Data shared lasso: a simple L1 regularization method for subgroup analyses

Vivian Viallon
IARC

Joint work with Edouard Ollier, Cedric Garcia and Nadim Ballout

International Agency for Research on Cancer
World Health Organization

Study of association in epidemiology/ clinical research

- General setting
  - Data: Y, $\mathbf{x} \in \mathbb{R}^p$
  - Objective: estimate $\boldsymbol{\beta}^*$, e.g. under a logistic model: $\text{logit}[\mathbb{P}(Y = 1)] = \mathbf{x}^T \boldsymbol{\beta}^*$.

Study of association in epidemiology/ clinical research

- General setting
  - Data: Y, $\mathbf{x} \in \mathbb{R}^p$
  - Objective: estimate $\boldsymbol{\beta}^*$, e.g. under a logistic model: $\text{logit}[\mathbb{P}(Y = 1)] = \mathbf{x}^T \boldsymbol{\beta}^*$.

- Subgroup analyses
  - The overall population = K predefined groups (or strata)
    - groups based on "additional" covariates (e.g., gender, age categories)
    - groups based on the outcome (e.g., disease subtypes)
  - Objective: estimate $(\boldsymbol{\beta}_k^*)_{k=1,\ldots,K}$: association between $\mathbf{x}$ and Y in group k

    $\Rightarrow$ to identify heterogeneities across the groups

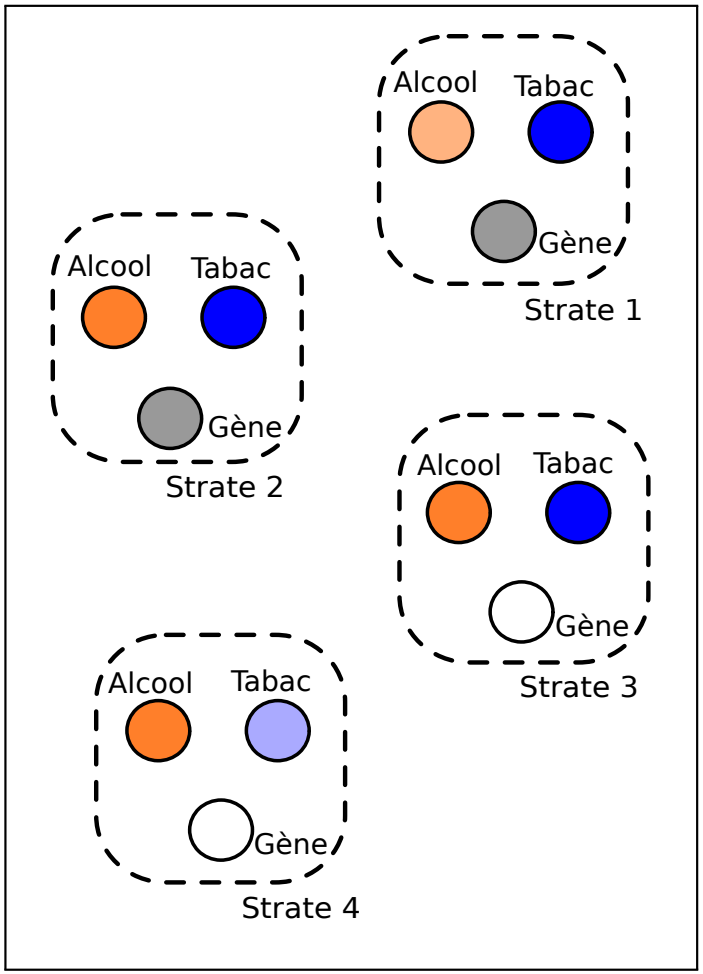
    $\Rightarrow$ to improve prediction accuracy

Example 1 : Linear regression on stratified data

- Association between $Y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$ on K predefined strata; $Z = 1, \ldots, K$.
- On the k-strata, n_k obs. :

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)} \boldsymbol{\beta}_k^* + \boldsymbol{\xi}^{(k)}$$

$\Rightarrow$ Kp parameters to be estimated



Example 2 : Case-control studies with multiple subtypes of disease

(Ballout, Garcia and V., submitted)

- $\sim$ Multinomial logistic regression
  - $Y \in \{0, 1, \ldots, K\}$
    - $Y = 0$: control
    - $Y = k > 0$: case, of subtype k.
  - Example : K breast cancer histological subtypes (e.g., based on ER/PR status).

$\Rightarrow$ No natural order among the subtypes

$$\log\left(\frac{\mathbb{P}(Y = k|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x})}\right) = \mathbf{x}^T \boldsymbol{\beta}_k^*$$

Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\dots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$

Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\ldots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$
    - $Z_i^j = k$: subtype of the case

Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\dots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$
    - $Z_i^j = k$: subtype of the case
  - The global study: K sub-studies
    1. m_1 pairs: Subtype 1 BC Vs Control
    2. m_2 pairs: Subtype 2 BC Vs Control
    3. ...
    4. m_K pairs: Subtype K BC Vs Control

Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\dots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$
    - $Z_i^j = k$: subtype of the case
  - The global study: K sub-studies
    1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \boldsymbol{\beta}_1^*$
    2. m_2 pairs: Subtype 2 BC Vs Control
    3. ...
    4. m_K pairs: Subtype K BC Vs Control

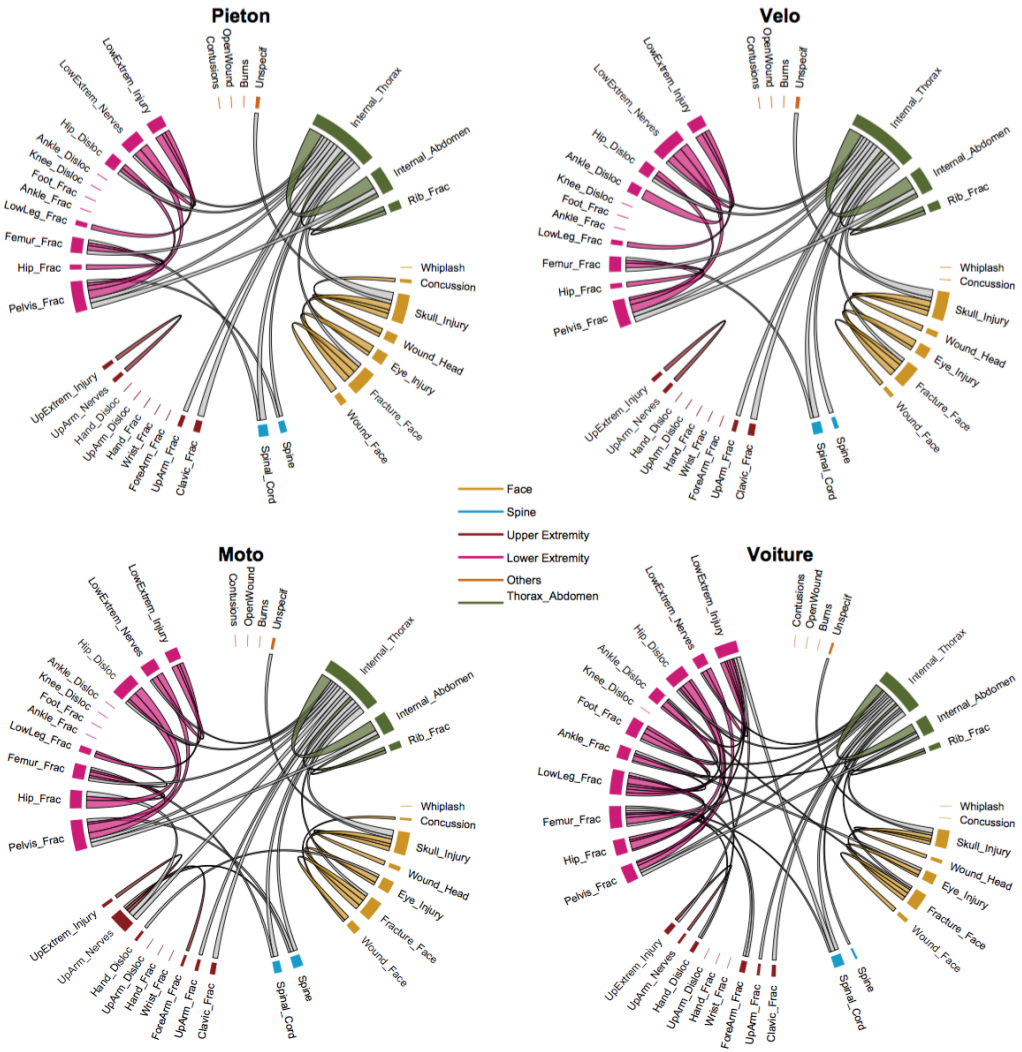
Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\dots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$
    - $Z_i^j = k$: subtype of the case
  - The global study: K sub-studies
    1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \boldsymbol{\beta}_1^*$
    2. m_2 pairs: Subtype 2 BC Vs Control $\Rightarrow \boldsymbol{\beta}_2^*$
    3. ...
    4. m_K pairs: Subtype K BC Vs Control

Example 3 : matched case-control studies (Ballout, Garcia and V., submitted)

- Same as before, but for each case, one matched control
  - $m = n/2$ pairs of observations, $(\mathbf{x}_i^j, Y_i^j, Z_i^j)_{i=1,\dots,m}^{j=1,2}$
    - one case, i.e. $Y_i^1 = 1$.
    - one matched control, i.e. $Y_i^2 = 0$
    - $Z_i^j = k$: subtype of the case
  - The global study: K sub-studies
    1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \boldsymbol{\beta}_1^*$
    2. m_2 pairs: Subtype 2 BC Vs Control $\Rightarrow \boldsymbol{\beta}_2^*$
    3. ...
    4. m_K pairs: Subtype K BC Vs Control $\Rightarrow \boldsymbol{\beta}_K^*$

Example 4 : K binary graphical models (Ballout and V., Statist. Med., 2019)



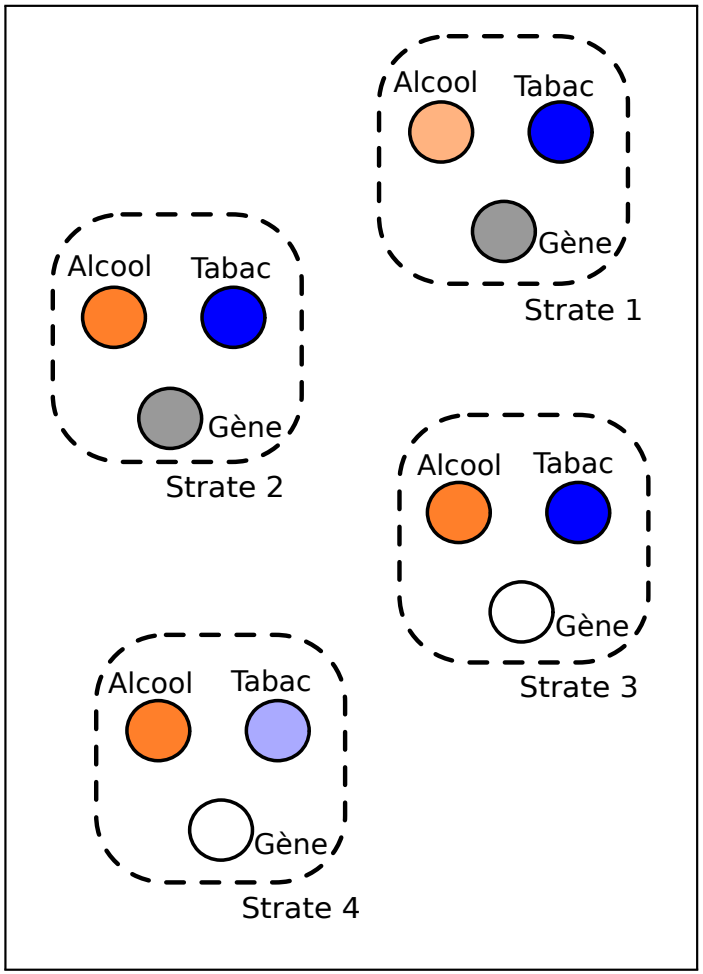
- Association among injuries suffered by victims of road accidents
- groups: $\sim$ road user type

Example 1 : Linear regression on stratified data

- Association between $Y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$ on K predefined strata.
- On the k-strata, n_k obs. :

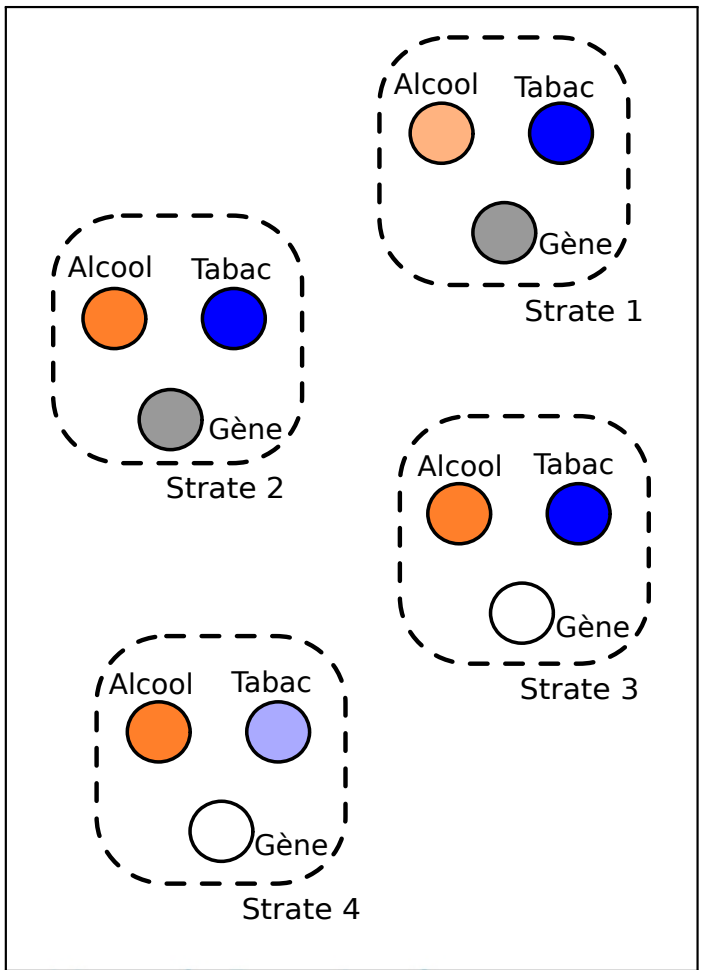
$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)} \boldsymbol{\beta}_k^* + \boldsymbol{\xi}^{(k)}$$

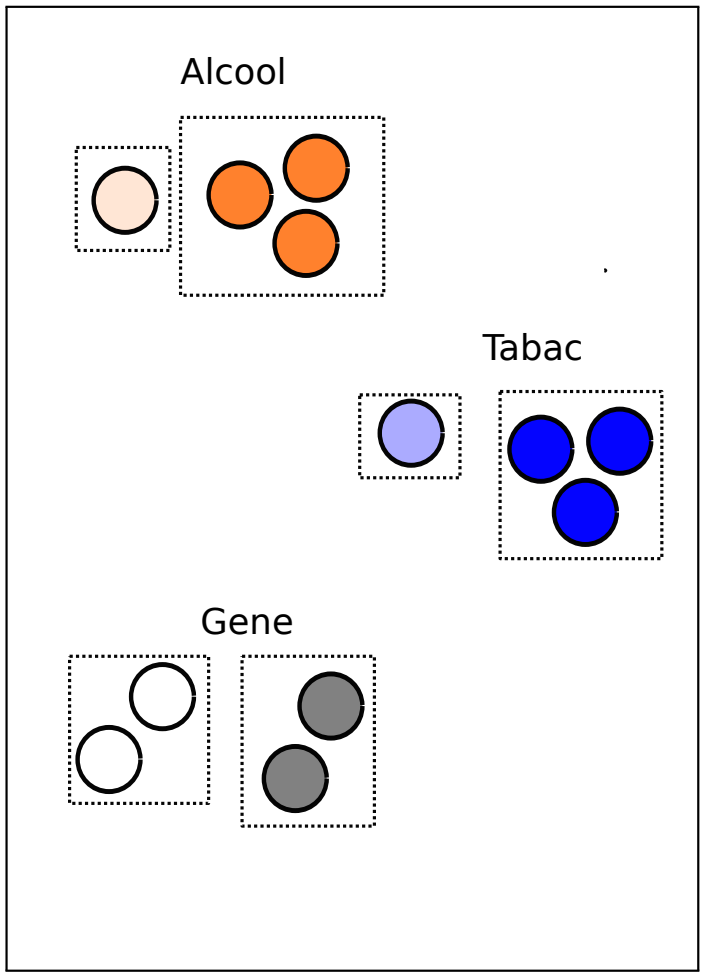
$\Rightarrow$ Kp parameters to be estimated



Main objective

- for each covariate: to identify the structure of its effects across the strata





Matrix formulation of the model

We have

$$\mathcal{Y} = \mathcal{X}\mathbf{b}^* + \boldsymbol{\xi},$$

with

$$\mathcal{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \vdots \\ \mathbf{Y}^{(K)} \end{pmatrix}, \ \mathcal{X} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{X}^{(K)} \end{pmatrix}, \ \mathbf{b}^* = \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \vdots \\ \boldsymbol{\beta}_K^* \end{pmatrix}$$

which are elements of $\mathbb{R}^n$, $\mathbb{R}^{n \times Kp}$ and $\mathbb{R}^{Kp}$,
where $n = \sum_{k=1}^{K} n_k$.

Oracle procedure (p=1)

- $\mathbf{b}^* = (\beta_1^*, \ldots, \beta_K^*) \in \mathbb{R}^K$, with
  - $\beta_1^* = 0$
  - β_2^*
  - $\beta_3^* = \ldots = \beta_K^*$

$$\mathbf{y} = \begin{pmatrix} X^{(1)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & X^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & X^{(K)} \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \vdots \\ \beta_K^* \end{pmatrix} + \boldsymbol{\xi}$$

Oracle procedure (p=1)

- $\mathbf{b}^* = (\beta_1^*, \ldots, \beta_K^*) \in \mathbb{R}^K$, with
  - $\beta_1^* = 0$
  - β_2^*
  - $\beta_3^* = \ldots = \beta_K^*$

$$\mathbf{y} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ X^{(2)} & \mathbf{0} \\ \mathbf{0} & X^{(3)} \\ \vdots & \vdots \\ \mathbf{0} & X^{(K)} \end{pmatrix} \begin{pmatrix} \beta_2^* \\ \mu^* \end{pmatrix} + \boldsymbol{\xi}$$

- Oracle procedure: by taking advantage of the homogeneity,
  - reduced complexity: better estimation accuracy
    - greater prediction accuracy
    - greater power to detect $\beta_2^* \neq \mu^*$

Penalized criteria

$$(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K) \in \underset{(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)}{\operatorname{argmin}} \left\{ \sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)} \boldsymbol{\beta}_k\|_2^2}{2n} + \operatorname{pen}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \right\}$$

Aim: To achieve a trade-off between

- goodness-of-fit
- match with the expected structure among vectors $\boldsymbol{\beta}_k^*$:
  - Sparsity: each $\boldsymbol{\beta}_k^*$ is sparse
  - Homogeneity: for some (k, k', j), $\beta_{k,j}^* = \beta_{k',j}^*$

Existing approaches... (1/2)

- Group-lasso, trace-norm:
  - ok for prediction error,
  - but not well suited for the identification of partitions

- ▸ Generalized Fused Lasso (Gertheiss et Tutz, 2012, V. et al. 2016):
  - oracular properties (adaptive version) when Kp is fixed ($n \to \infty$);
  - what about $Kp \to \infty$?
  - implementation is "not easy" (extensions to other models, etc.)

Existing approaches... (2/2)

- Standard approaches in epidemiology
  - either estimation on each stratum "independently",
  - or selection of one reference stratum + interaction tests

Existing approaches... (2/2)

- Standard approaches in epidemiology
  - either estimation on each stratum "independently",
  - or selection of one reference stratum + interaction tests

  - RefLasso with $r = 3$: $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_3^* + \boldsymbol{\gamma}_k^*$, avec $\boldsymbol{\gamma}_3^* = \mathbf{0}_p$

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\beta}_3 + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\beta}_3\|_1 + \sum_{k \neq 3} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1$$

Existing approaches... (2/2)

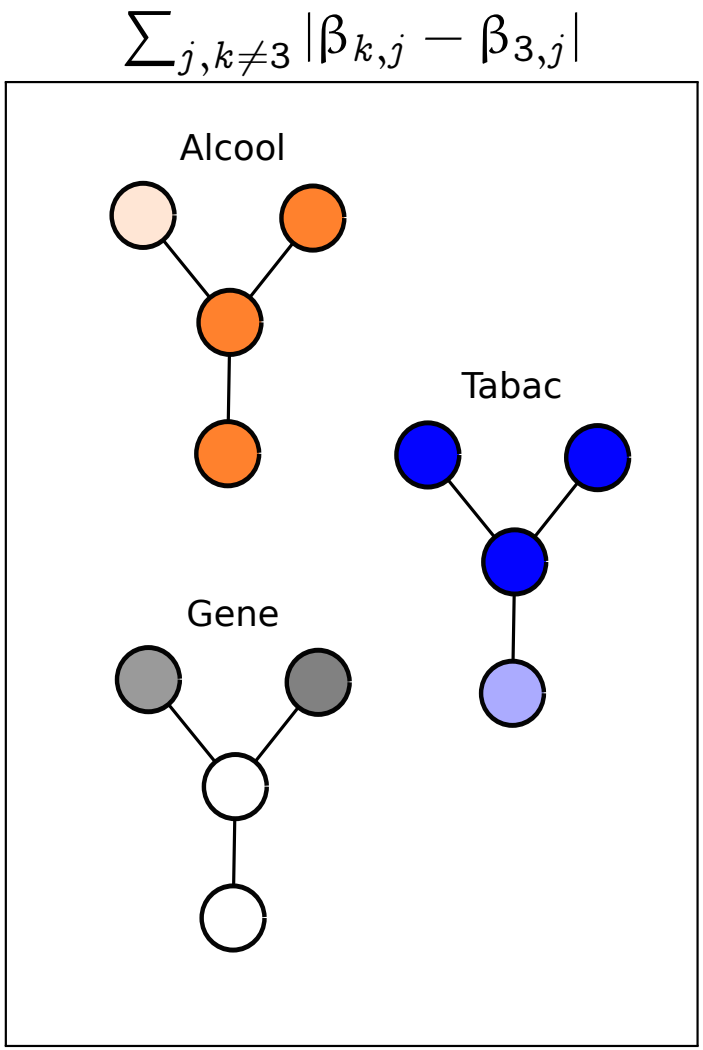
- Standard approaches in epidemiology
  - either estimation on each stratum "independently",
  - or selection of one reference stratum + interaction tests

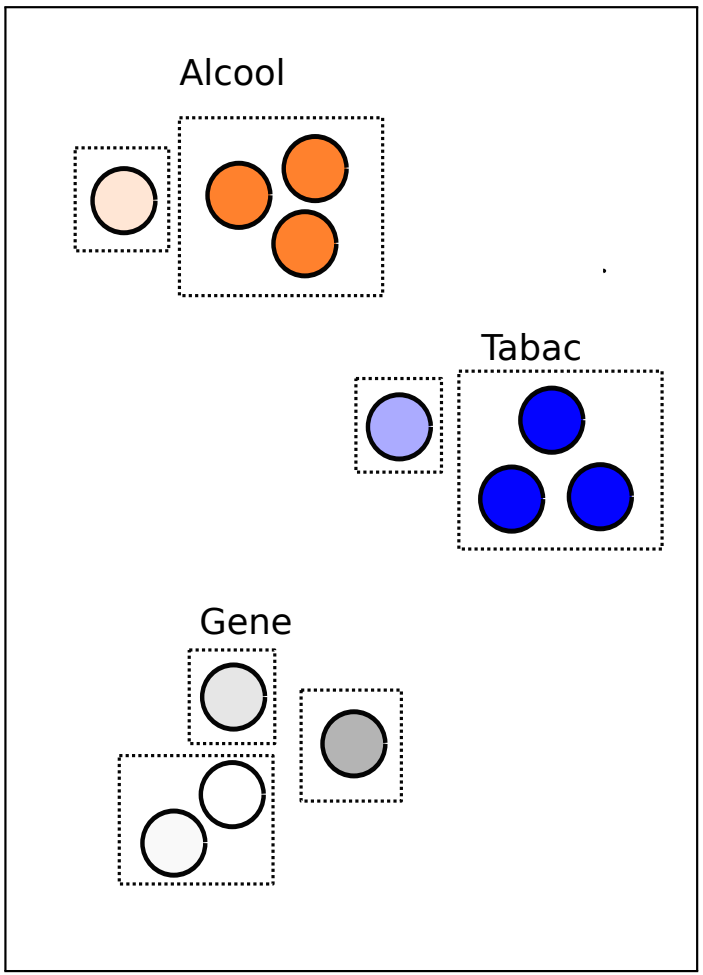
  - RefLasso with $r = 3$: $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_3^* + \boldsymbol{\gamma}_k^*$, avec $\boldsymbol{\gamma}_3^* = \mathbf{0}_p$

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\beta}_3 + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\beta}_3\|_1 + \sum_{k \neq 3} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1$$

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\beta}_3\|_1 + \sum_{k \neq 3} \lambda_{2,k} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_3\|_1$$

RefLasso as a Star-Fused Lasso; example with $r = 3$

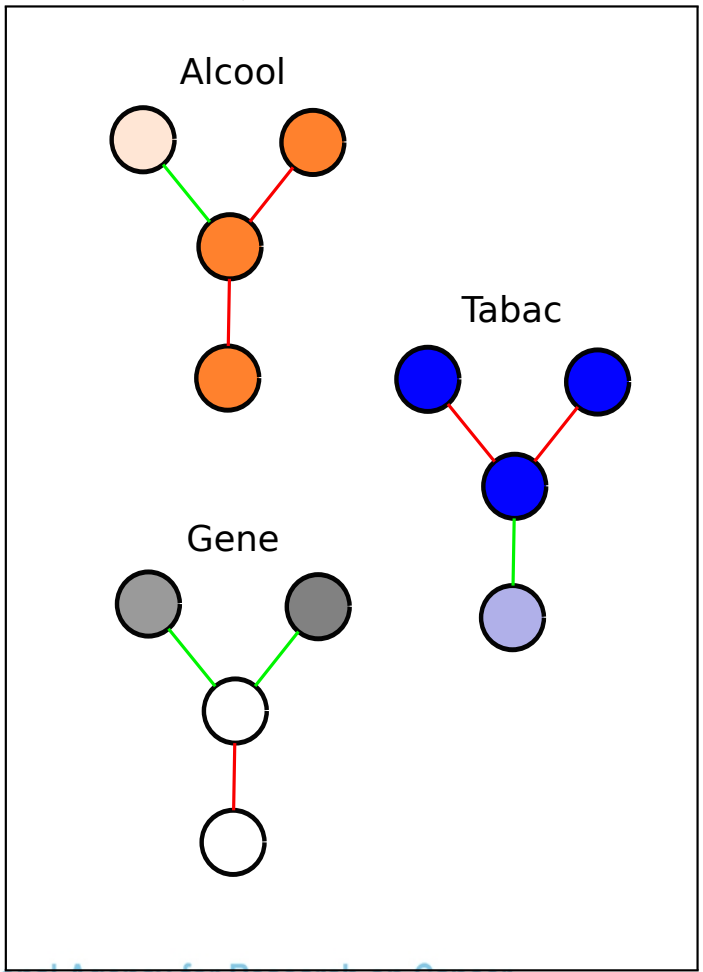




RefLasso: influence of the reference group

Reference $r = 3$

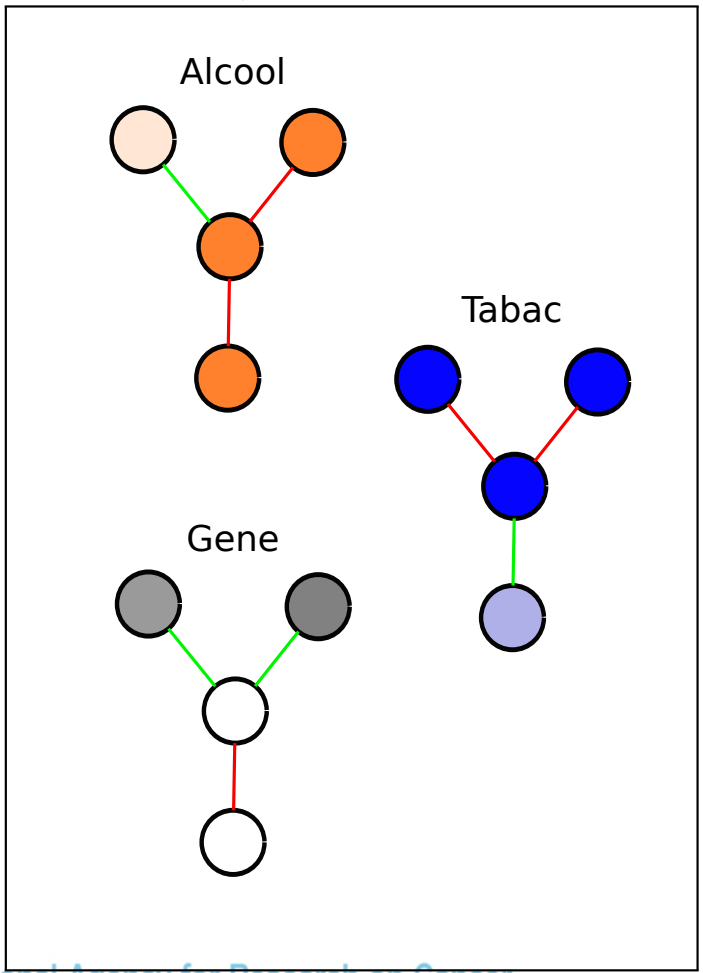
$\sum_{j,k \neq 3} |\beta_{k,j} - \beta_{3,j}|$



RefLasso: influence of the reference group

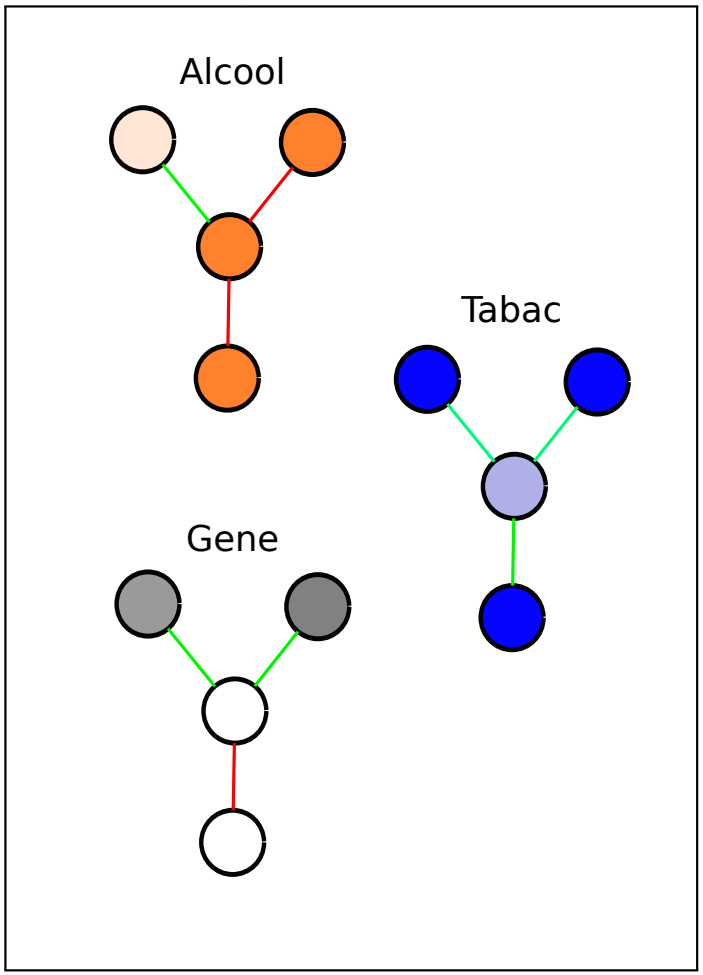
Reference $r = 3$

$\sum_{j,k \neq 3} |\beta_{k,j} - \beta_{3,j}|$

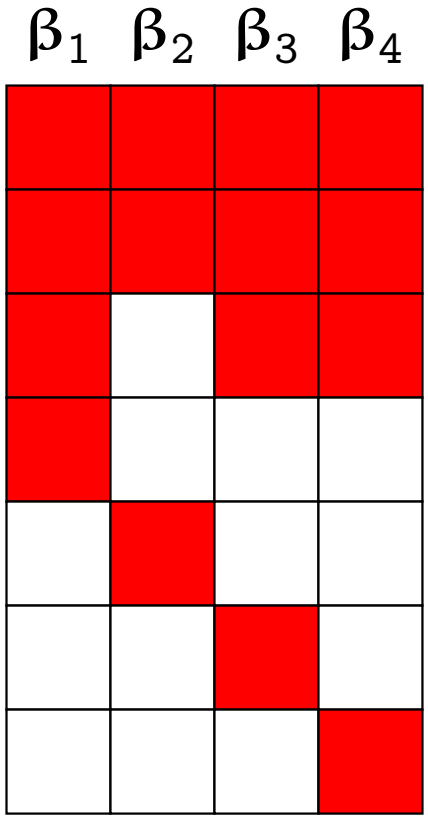


Reference $r = 4$

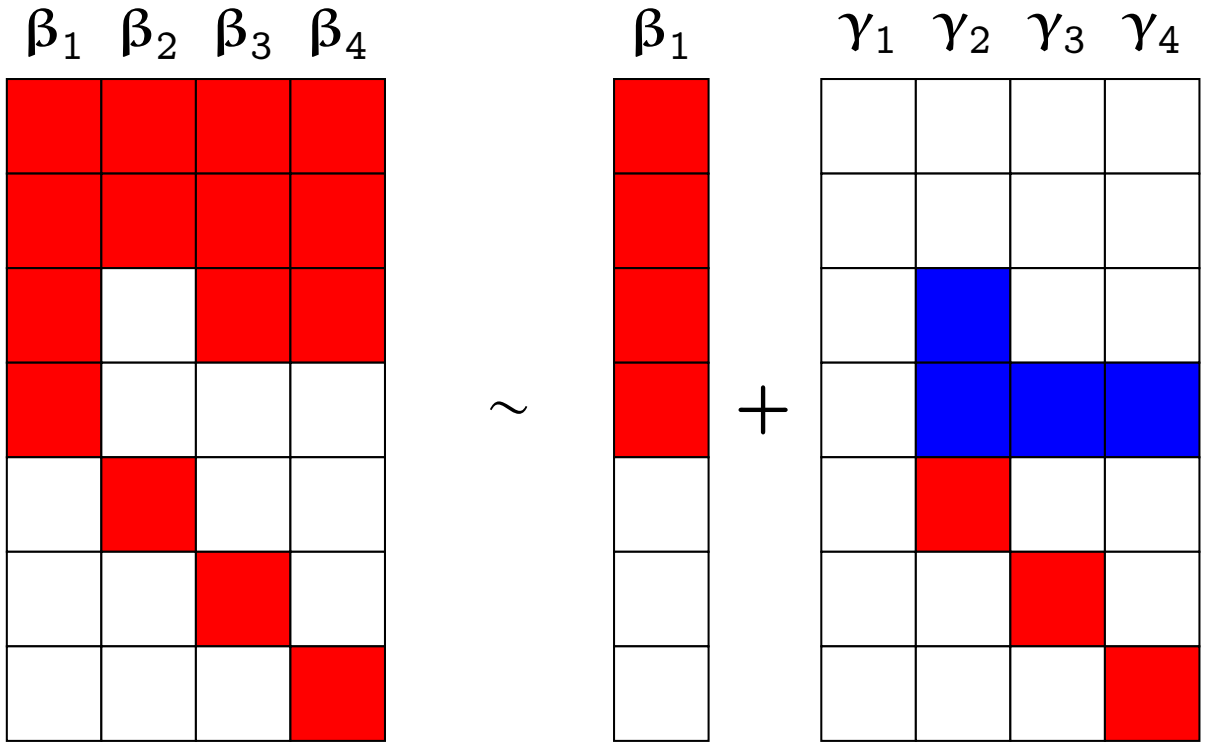
$\sum_{j,k \neq 4} |\beta_{k,j} - \beta_{4,j}|$



RefLasso = Rank 1 matrix + sparse matrix decomposition

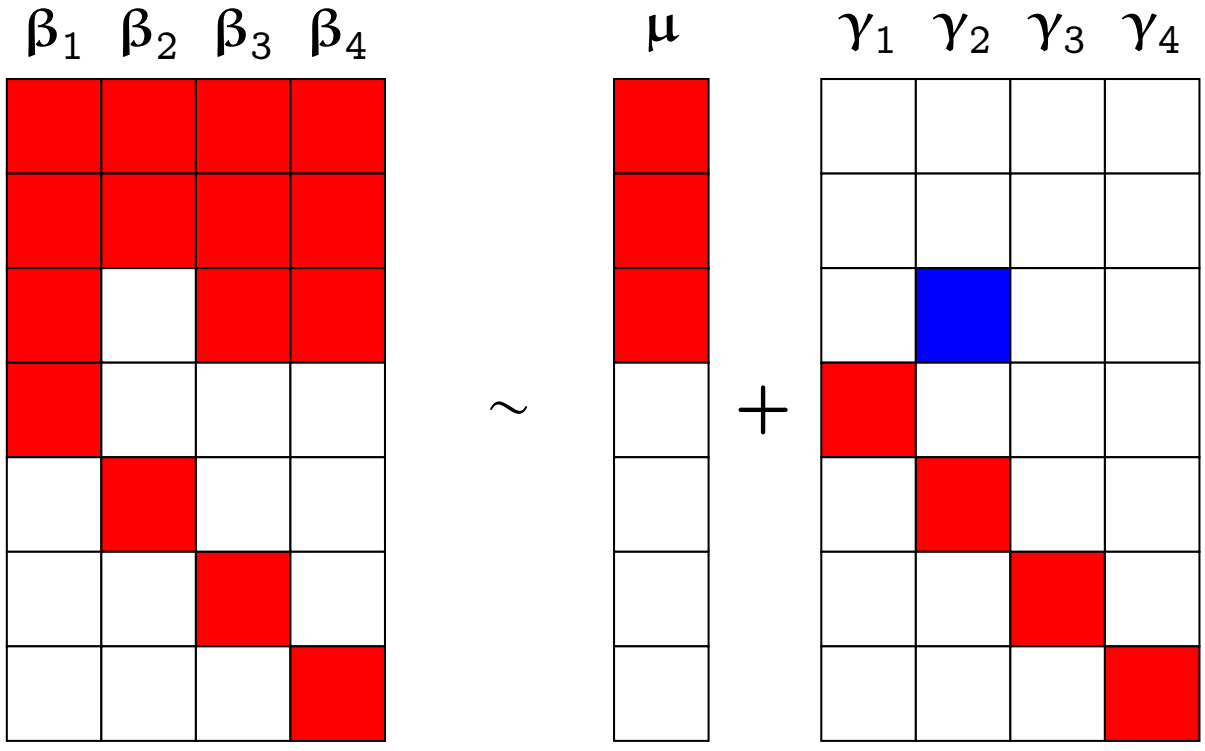


RefLasso = Rank 1 matrix + sparse matrix decomposition



$C = 15$ $C^* = 4 + 7 = 11$

RefLasso = Rank 1 matrix + sparse matrix decomposition



Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

- Particular constraints:
  - $\boldsymbol{\mu}^* = \mathbf{0}_p$: independent estimations: complexity $\sum_k \|\boldsymbol{\beta}_k^*\|_0$

Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

- Particular constraints:
  - $\boldsymbol{\mu}^* = \mathbf{0}_p$: independent estimations: complexity $\sum_k \|\boldsymbol{\beta}_k^*\|_0$
  - $\boldsymbol{\gamma}_1^* = \mathbf{0}_p$: reference stratum: $\boldsymbol{\mu}^* = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\gamma}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$.

Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

- Particular constraints:
  - $\boldsymbol{\mu}^* = \mathbf{0}_p$: independent estimations: complexity $\sum_k \|\boldsymbol{\beta}_k^*\|_0$
  - $\boldsymbol{\gamma}_1^* = \mathbf{0}_p$: reference stratum: $\boldsymbol{\mu}^* = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\gamma}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$.
    - Complexity : $\|\boldsymbol{\beta}_1^*\|_0 + \sum_k \|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*\|_0$.

Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

- Particular constraints:
  - $\boldsymbol{\mu}^* = \mathbf{0}_p$: independent estimations: complexity $\sum_k \|\boldsymbol{\beta}_k^*\|_0$
  - $\boldsymbol{\gamma}_1^* = \mathbf{0}_p$: reference stratum: $\boldsymbol{\mu}^* = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\gamma}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$.
    - Complexity : $\|\boldsymbol{\beta}_1^*\|_0 + \sum_k \|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*\|_0$.

    $\Rightarrow$ generally sub-optimal

Over-parametrized decomposition of the $\boldsymbol{\beta}_k^*$'s

- Consider the over-parametrization

$$(\boldsymbol{\mu}^*, \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^*) \in \mathbb{R}^{(K+1)p} \quad \text{tq}: \quad \boldsymbol{\beta}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\gamma}_k^*$$

- Particular constraints:
  - $\boldsymbol{\mu}^* = \mathbf{0}_p$: independent estimations: complexity $\sum_k \|\boldsymbol{\beta}_k^*\|_0$
  - $\boldsymbol{\gamma}_1^* = \mathbf{0}_p$: reference stratum: $\boldsymbol{\mu}^* = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\gamma}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$.
    - Complexity : $\|\boldsymbol{\beta}_1^*\|_0 + \sum_k \|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*\|_0$.

    $\Rightarrow$ generally sub-optimal

  - Optimal constraint: $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*_{\boldsymbol{r}^*}$ where

$$\mu^*_{\boldsymbol{r}^*,j} = \text{mode}(0, \beta^*_{1,j}, \ldots, \beta^*_{K,j}) = \beta^*_{r_j^*,j}$$

ie. such that $\{\|\boldsymbol{\mu}^*\|_0 + \sum_k \|\boldsymbol{\beta}_k^* - \boldsymbol{\mu}^*\|_0\}$ is minimized.

Data Shared Lasso

(Ollier and V., Biometrika, 2017), Gross and Tibshirani (CSDA, 2016)

- Objective: to mimic this "oracle".

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_k \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1$$

Data Shared Lasso

(Ollier and V., Biometrika, 2017), Gross and Tibshirani (CSDA, 2016)

- Objective: to mimic this "oracle".

or, equivalently:

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_k \lambda_{2,k} \|\boldsymbol{\beta}_k - \boldsymbol{\mu}\|_1$$

$$\begin{aligned} \hat{\mu}_j &= \min_m \left\{ \lambda_1 |m| + \textstyle\sum_k \lambda_{2,k} |\widehat{\beta}_{k,j} - m| \right\} \\ &= \text{WSmedian}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}) \end{aligned}$$

Implementation

- Set $\gamma_k = \boldsymbol{\beta}_k - \boldsymbol{\mu}$, $\tau_k = \lambda_{2,k}/\lambda_1$, and introduce

$$\mathfrak{X}_0 = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix} \quad \text{and} \quad \theta_0 = \begin{pmatrix} \boldsymbol{\mu} \\ \tau_1 \boldsymbol{\gamma}_1 \\ \vdots \\ \tau_K \boldsymbol{\gamma}_K \end{pmatrix}$$

which belong to $\mathbb{R}^{n \times (K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.

Implementation

- Set $\gamma_k = \beta_k - \mu$, $\tau_k = \lambda_{2,k}/\lambda_1$, and introduce

$$\mathcal{X}_0 = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix} \quad \text{and} \quad \theta_0 = \begin{pmatrix} \mu \\ \tau_1 \gamma_1 \\ \vdots \\ \tau_K \gamma_K \end{pmatrix}$$

  which belong to $\mathbb{R}^{n \times (K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.

- Data Shared Lasso minimizes

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\mu + \gamma_k)\|_2^2}{2n} + \lambda_1 \|\mu\|_1 + \sum_k \lambda_{2,k} \|\gamma_k\|_1.$$

Implementation

- Set $\gamma_k = \beta_k - \mu$, $\tau_k = \lambda_{2,k}/\lambda_1$, and introduce

$$\mathcal{X}_0 = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix} \quad \text{and} \quad \theta_0 = \begin{pmatrix} \mu \\ \tau_1 \gamma_1 \\ \vdots \\ \tau_K \gamma_K \end{pmatrix}$$

  which belong to $\mathbb{R}^{n \times (K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.

- Data Shared Lasso minimizes

$$\frac{\|\mathcal{Y} - \mathcal{X}_0 \theta_0\|_2^2}{2n} + \lambda_1 \|\theta_0\|_1.$$

  $\Rightarrow$ Implementation: straightforward, under a variety of models (GLMs, survival models, ...)

RefLasso: given one reference stratum (for each j)

- Let $\boldsymbol{r} = (r_1, \ldots, r_p)$ be one vector, indicating the reference stratum for each j.
  1. example 1 : $\boldsymbol{r} = (1, \ldots, 1)$
  2. example 2 : $\boldsymbol{r} = \boldsymbol{r}^* = (r_1^*, \ldots, r_p^*)$

RefLasso: given one reference stratum (for each j)

- Let $\boldsymbol{r} = (r_1, \ldots, r_p)$ be one vector, indicating the reference stratum for each j.
  1. example 1 : $\boldsymbol{r} = (1, \ldots, 1)$
  2. example 2 : $\boldsymbol{r} = \boldsymbol{r}^* = (r_1^*, \ldots, r_p^*)$
- Consider the decomposition $\boldsymbol{\beta}_k^* = \boldsymbol{\mu}_{\boldsymbol{r}}^* + \boldsymbol{\gamma}_{\boldsymbol{r},k}^*$ with
  - $\mu_{\boldsymbol{r}j}^* = \beta_{r_j,j}^*$
  - $\boldsymbol{\gamma}_{\boldsymbol{r},k}^* = \boldsymbol{\beta}_k^* - \boldsymbol{\mu}_{\boldsymbol{r}}^*$

RefLasso: given one reference stratum (for each j)

- Let $\boldsymbol{r} = (r_1, \ldots, r_p)$ be one vector, indicating the reference stratum for each j.
  1. example 1 : $\boldsymbol{r} = (1, \ldots, 1)$
  2. example 2 : $\boldsymbol{r} = \boldsymbol{r}^* = (r_1^*, \ldots, r_p^*)$
- Consider the decomposition $\boldsymbol{\beta}_k^* = \boldsymbol{\mu}_{\boldsymbol{r}}^* + \boldsymbol{\gamma}_{\boldsymbol{r},k}^*$ with
  - $\mu_{\boldsymbol{r}j}^* = \beta_{r_j,j}^*$
  - $\boldsymbol{\gamma}_{\boldsymbol{r},k}^* = \boldsymbol{\beta}_k^* - \boldsymbol{\mu}_{\boldsymbol{r}}^*$
- Estimation of these parameters can be obtained by minimizing

$$\sum_{k=1}^{K} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1,$$

under the constraints $\gamma_{r_j,j} = 0$ for all $j \in [p]$

- Recall $\gamma_k = \beta_k - \mu$, $\tau_k = \lambda_{2,k}/\lambda_1$

$$\mathcal{X}_0 = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix} \quad \text{and} \quad \theta_0 = \begin{pmatrix} \mu \\ \tau_1 \gamma_1 \\ \vdots \\ \tau_K \gamma_K \end{pmatrix}$$

  qui appartiennent à $\mathbb{R}^{n \times (K+1)p}$ et $\mathbb{R}^{(K+1)p}$.

- Data Shared Lasso (AutoRefLasso), AprioriRefLasso (r chosen a priori) et OracleRefLasso ($r = r^*$ optimal) all minimize

$$\frac{\|y - \mathcal{X}\theta\|_2^2}{2n} + \lambda_1 \|\theta\|_1$$

  for one particular choice "$\mathcal{X} \subseteq \mathcal{X}_0$".

  $\Rightarrow$ Unified framework to study and compare them (and implement them)

Irrepresentability conditions (Ollier and V., Biometrika, 2017)

- A priori : $\theta^*_{\boldsymbol{r}} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},K})^T \in \mathbb{R}^{Kp}$.
- Oracle : $\theta^*_{\boldsymbol{r}^*} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,K})^T \in \mathbb{R}^{Kp}$
- Auto : $\theta^*_0 = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{0,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{0,K})^T \in \mathbb{R}^{(K+1)p}$

Irrepresentability conditions (Ollier and V., Biometrika, 2017)

- A priori : $\theta^*_{\boldsymbol{r}} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},K})^T \in \mathbb{R}^{Kp}$.
- Oracle : $\theta^*_{\boldsymbol{r}^*} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,K})^T \in \mathbb{R}^{Kp}$
- Auto : $\theta^*_0 = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{0,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{0,K})^T \in \mathbb{R}^{(K+1)p}$
- $supp(\theta^*_{\boldsymbol{r}}) = J_{\boldsymbol{r}}$ and $supp(\theta^*_0) = J_0$

$\Leftrightarrow$ $S_{\boldsymbol{r}} = \{j : \beta^*_{r_j,j} \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}$

- $\theta^*_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}} = \theta^*_{0 J_0}$ and so $\mathcal{X}_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}} = \mathcal{X}_{0 J_0}$

Irrepresentability conditions (Ollier and V., Biometrika, 2017)

- A priori : $\theta^*_{\boldsymbol{r}} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r},K})^T \in \mathbb{R}^{Kp}$.
- Oracle : $\theta^*_{\boldsymbol{r}^*} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{\boldsymbol{r}^*,K})^T \in \mathbb{R}^{Kp}$
- Auto : $\theta^*_0 = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}^*}, \tau_1 \boldsymbol{\gamma}^{*T}_{0,1}, \ldots, \tau_K \boldsymbol{\gamma}^{*T}_{0,K})^T \in \mathbb{R}^{(K+1)p}$
- $supp(\theta^*_{\boldsymbol{r}}) = J_{\boldsymbol{r}}$ and $supp(\theta^*_0) = J_0$

$\Leftrightarrow$ $S_{\boldsymbol{r}} = \{j : \beta^*_{r_j,j} \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}$

- $\theta^*_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}} = \theta^*_{0 J_0}$ and so $\mathcal{X}_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}} = \mathcal{X}_{0 J_0}$

$$
\begin{aligned}
(IC)_{\boldsymbol{r}} : \quad & \Lambda_{\min}(\mathcal{X}^T_{\boldsymbol{r} J_{\boldsymbol{r}}} \mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}}) \geqslant C_{\boldsymbol{r}} > 0 \text{ et} \\
& c_{\boldsymbol{r}} = \max_{j \notin J_{\boldsymbol{r}}} \|(\mathcal{X}^T_{\boldsymbol{r} J_{\boldsymbol{r}}} \mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}})^{-1} \mathcal{X}^T_{\boldsymbol{r} J_{\boldsymbol{r}}} \mathcal{X}_{\boldsymbol{r} j}\|_1 < 1, \\
(IC)_0 : \quad & \Lambda_{\min}(\mathcal{X}^T_{0 J_0} \mathcal{X}_{0 J_0}) \geqslant C_0 > 0 \text{ et} \\
& c_0 = \max_{j \notin J_0} \|(\mathcal{X}^T_{0 J_0} \mathcal{X}_{0 J_0})^{-1} \mathcal{X}^T_{0 J_0} \mathcal{X}_{0 j}\|_1 < 1.
\end{aligned}
$$

Irrepresentability conditions (Ollier and V., Biometrika, 2017)

- A priori : $\theta_{\boldsymbol{r}}^* = (\boldsymbol{\mu}_{\boldsymbol{r}}^{*T}, \tau_1 \boldsymbol{\gamma}_{\boldsymbol{r},1}^{*T}, \ldots, \tau_K \boldsymbol{\gamma}_{\boldsymbol{r},K}^{*T})^T \in \mathbb{R}^{Kp}$.
- Oracle : $\theta_{\boldsymbol{r}^*}^* = (\boldsymbol{\mu}_{\boldsymbol{r}^*}^{*T}, \tau_1 \boldsymbol{\gamma}_{\boldsymbol{r}^*,1}^{*T}, \ldots, \tau_K \boldsymbol{\gamma}_{\boldsymbol{r}^*,K}^{*T})^T \in \mathbb{R}^{Kp}$
- Auto : $\theta_0^* = (\boldsymbol{\mu}_{\boldsymbol{r}^*}^{*T}, \tau_1 \boldsymbol{\gamma}_{0,1}^{*T}, \ldots, \tau_K \boldsymbol{\gamma}_{0,K}^{*T})^T \in \mathbb{R}^{(K+1)p}$
- $supp(\theta_{\boldsymbol{r}}^*) = J_{\boldsymbol{r}}$ and $supp(\theta_0^*) = J_0$

$\Leftrightarrow$ $S_{\boldsymbol{r}} = \{j : \beta_{r_j,j}^* \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta_{k,j}^* \neq \beta_{r_j,j}^*\}$

- $\theta_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}}^* = \theta_{0 J_0}^*$ and so $\mathcal{X}_{\boldsymbol{r}^* J_{\boldsymbol{r}^*}} = \mathcal{X}_{0 J_0}$

$$
\begin{aligned}
(IC)_{\boldsymbol{r}} : \quad & \Lambda_{\min}(\mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}}^T \mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}}) \geqslant C_{\boldsymbol{r}} > 0 \text{ et} \\
& c_{\boldsymbol{r}} = \max_{j \notin J_{\boldsymbol{r}}} \|(\mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}}^T \mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}})^{-1} \mathcal{X}_{\boldsymbol{r} J_{\boldsymbol{r}}}^T \mathcal{X}_{\boldsymbol{r} j}\|_1 < 1, \\
(IC)_0 : \quad & \Lambda_{\min}(\mathcal{X}_{0 J_0}^T \mathcal{X}_{0 J_0}) \geqslant C_0 > 0 \text{ et} \\
& c_0 = \max_{j \notin J_0} \|(\mathcal{X}_{0 J_0}^T \mathcal{X}_{0 J_0})^{-1} \mathcal{X}_{0 J_0}^T \mathcal{X}_{0 j}\|_1 < 1.
\end{aligned}
$$

International Agency for Research on Cancer

World Health Organization

$\Rightarrow$ $(IC)_{\boldsymbol{r}^*}$ and $(IC)_0$ are *almost* identical.

The particular case where $n_k = n/K$ and $(\mathbf{X}^{(k)^T}\mathbf{X}^{(k)})/n_k = \mathbf{I}_{n_k}$

- $S_{\boldsymbol{r}} = \{j : \beta^*_{r_j,j} \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}$
- $\mathcal{D}_{\boldsymbol{r},0} = \max_{j \notin S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or 0)
- $\mathcal{D}_{\boldsymbol{r},1} = \max_{j \in S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or $-\infty$)

The particular case where $n_k = n/K$ and $(\mathbf{X}^{(k)^T}\mathbf{X}^{(k)})/n_k = \mathbf{I}_{n_k}$

- $S_{\boldsymbol{r}} = \{j : \beta^*_{r_j,j} \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}$
- $\mathcal{D}_{\boldsymbol{r},0} = \max_{j \notin S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or 0)
- $\mathcal{D}_{\boldsymbol{r},1} = \max_{j \in S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or $-\infty$)
- Let $\tau_k = \tau_0 K^{-1/2}$ for some $\tau_0 > 0$,

$$(sIC)_{\boldsymbol{r}}: \qquad 0 \leqslant \frac{K^{1/2}}{K - 2\mathcal{D}_{\boldsymbol{r},1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\boldsymbol{r},0}}.$$

$$(sIC)_0: \qquad 0 \leqslant \frac{K^{1/2}}{K - 2\mathcal{D}_{\boldsymbol{r}^*,1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\boldsymbol{r}^*,0}}.$$

The particular case where $n_k = n/K$ and $(\mathbf{X}^{(k)^T}\mathbf{X}^{(k)})/n_k = \mathbf{I}_{n_k}$

- $S_{\boldsymbol{r}} = \{j : \beta^*_{r_j,j} \neq 0\}$, $T_{\boldsymbol{r}} = \{(k,j) \in [K] \times [p] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}$
- $\mathcal{D}_{\boldsymbol{r},0} = \max_{j \notin S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or 0)
- $\mathcal{D}_{\boldsymbol{r},1} = \max_{j \in S_{\boldsymbol{r}}} |\{k \in [K] : \beta^*_{k,j} \neq \beta^*_{r_j,j}\}|$ (or $-\infty$)
- Let $\tau_k = \tau_0 K^{-1/2}$ for some $\tau_0 > 0$,

$$(sIC)_{\boldsymbol{r}}: \qquad 0 \leqslant \frac{K^{1/2}}{K - 2\mathcal{D}_{\boldsymbol{r},1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\boldsymbol{r},0}}.$$

$$(sIC)_0: \qquad 0 \leqslant \frac{K^{1/2}}{K - 2\mathcal{D}_{\boldsymbol{r}^*,1}} < \tau_0 < \frac{K^{1/2}}{\mathcal{D}_{\boldsymbol{r}^*,0}}.$$

- $(sIC)_0 = (sIC)_{\boldsymbol{r}^*}$
- If $\mathcal{D}_{\boldsymbol{r},0} = \mathcal{D}_{\boldsymbol{r},1} = \mathcal{D}_{\boldsymbol{r}}$ then $(sIC)_{\boldsymbol{r}} \Rightarrow \mathcal{D}_{\boldsymbol{r}} < K/3$
- $(sIC)_{\boldsymbol{r}}$ generally stronger than $(sIC)_{\boldsymbol{r}^*}$

Theorem 1 (Ollier and V., Biometrika, 2017)

- *Assume that* $\varepsilon_i^{(k)}$ *: i.i.d. centered sub-Gaussian variables with parameter* $\sigma > 0$.
- *Under* $(sIC)_0$, *introduce:*

$$\gamma = \min\left\{1 - \mathcal{D}_0\tau_0 K^{-1/2}, 1 - \frac{K^{1/2} + \mathcal{D}_1\tau_0}{(K - \mathcal{D}_1)\tau_0}\right\}$$

$$C_{\min} = \min\left(1, \tau_0^{-2}, \frac{1}{2}\left[\left(\tau_0^{-2} + 1\right) - \left\{\left(\tau_0^{-2} - 1\right)^2 + \frac{4\mathcal{D}_1}{\tau_0^2 K}\right\}\right]\right)$$

$$\lambda_1 > \frac{2}{\gamma \min(1, \tau_0)}\left[\frac{2\sigma^2 \log\{(K+1)p\}}{n}\right]^{1/2}, \quad \lambda_{2,k} = \tau_k \lambda_1$$

$$\beta_{\min} = \lambda_1\left[(|S_{\boldsymbol{r}^*}| + |T_{\boldsymbol{r}^*}|)^{1/2} C_{\min}^{-1} + 4\sigma C_{\min}^{-1/2}\right],$$

- *Assume that* $|\beta^*_{r^*_j, j}| > \beta_{\min}$ *and* $|\beta^*_{k,j} - \beta^*_{r^*_j, j}| > \frac{K^{1/2}\beta_{\min}}{\tau_0}$.

Then, with high probability,

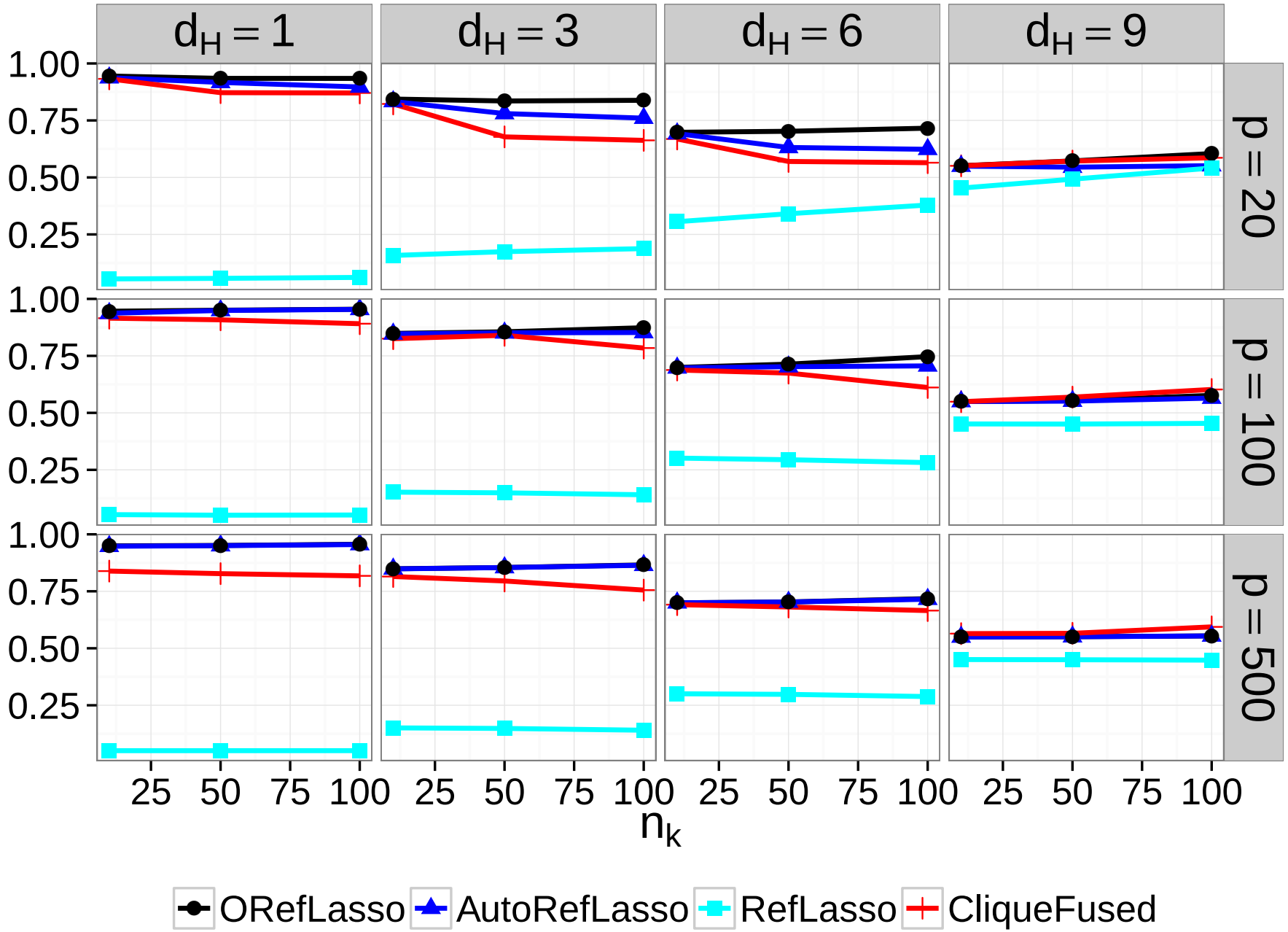
- $S_{\boldsymbol{r}^*}$ *and* $T_{\boldsymbol{r}^*}$ *are both recovered;*
- $\|\widehat{\theta}_{0J_0} - \theta^*_{0J_0}\|_\infty \leqslant \beta_{\min}$.

Simulations in the Linear regression case

- $K = 20$; $\mathbf{x}_i^{(k)} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$.
- $P_0 \subseteq [p]$, with $|P_0| = 15$

$$\beta^*_{k,j} = \begin{cases} 0 & \forall j \in [p] \setminus P_0 \\ 1 & j \in P_0 \ \& \ k > d_H \\ 1 + K^{1/2} & j \in P_0 \ \& \ k \leqslant d_H \end{cases}$$

- Accuracy regarding:
  - $T^*_{P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{r^*_j,j}\}$
  - $T^*_{1,P_0} = \{(k,j) \in [K] \times P_0 : \beta^*_{k,j} \neq \beta^*_{1,j}\}$


$d_H = 1$
$d_H = 3$
$d_H = 6$
$d_H = 9$
p = 20
p = 100
p = 500
1.00
0.75
0.50
0.25
25 50 75 100
n_k
ORefLasso
AutoRefLasso
RefLasso
CliqueFused

Sparse multinomial Logistic regression (Ballout, Garcia and V., submitted)

- In unmatched designs, if
  - $Y = 0$: control ["natural" reference category]
  - $Y = k$: case of subtype k, $k = 1, \ldots, K$
  - no natural order among the subtypes.

$\Rightarrow$ Sparse multinomial logistic regression is a natural extension of sparse (binary) logistic regression.

Sparse multinomial Logistic regression (Ballout, Garcia and V., submitted)

- In unmatched designs, if
  - $Y = 0$: control ["natural" reference category]
  - $Y = k$: case of subtype k, $k = 1, \ldots, K$
  - no natural order among the subtypes.

$\Rightarrow$ Sparse multinomial logistic regression is a natural extension of sparse (binary) logistic regression.

- Two formulations exist. No clear guidance in the literature on which one to chose in practice.

Two formulations

1. Select a reference category (say 0) and then assume the existence of K parameter vectors $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^* \in \mathbb{R}^p$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\{\exp(\mathbf{x}^T \boldsymbol{\delta}_\ell^*)\}^{\mathbf{1}(\ell \neq 0)}}{1 + \sum_{k=1}^K \exp(\mathbf{x}^T \boldsymbol{\delta}_k^*)}.$$

Two formulations

1. Select a reference category (say 0) and then assume the existence of K parameter vectors $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^* \in \mathbb{R}^p$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\{\exp(\mathbf{x}^T \boldsymbol{\delta}_\ell^*)\}^{\mathbf{1}(\ell \neq 0)}}{1 + \sum_{k=1}^K \exp(\mathbf{x}^T \boldsymbol{\delta}_k^*)}.$$

SparseMultinomialRef0 maximizes

$$L(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) - \lambda \sum_{k=1}^K \|\boldsymbol{\gamma}_k\|_1.$$

Two formulations

1. Select a reference category (say 0) and then assume the existence of K parameter vectors $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^* \in \mathbb{R}^p$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\{\exp(\mathbf{x}^T \boldsymbol{\delta}_\ell^*)\}^{\mathbf{1}(\ell \neq 0)}}{1 + \sum_{k=1}^K \exp(\mathbf{x}^T \boldsymbol{\delta}_k^*)}.$$

2. Use a symmetric over-parametrization instead: assume the existence of $K+1$ vectors $(\boldsymbol{\beta}_0^*, \ldots, \boldsymbol{\beta}_K^*\}$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_\ell^*)}{\sum_{k=0}^K \exp(\mathbf{x}^T \boldsymbol{\beta}_k^*)}.$$

$(\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_0^*)$

Two formulations

1. Select a reference category (say 0) and then assume the existence of K parameter vectors $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_K^* \in \mathbb{R}^p$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\{\exp(\mathbf{x}^T \boldsymbol{\delta}_\ell^*)\}^{\mathbf{1}(\ell \neq 0)}}{1 + \sum_{k=1}^K \exp(\mathbf{x}^T \boldsymbol{\delta}_k^*)}.$$

2. Use a symmetric over-parametrization instead: assume the existence of $K+1$ vectors $\{\boldsymbol{\beta}_0^*, \ldots, \boldsymbol{\beta}_K^*\}$ s.t.

$$\mathbb{P}(Y = \ell | \mathbf{x}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_\ell^*)}{\sum_{k=0}^K \exp(\mathbf{x}^T \boldsymbol{\beta}_k^*)}.$$

$(\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_0^*)$
glmnet maximizes

$$\mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k=0}^K \|\boldsymbol{\beta}_k\|_1.$$

Link between the two formulations

- Data Shared Version of MultinomRef0:
  - write $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$
  - and maximize

$$L(\boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_K) - \lambda \|\boldsymbol{\mu}\|_1 - \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1$$

Link between the two formulations

- Data Shared Version of MultinomRef0:
  - write $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$
  - and maximize

$$L(\boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_K) - \lambda \|\boldsymbol{\mu}\|_1 - \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1$$

$$= \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k \geqslant 0} \|\boldsymbol{\beta}_k\|_1$$

Link between the two formulations

- Data Shared Version of MultinomRef0:
  - write $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$
  - and maximize

$$L(\boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_K) - \lambda \|\boldsymbol{\mu}\|_1 - \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1$$

$$= \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k \geqslant 0} \|\boldsymbol{\beta}_k\|_1$$

- glmnet returns $(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$, which is exactly the estimates $(-\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_K)$ we would obtain by performing MultinomRef0 with Data Shared lasso

Link between the two formulations

- Data Shared Version of MultinomRef0:
  - write $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$
  - and maximize

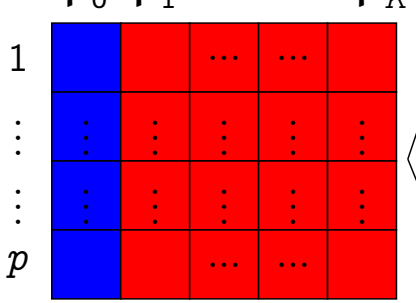
$$L(\boldsymbol{\mu} + \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\mu} + \boldsymbol{\gamma}_K) - \lambda \|\boldsymbol{\mu}\|_1 - \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1$$

$$= \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k \geqslant 0} \|\boldsymbol{\beta}_k\|_1$$

- glmnet returns $(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$, which is exactly the estimates $(-\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_K)$ we would obtain by performing MultinomRef0 with Data Shared lasso

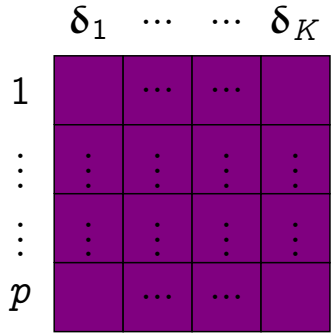
$\Rightarrow$ The symmetric formulation (glmnet) encourages homogeneity

Symmetric formulation



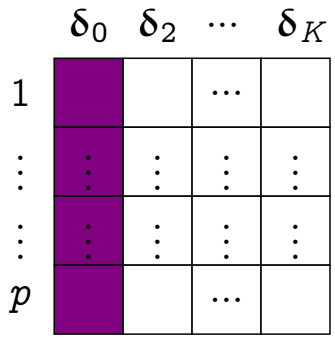
Standard formulation

with $\delta_k = \beta_k - \beta_0$



$C(0) = p(K-1)$

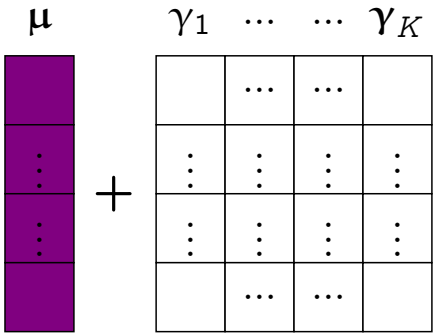
with $\delta_k = \beta_k - \beta_1$



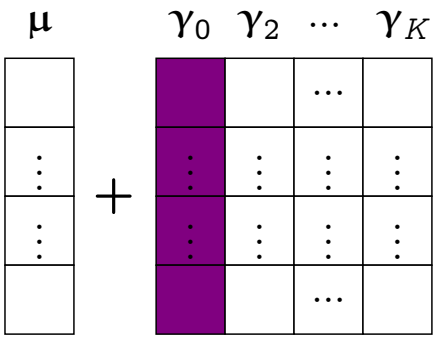
$C(1) = p$

Decomposition targeted by

MultinomDataSharedRef



$C(0) = p + 0 = p$



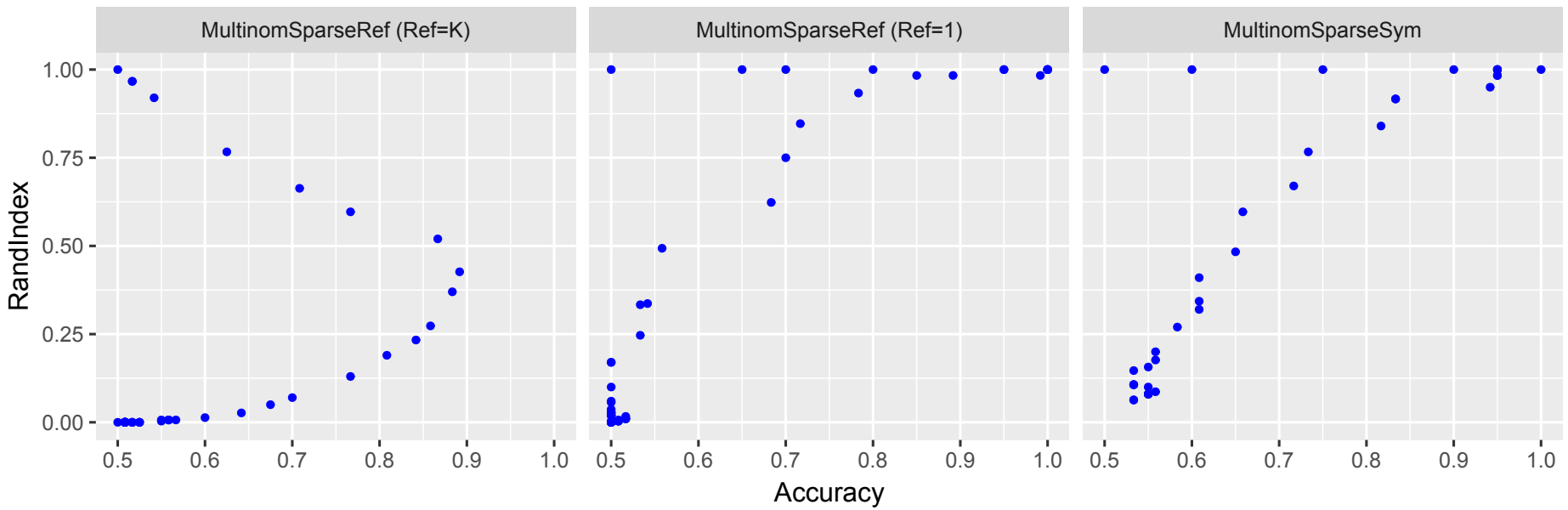
$C(1) = 0 + p = p$

International Agency for Research on Cancer

World Health Organization

A simple example in the homogenous case: $\beta_0^* = 0$, $\beta_1^* = \ldots = \beta_K^* \neq 0$

- We make λ vary and compute, for each value
  - Accuracy : accuracy in the identification of null/non-null entries in $B^* = (\beta_1^*, \ldots, \beta_K^*)$
  - Rand Index : $\sim$ accuracy in the identification of groups of disease subtypes for which any one given metabolite shows the same level of association.



Conclusion/Discussion

- Data Shared Lasso has some nice properties
  - easy to implement, under a variety of models ($\sim$ Lasso with interaction terms)
  - $\approx$ Oracle RefLasso regarding theoretical properties
  - $\approx$ A priori RefLasso regarding implementation
  - can partially describe the structure of the effects in situations where CliqueFused may fail to fully describe it.

$\Rightarrow$ Nice alternative/complement to "A priori RefLasso"

- Provides insights into the sparse multinomial regression models

CliqueFused

Back

- Penalized criterion:

$$\frac{\|\mathbf{y} - \mathbf{x}\mathbf{b}\|_2^2}{2n} + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{j_1 \sim j_2} |b_{j_1} - b_{j_2}|$$

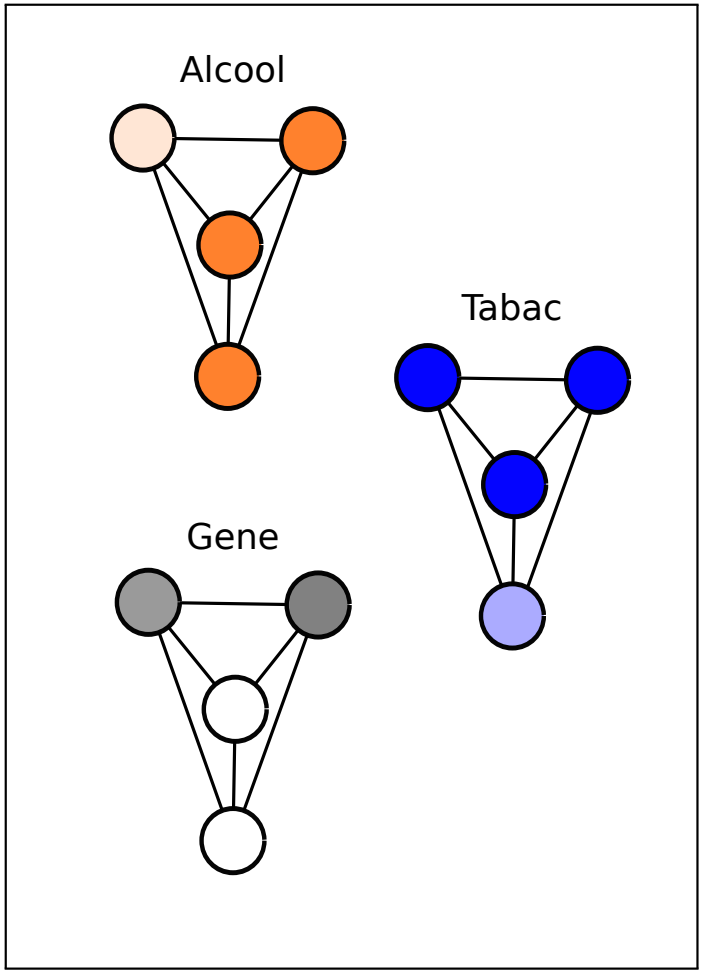
CliqueFused

Back

- Penalized criterion:

$$\frac{\|\mathbf{y} - \mathbf{x}\mathbf{b}\|_2^2}{2n} + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{j_1 \sim j_2} |b_{j_1} - b_{j_2}|$$

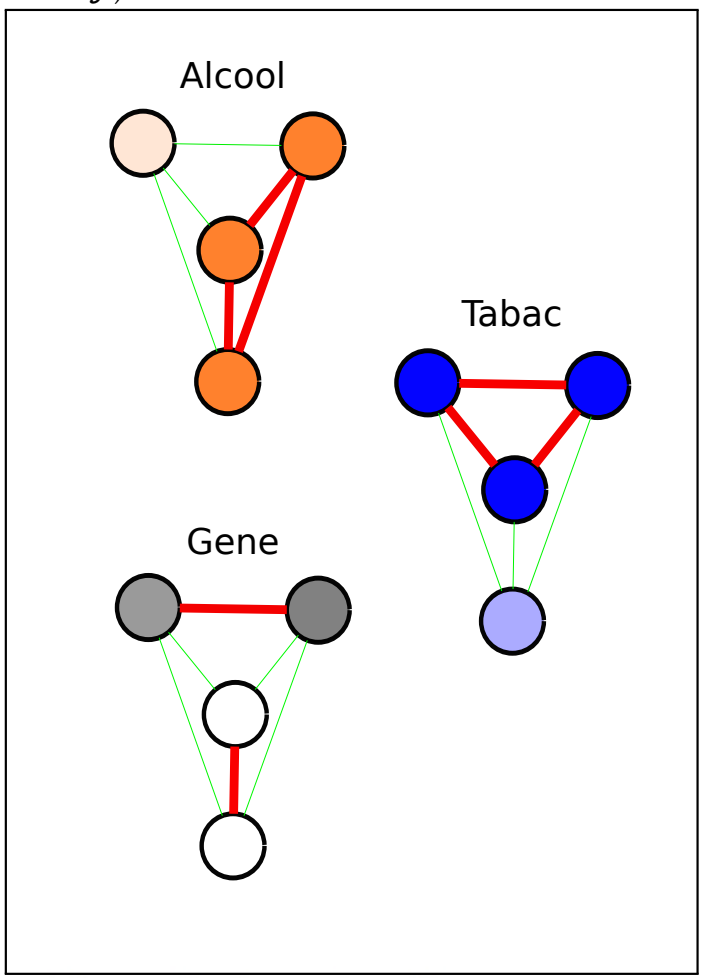
$\sum_{j,k>k'} |\beta_{k,j} - \beta_{k',j}|$



Adaptive Clique Fused (V. et al., Stat. Comp., 2016)

Back

$$\sum_{j,k>k'} v_{k,k',j}|\beta_{k,j} - \beta_{k',j}|$$

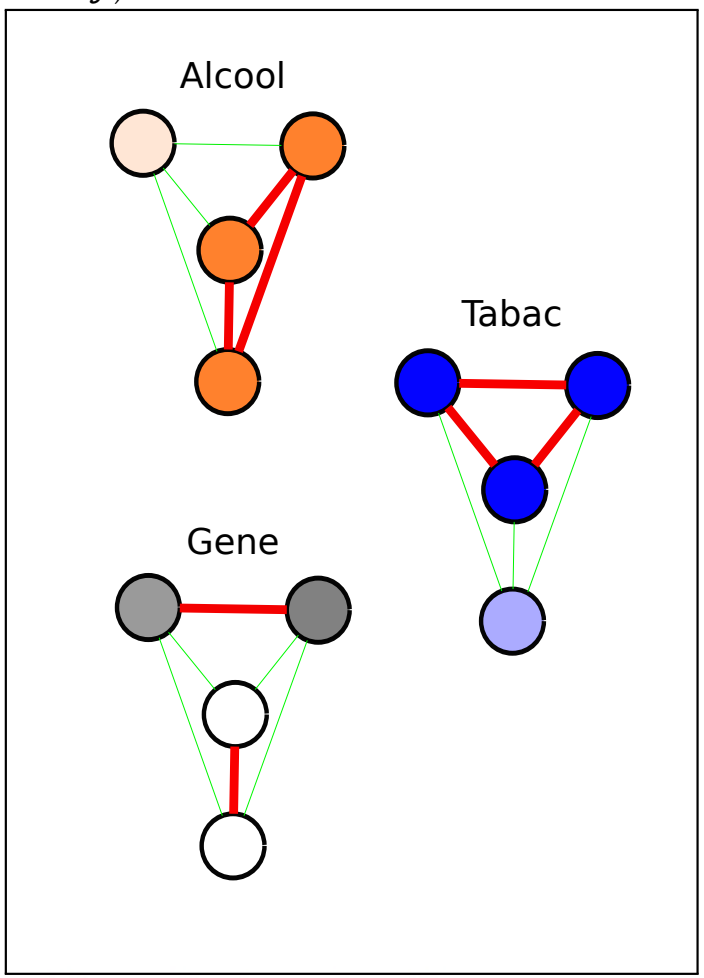


$$v_{k,k',j} = |\tilde{\beta}_{k,j} - \tilde{\beta}_{k',j}|^{-1}$$

Adaptive Clique Fused (V. et al., Stat. Comp., 2016)

Back

$$\sum_{j,k>k'} v_{k,k',j}|\beta_{k,j} - \beta_{k',j}|$$



$$v_{k,k',j} = |\tilde{\beta}_{k,j} - \tilde{\beta}_{k',j}|^{-1}$$

Corollaire 1

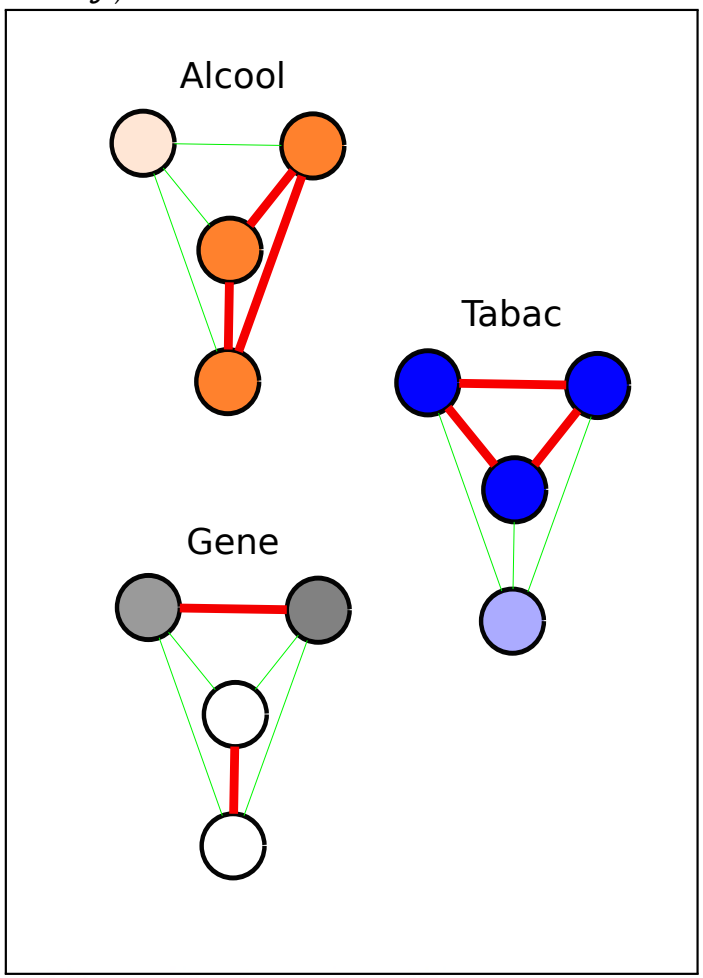
If Kp is fixed, and under some regularity conditions,

1. $\mathbb{P}(\textit{structure identified, for all } j) \to 1$;
2. $\sqrt{n}(\hat{\mathbf{b}}^{ad}_{\mathcal{A}} - \mathbf{b}^{*}_{\mathcal{A}}) \longrightarrow_d \mathcal{N}(\mathbf{0}_{s_0}, \sigma^2 \tilde{\mathbf{C}}^{-1}_{\mathcal{A}})$.

Adaptive Clique Fused (V. et al., Stat. Comp., 2016)

Back

$$\sum_{j,k>k'} v_{k,k',j}|\beta_{k,j} - \beta_{k',j}|$$



$$v_{k,k',j} = |\tilde{\beta}_{k,j} - \tilde{\beta}_{k',j}|^{-1}$$

Corollaire 1

If Kp is fixed, and under some regularity conditions,

1. $\mathbb{P}(\textit{structure identified, for all } j) \to 1$;
2. $\sqrt{n}(\hat{\mathbf{b}}^{ad}_{\mathcal{A}} - \mathbf{b}^{*}_{\mathcal{A}}) \longrightarrow_d \mathcal{N}(\mathbf{0}_{s_0}, \sigma^2 \tilde{\mathbf{C}}^{-1}_{\mathcal{A}})$.

- If $n_k \gg p$, Adaptive Clique Fused ok

Pénalités classiques en apprentissage multi-tâche

- (sparse) GroupLasso : soit $\boldsymbol{\beta}^j = (\beta_{1,j}, \ldots, \beta_{K,j})$, et

$$\text{pen}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \sum_j \|\boldsymbol{\beta}^{(j)}\|_2$$

- Nuclear(Trace)-Norm ($\sim$ reduced-rank regression). Soit $\mathbf{B}$ la matrice de dimension (p, K) avec $(\mathbf{B})_{k,j} = \beta_{k,j}$, et

$$\text{pen}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = tr(\sqrt{\mathbf{B}^T\mathbf{B}}).$$

Pénalités classiques en apprentissage multi-tâche

- (sparse) GroupLasso : soit $\boldsymbol{\beta}^j = (\beta_{1,j}, \ldots, \beta_{K,j})$, et

$$\text{pen}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \sum_j \|\boldsymbol{\beta}^{(j)}\|_2$$

- Nuclear(Trace)-Norm ($\sim$ reduced-rank regression). Soit $\mathbf{B}$ la matrice de dimension (p, K) avec $(\mathbf{B})_{k,j} = \beta_{k,j}$, et

$$\text{pen}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = tr(\sqrt{\mathbf{B}^T\mathbf{B}}).$$

$\Rightarrow$ Ok pour prédiction, mais pas adaptées à notre objectif d'identification des partitions (structure des effets) pour chaque covariable

- Soit $\tilde{\mathbf{X}}_{\boldsymbol{r}}^{(k)} = \mathbf{X}^{(k)}_{P_{\boldsymbol{r}}^{(k)}}$ où $P_{\boldsymbol{r}}^{(k)} = \{j \in [p] : k \neq r_j\}$
- Soit alors la matrice de taille $n \times Kp$:

$$\mathfrak{X}_{\boldsymbol{r}} = \begin{pmatrix} \mathbf{X}^{(1)} & \tilde{\mathbf{X}}_{\ell}^{(1)}/\tau_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & 0 & \dots & \tilde{\mathbf{X}}_{\ell}^{(K)}/\tau_K \end{pmatrix}$$

- Soit $\tilde{\mathbf{X}}_{\boldsymbol{r}}^{(k)} = \mathbf{X}^{(k)}_{P_{\boldsymbol{r}}^{(k)}}$ où $P_{\boldsymbol{r}}^{(k)} = \{j \in [p] : k \neq r_j\}$
- Soit alors la matrice de taille $n \times Kp$:

$$\mathfrak{X}_{\boldsymbol{r}} = \begin{pmatrix} \mathbf{X}^{(1)} & \tilde{\mathbf{X}}_{\ell}^{(1)}/\tau_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & 0 & \dots & \tilde{\mathbf{X}}_{\ell}^{(K)}/\tau_K \end{pmatrix}$$

de manière équivalente, on peut minimiser en $\boldsymbol{\theta}_{\boldsymbol{r}} \in \mathbb{R}^{Kp}$,

$$\frac{1}{2n}\|\mathcal{Y} - \mathfrak{X}_{\boldsymbol{r}}\boldsymbol{\theta}_{\boldsymbol{r}}\|_2^2 + \lambda_1\|\boldsymbol{\theta}_{\boldsymbol{r}}\|_1$$

qui retourne un estimateur de

$$\boldsymbol{\theta}^*_{\boldsymbol{r}} = (\boldsymbol{\mu}^{*T}_{\boldsymbol{r}}, \tau_1\boldsymbol{\gamma}^{*T}_{\boldsymbol{r},1}, \dots, \tau_K\boldsymbol{\gamma}^{*T}_{\boldsymbol{r},K})^T \in \mathbb{R}^{Kp}.$$

avec $\boldsymbol{\gamma}^{*T}_{\boldsymbol{r},k} = (\boldsymbol{\beta}^*_k - \boldsymbol{\mu}^*_{\boldsymbol{r}})_{P_{\boldsymbol{r}}^{(k)}}$