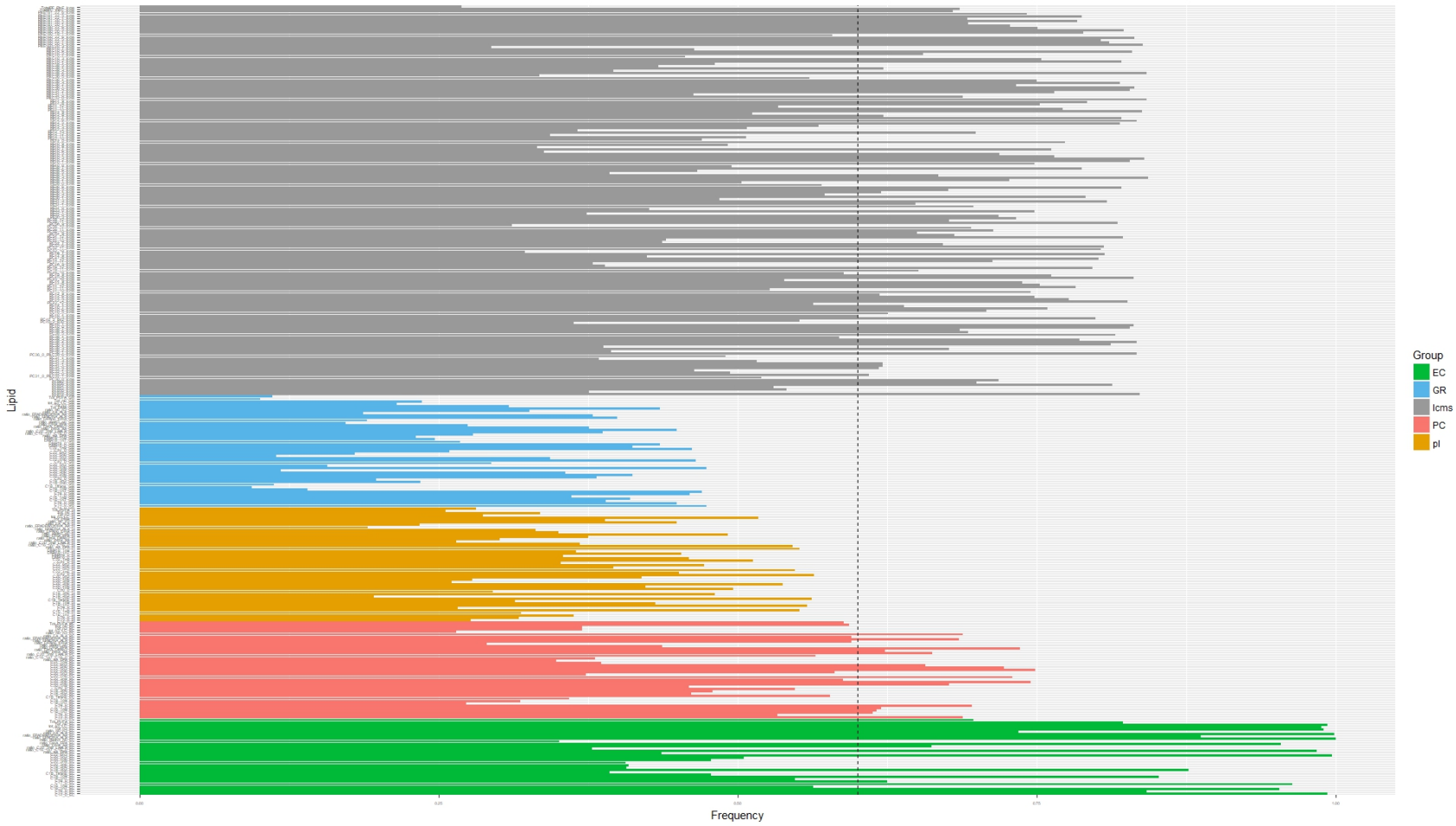
Results

Lipids frequency selection with **sgPLS** over 100 runs. The vertical dashed line correspond to a frequency selection of 60 %



Results

Lipids frequency selection with **sgLasso** over 100 runs. The vertical dashed line correspond to a frequency selection of 60 %



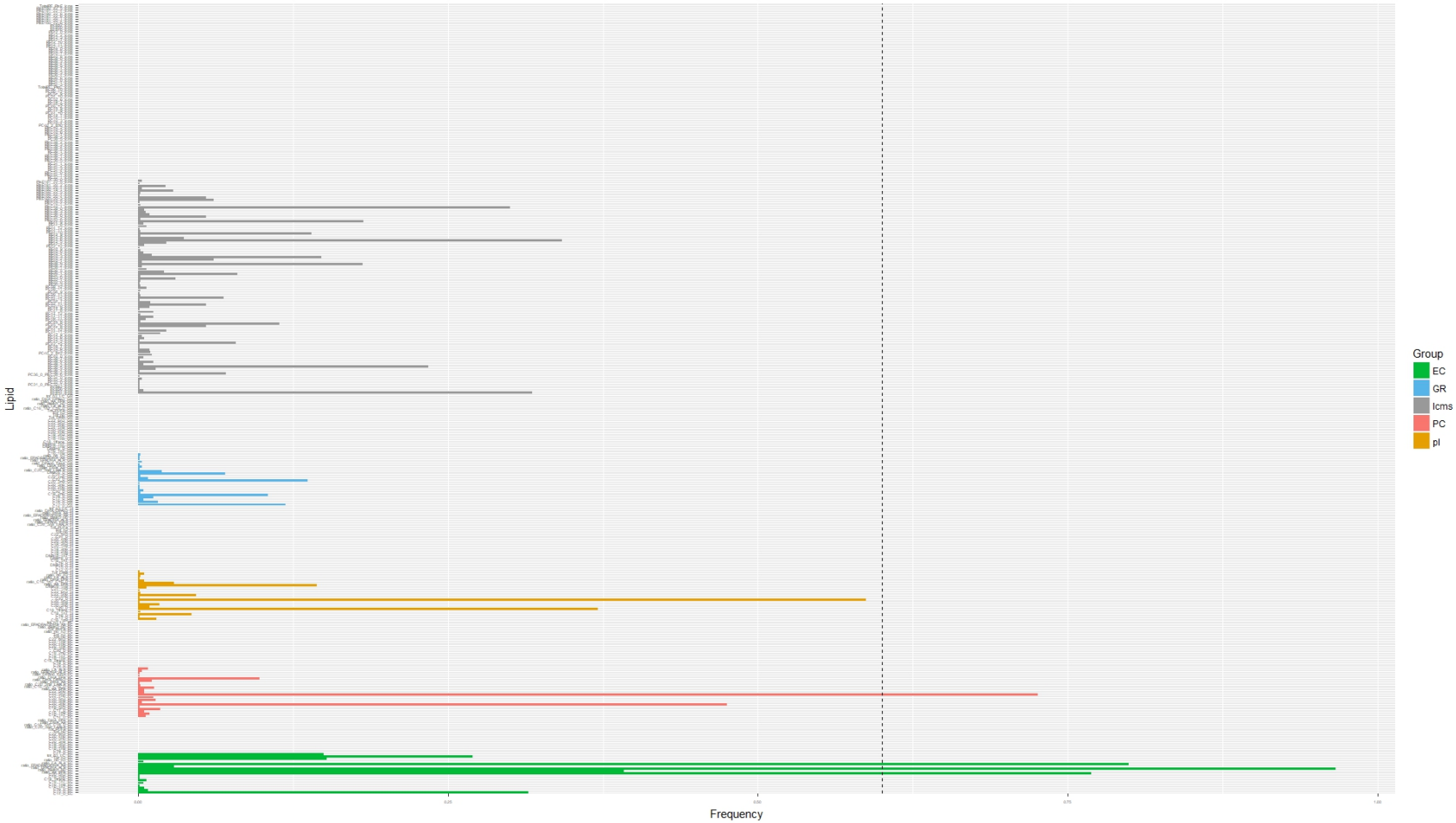
Discussion

- Prediction performance $\Rightarrow$ sPLS and sgPLS
- Variable selection $\Rightarrow$ sgPLS (only one group selected)

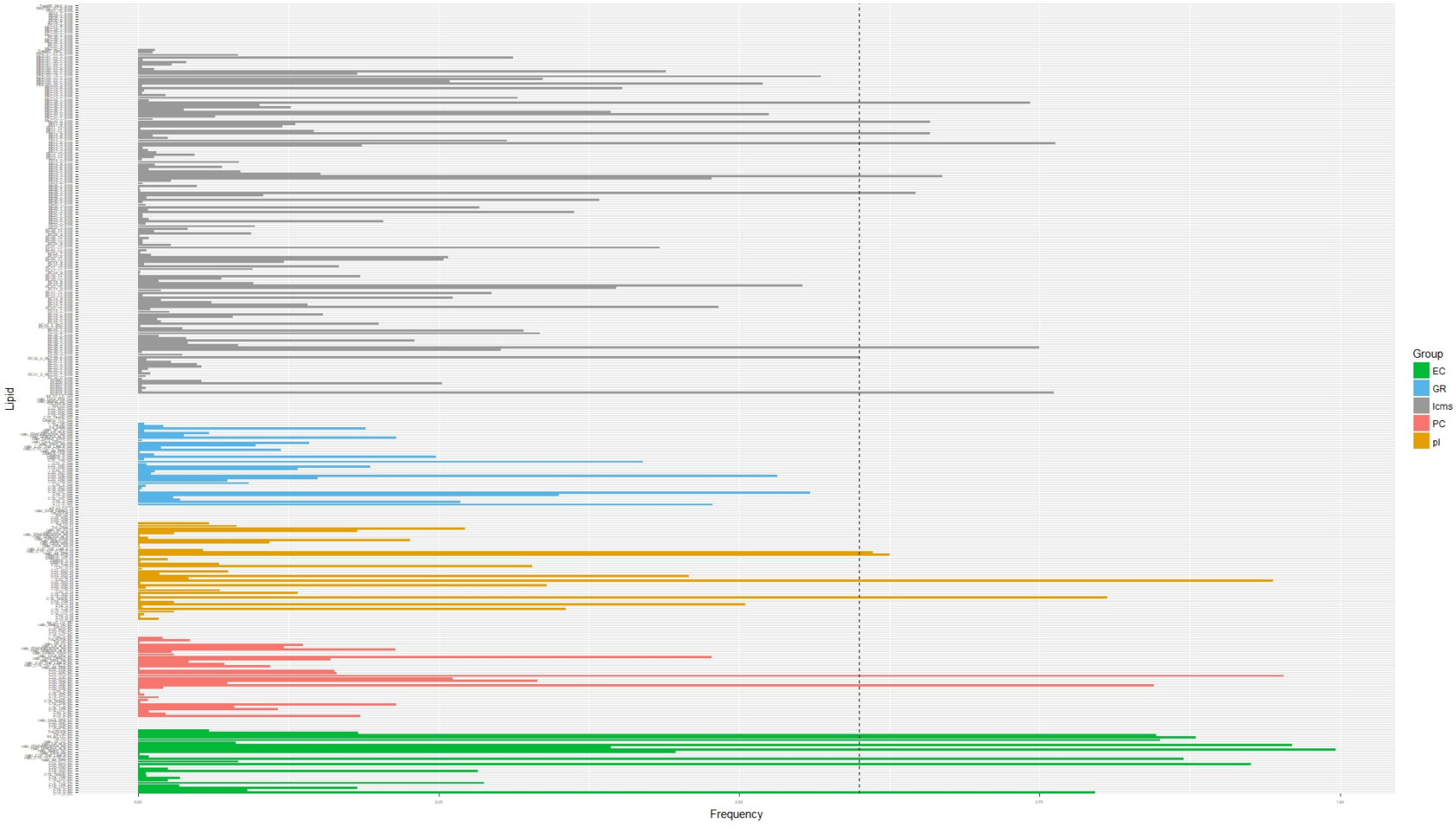
  $\Rightarrow$ sparse data with some relevant predictors
  $\Rightarrow$ biologically relevant predictors
  $\Rightarrow$ lower costs

- Generally, Latent variables methods outperformed penalized regression methods in this application
- Small sample size + High correlations + Group structure $\Rightarrow$ sgPLS +++

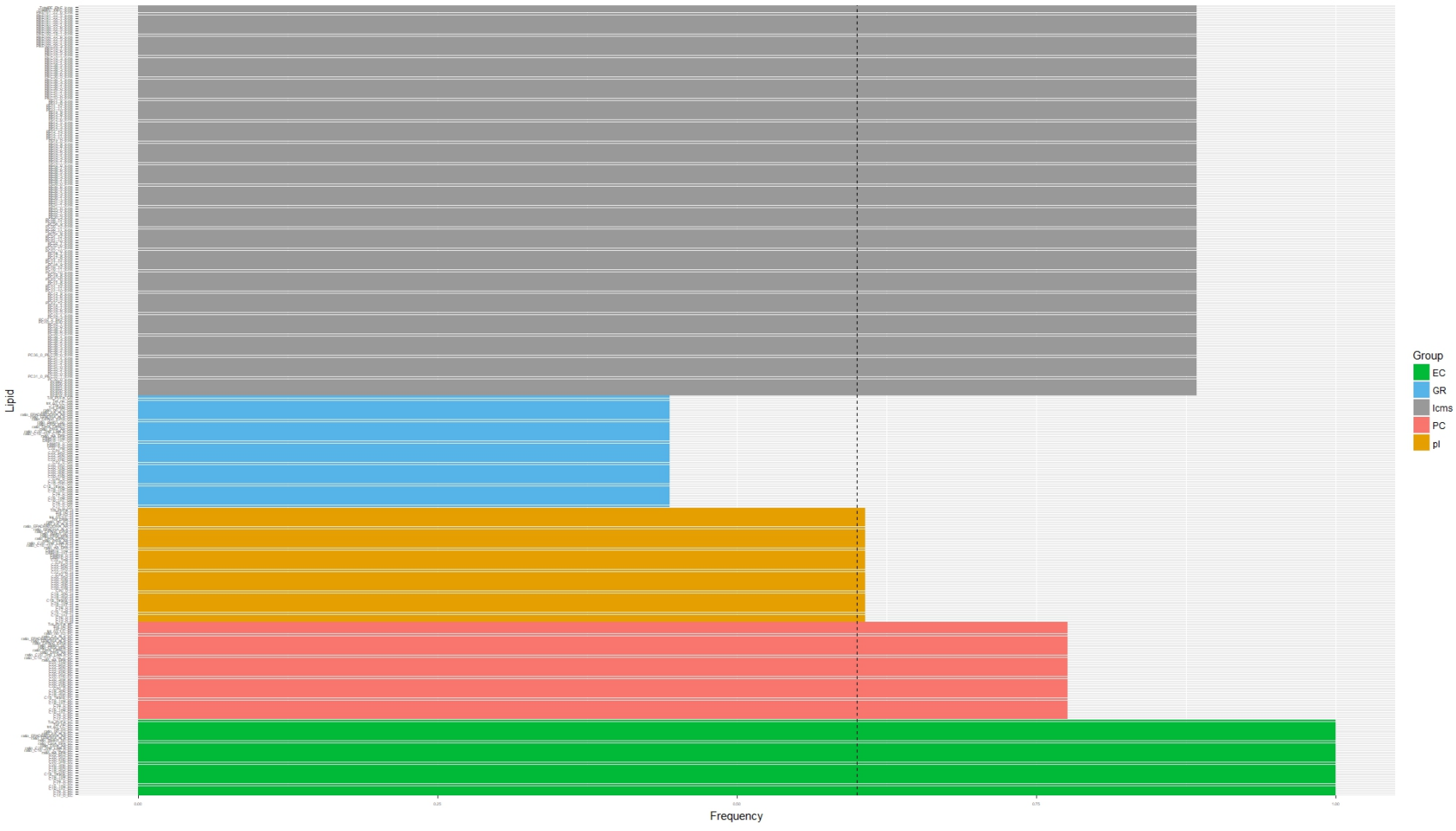
Lasso



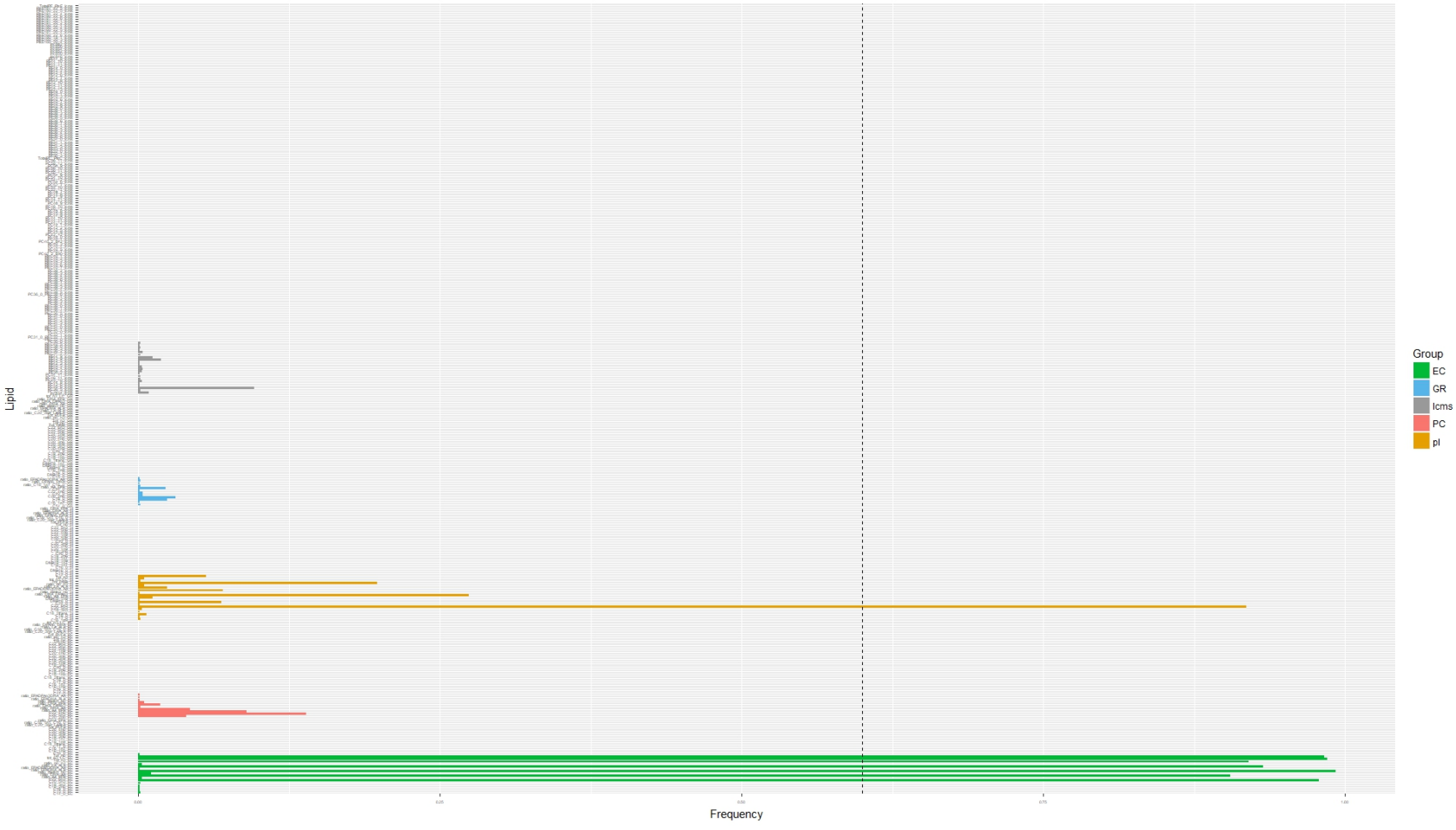
Elastic net



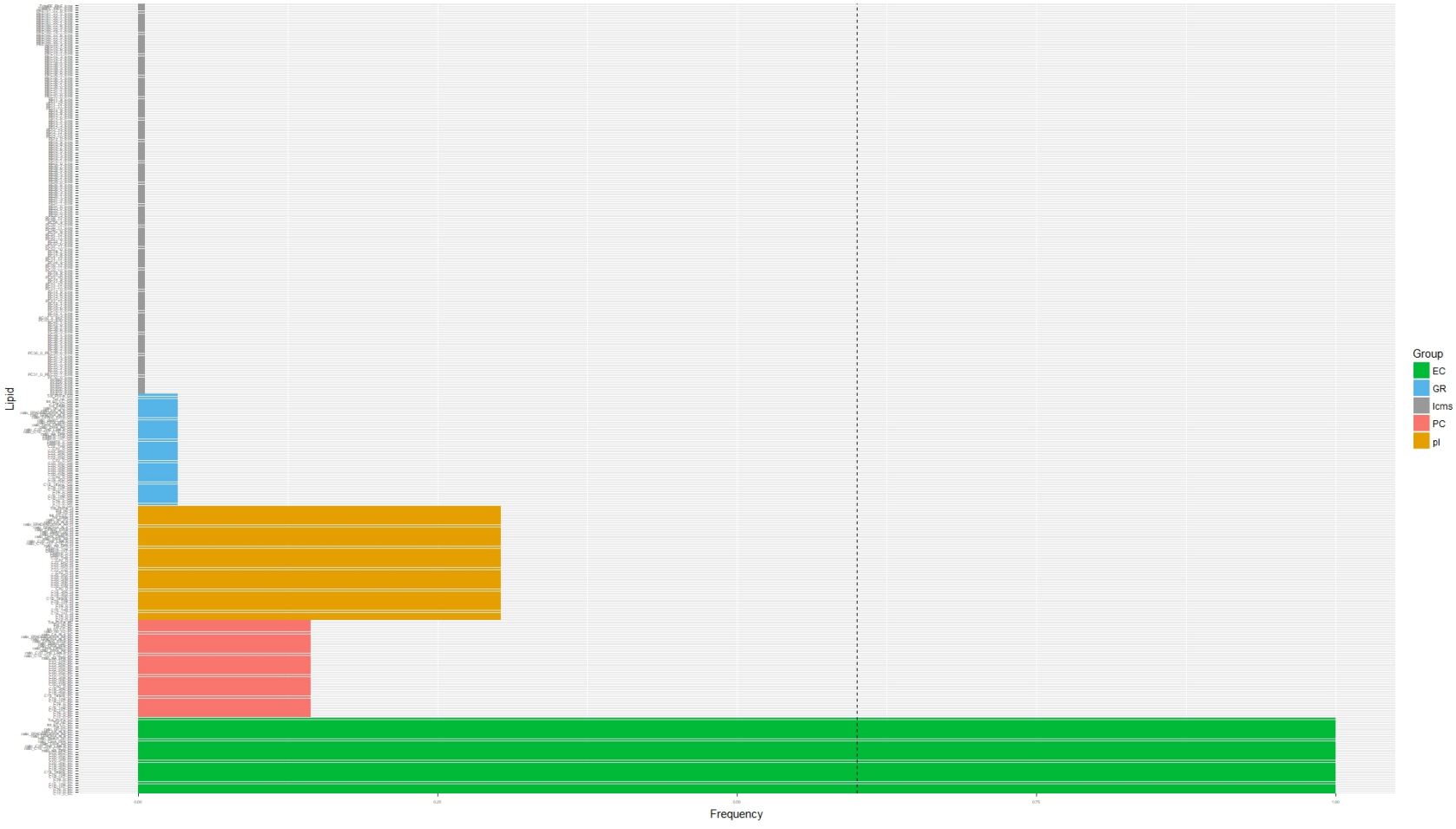
gLasso



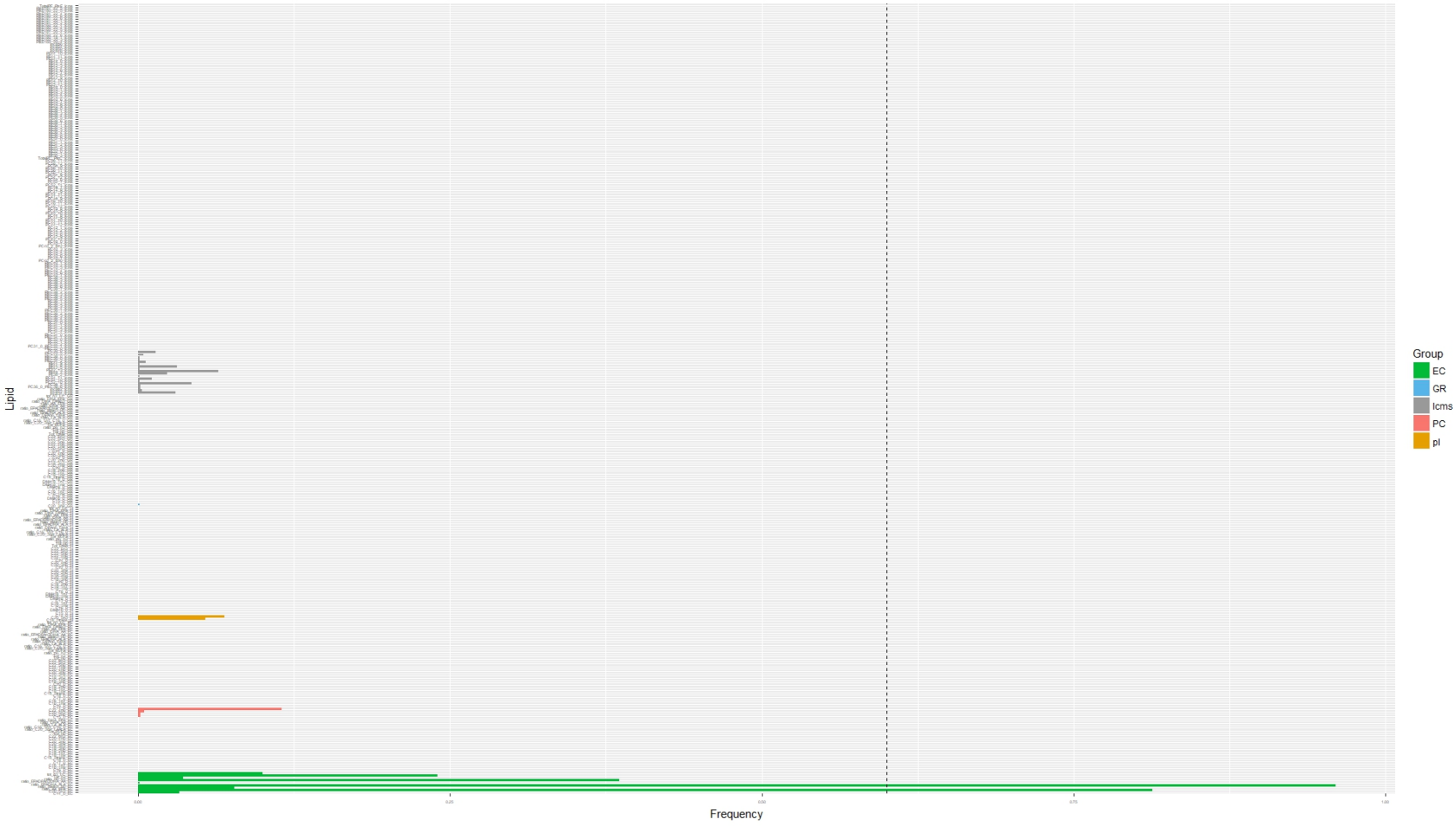
sPLS



gPLS



AdaLasso



AdagLasso