Generalization of the Lasso Penalty:

Example of the Group Lasso

POLL Luca & BENAMI Saloua

Toulouse School of Economics

December 3, 2021

Outline

Introduction - Shortcomings of Lasso

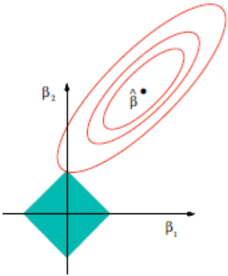


Figure: Lasso estimation graphically
Source: Hastie et al. (2015)

Lasso minimization problem:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Problem: natural group structure of covariates:

1. **Multicollinearity**: if matrix of regressors is not full rank, Lasso does not have a unique solution
   - $\Rightarrow$ Lasso might empirically perform poorly with correlated variables
2. **Factor variables**: desirable to shrink all dummies belonging to a factor variable simultaneously to zero

Group Lasso

- New model that takes into account the correlation structure, proposed by Yuan and Lin (2006) : the Group Lasso
- Prominent for factor variables, where levels are represented in set of dummies
- Group Lasso solves the convex problem:

$$\min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^{P_j}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T \theta_j \right)^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\} \quad (1)$$

Where $j = 1, ..., J$ represents the J groups of variables, $Z_j \in \mathbb{R}^{p_j}$ the covariates in group j, and $\theta_j \in \mathbb{R}^{p_j}$ the p_j regression coefficients of group j.
$\|\theta_j\|_2$ represents the Euclidean norm.

Group Lasso

- $\|\theta_j\|_2 = |\theta_j|$ for $p_j = 1$: if all groups are singletons, the optimization problem reduces to the ordinary Lasso
- If only one group, optimization problem is the Ridge regression
- The corresponding FOC is given as:

$$-\mathrm{Z}_j^T \left(\mathrm{y} - \sum_{\ell=1}^{J} \mathrm{Z}_\ell \widehat{\theta}_\ell \right) + \lambda \widehat{s}_j = 0, \qquad \text{for } j = 1, \cdots J,$$

  where $\widehat{s}_j \in \mathbb{R}^{p_j}$ is an element of the subdifferential of the norm $|| \cdot ||_2$
- Block coordinate descent : requires that Z_j is orthonormal (ORT)
- Groupwise-majorization-descent : relaxes the ORT

Group Lasso

- If all groups are equally penalized : larger groups more likely to be selected
- Yuan and Lin (2006) recommended weighting the penalties for each group according to their size:

$$\min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^{P_j}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T \theta_j \right)^2 + \lambda \sum_{j=1}^{J} m_j \left\| \theta_j \right\|_2 \right\} \tag{2}$$

  Where $m_j = \sqrt{p_j}$.

- For simplicity, the weights are omitted

Group Lasso

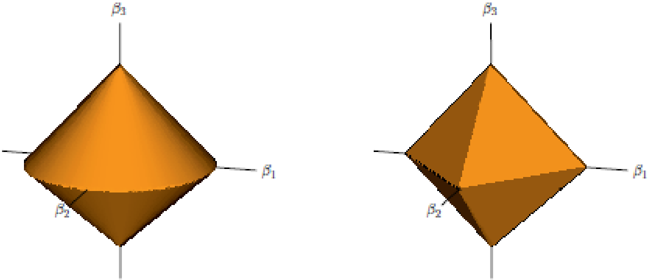


Figure: Group Lasso ball (left panel) versus Lasso ball (right panel)

- Three variables and two groups: $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$

Shortcomings & Alternatives

- Group Lasso allows to include or exclude whole groups instead of individual coefficients
- Groups : factors or highly correlated variables

But some drawbacks to note:

- Statistical performance that relies on the group structure
- Group Lasso can be inferior for groups with small sizes
- If groups are too large, the resulting model might still not be sparse enough
- If groups overlap, group Lasso performs poorly

Shortcomings & Alternatives

- **Sparse group Lasso**: allows sparsity across groups that are selected as well as which coefficients are nonzero within a group (through additional L1 norm)

$$\min_{(\theta_1,\ldots,\theta_J)} \left\{ \frac{1}{2} \left\| y - \sum_{j=1}^{J} Z_f \theta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} [(1-\alpha) \|\theta_j\|_2 + \alpha \|\theta_j\|_1] \right\} \tag{3}$$

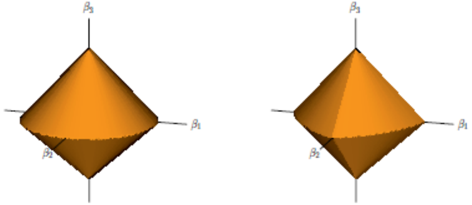


Figure: Group Lasso ball (left panel) vs Sparse Group Lasso ball (right panel)

Shortcomings & Alternatives

- **Overlap group Lasso**: variables can contribute to more than one group

$$\min_{\nu_j \in \mathcal{V}_{j,j=1,\ldots,J}} \left\{ \frac{1}{2} \left\| y - X \left(\sum_{j=1}^{J} \nu_j \right) \right\|_2^2 + \lambda \sum_{j=1}^{J} \|\nu_j\|_2 \right\} \qquad (4)$$

With $\beta = \sum_{j=1}^{J} \nu_j$ $\nu_j \in \mathbb{R}^p$ is a vector of all zeros except in positions corresponding to members of group j.

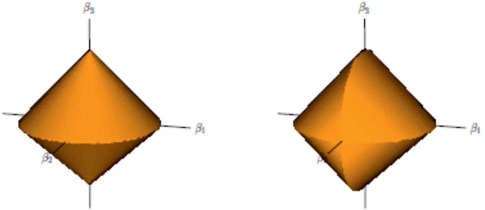


Figure: Group Lasso ball (left panel) vs Overlap Group Lasso ball (right panel)

Simulation

Simulation

Creating the data

Factor variable dataset

- J factor variables with $Z_j \in \mathbb{R}^{p_j}$ factor levels
- Coefficients $\theta_j \in \mathbb{R}^{p_j}$ drawn from a uniform distribution
- Group sparsity: $(1 - \text{sparsity}) \times J$ groups relevant
- Within sparsity: Relevant coefficients trough binomial distribution
- $Y = \sum_{j=1}^{J} Z_j' \theta_j + 0.1 \times \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$

Simulation

Creating the data

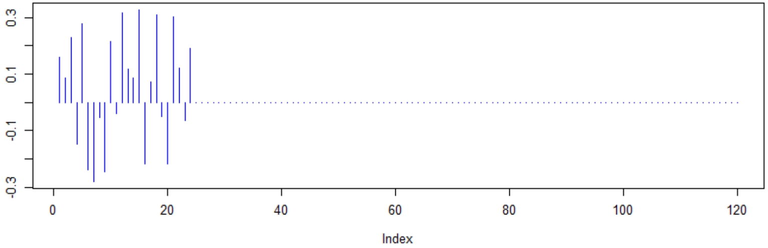
Correlated variable dataset

- J groups of $Z_j \in \mathbb{R}^{p_j}$ correlated covariates
- $p_1 = \gamma_1, \ldots, p_j = p_{j-1} + c \times \gamma_j$ where $\gamma_j \sim \mathcal{N}(0,\, 1)$
- Coefficients $\theta_j \in \mathbb{R}^{p_j}$ drawn from a uniform distribution
- Group sparsity: $(1 - \text{sparsity}) \times J$ groups relevant
- Within sparsity: Relevant coefficients trough binomial distribution
- $Y = \sum_{j=1}^{J} Z_j' \theta_j + 0.1 \times \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,\, 1)$

Simulation

Dataset with 120 covariates out of which 24 are nonzero

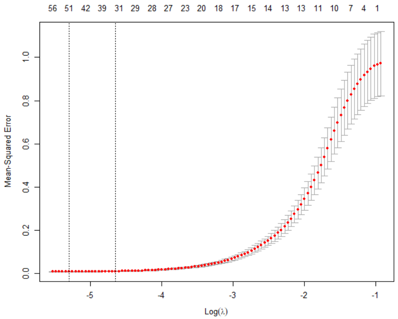
```r
data <- factor_sampler(levelsnr = 6, nrgroups = 20,
                       samplesize = 100, grsparsity = 0.8,
                       withinsparsity = 0)
plot(data$coefs_vec, col = "blue", type = "h", ylab = "")
```

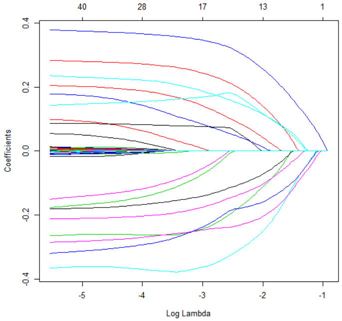


Simulation

Estimation via Lasso

```r
lasso <- cv.glmnet(x=data$x, y=data$y, family="gaussian")
plot(lasso)
plot(lasso$glmnet.fit, xvar="lambda")
```

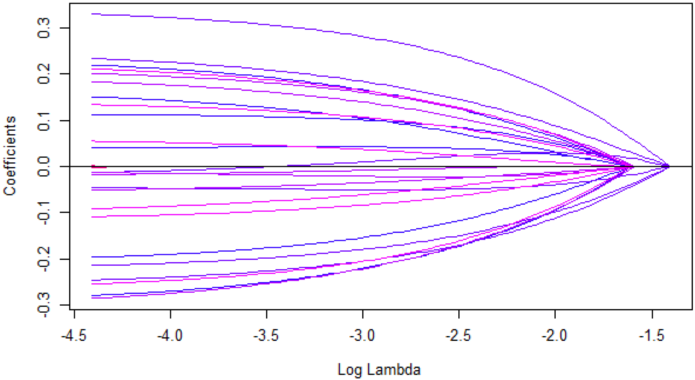




Simulation

Estimation via Group Lasso

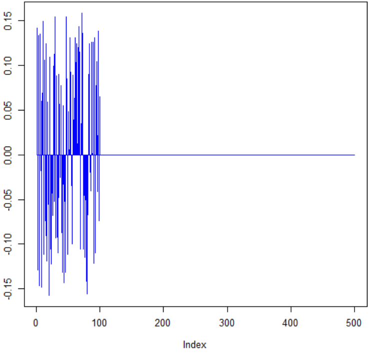
```
gglasso <- cv.gglasso(x=data$x, y=data$y,
                      group=data$group_structure)
plot(gglasso$gglasso.fit)
```

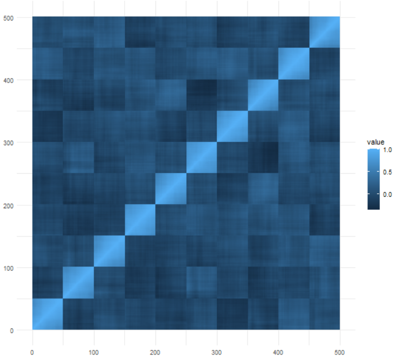


Simulation

Dataset with 500 covariates out of which 100 are nonzero (2 groups)

```
data <- correlated_sampler(levelsnr = 50, nrgroups = 10,
                           samplesize = 100, grsparsity=0.8,
                           withinsparsity = 0, corr = 0.2)
plot(data$coefs_vec, col = "blue", type = "h", ylab = "")
data$corrplot + xlab("") + ylab("")
```

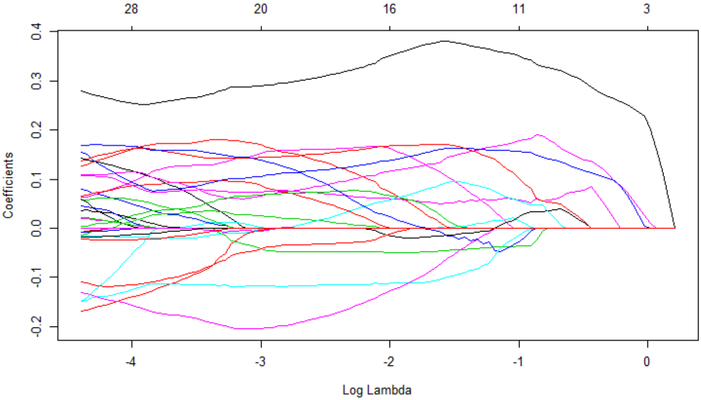




Simulation

Estimation via Lasso

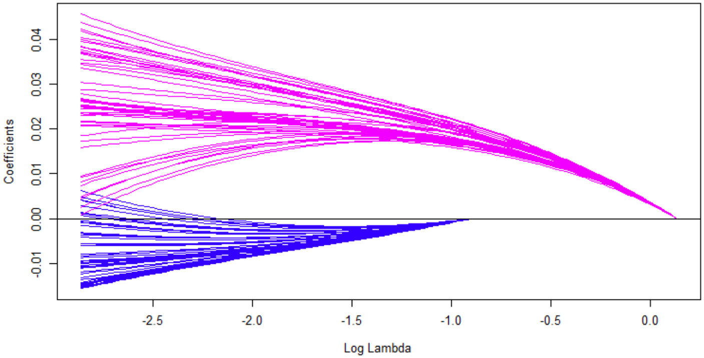
```r
lasso <- cv.glmnet(x=data$x, y=data$y, family="gaussian")
plot(lasso$glmnet.fit, xvar="lambda")
```



Simulation

Estimation via Group Lasso

```
gglasso <- cv.gglasso(x=data$x, y=data$y,
                      group=data$group_structure)
plot(gglasso$gglasso.fit)
```

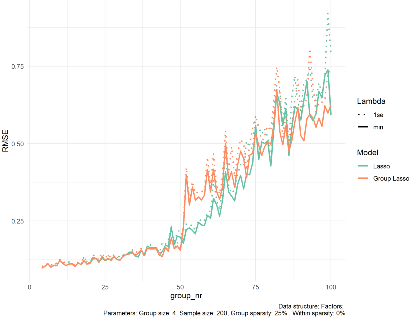


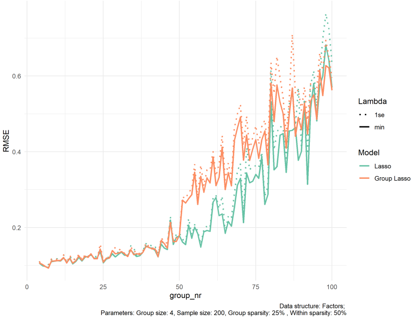
Simulation

Factor variable simulation - RMSE evaluation

- Iterate over Number of groups, group size, group sparsity and within group sparsity

$\Rightarrow$ Group Lasso performs poor in small groups

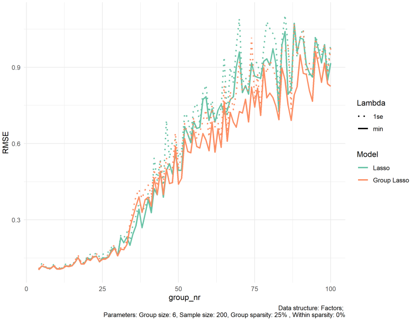




Simulation

Factor variable simulation - RMSE evaluation

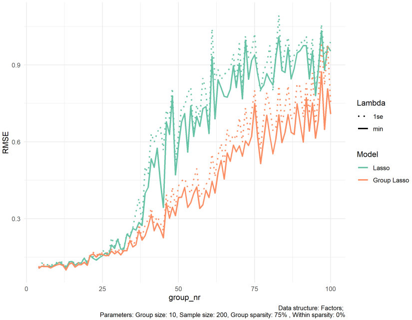
⇒ Changing pattern if group size increases

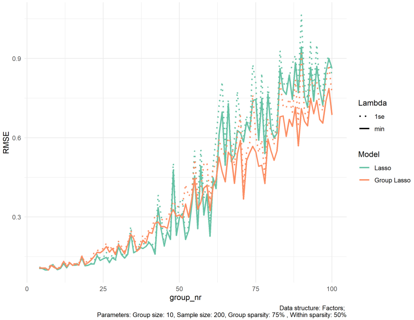


Simulation

Factor variable simulation - RMSE evaluation

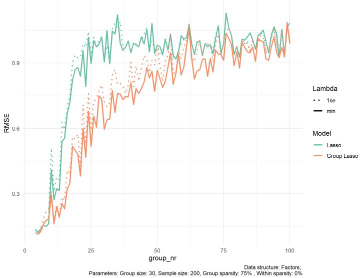
⇒ Group Lasso superior, within sparsity relevant

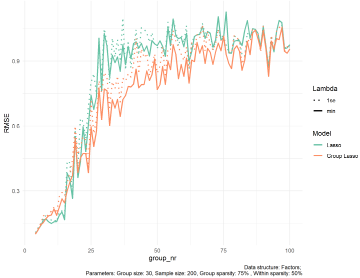


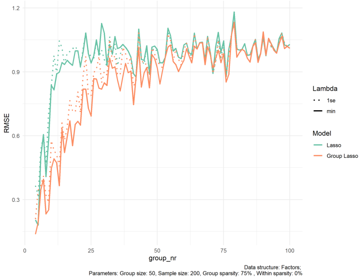


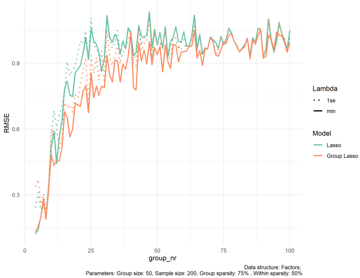
Simulation

Factor variable simulation - RMSE evaluation







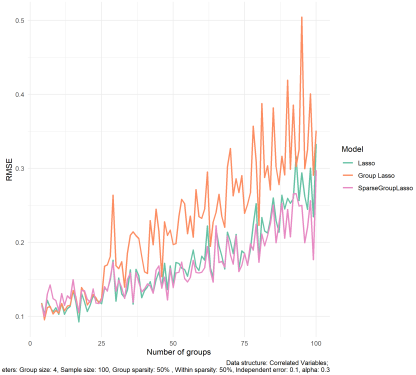


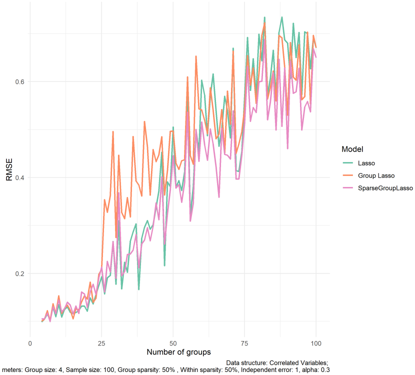
Simulation

Correlated variables simulation - RMSE evaluation

- Iterate over Number of groups, group size, group sparsity, within group sparsity + wight for correlation α for the sparse group lasso

$\Rightarrow$ Group Lasso again bad performance

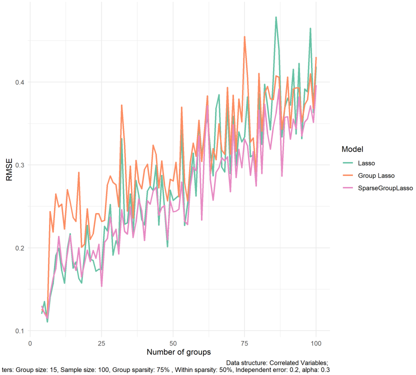


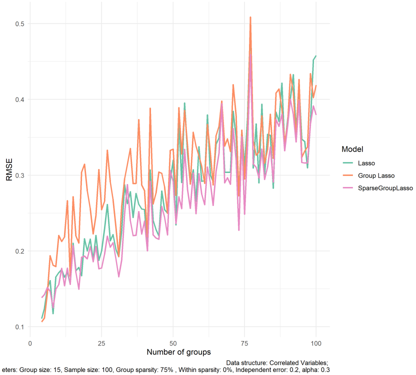


Simulation

Correlated variables simulation - RMSE evaluation

$\Rightarrow$ For fixed correlation level and α pattern the same across group size and group sparsity:





Simulation

Difficulties and challenges

- Evaluation of correlated variables with sparse group lasso challenging
- Parameters to vary:
  - Correlation of variables within group
  - Group sparsity
  - Within group sparsity
  - Group size
  - Number of groups
  - α parameter for Sparse Group Lasso
  - Sample size

Thank you for your attention!

Link to the Bitbucket Repository