Méthodes pénalisées pour l'apprentissage supervisé en grande dimension ; applications en cancérologie et en traumatologie

Vivian Viallon, Centre International de Recherche sur le Cancer

Séminaires "Grosses données"

International Agency for Research on Cancer

World Health Organization

Contents

Supervised Learning and overfitting

Overview

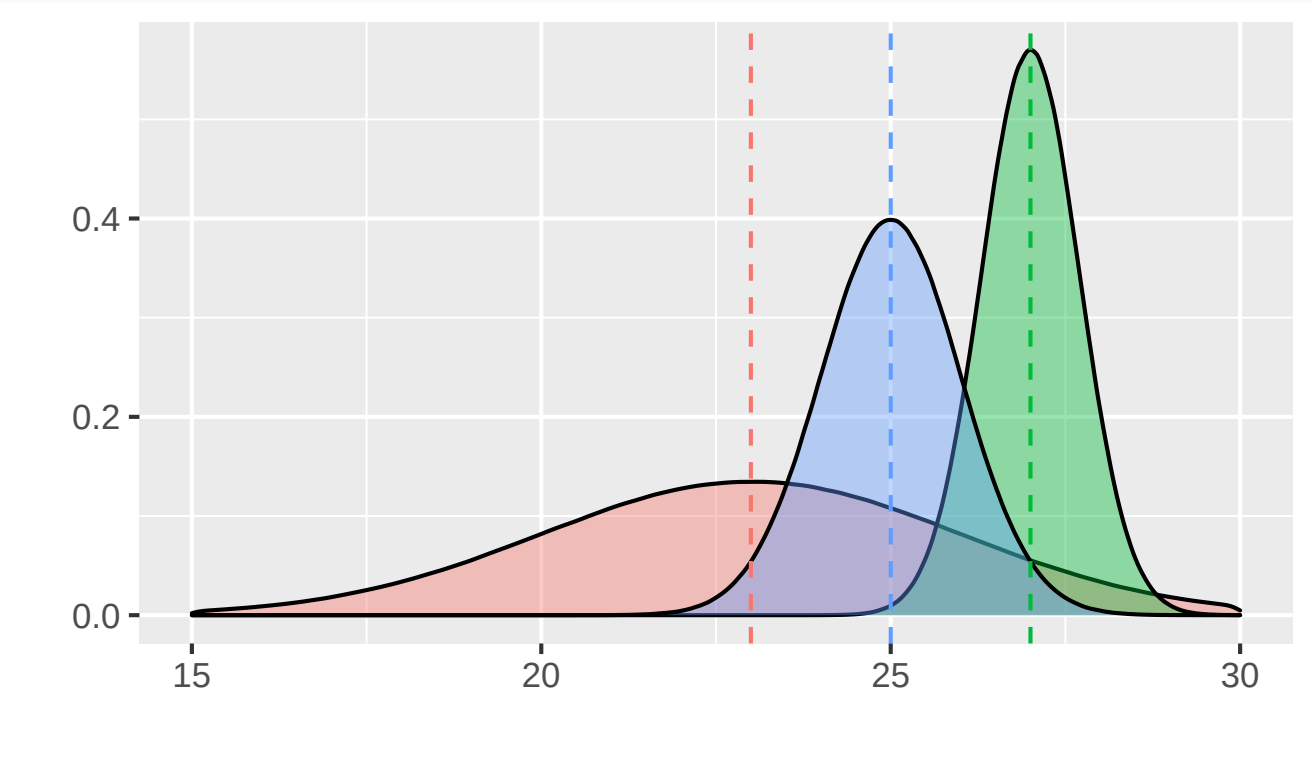
- y: outcome of interest
  - continuous: biomarker level, survival time, consumption of electricity, traffic, etc.
  - categorical/binary: diseased/"healthy", "qualifiers" of an image, spam/regular email, etc.

Overview

- y: outcome of interest
- What is the best prediction we can make for a new individual/observation?

Overview

- y: outcome of interest
- What is the best prediction we can make for a new individual/observation?
  - Intuition with a continuous y; example: biomarker level



Whole population High-caloric diet

High-caloric diet + low physical activity + "bad" genes + ...

Formalism

- **Data**
  - y: **continuous** outcome ($\sim$ **label**)
  - $\mathbf{x} = (x_1, \ldots, x_p)$: p features (predictors)

- **Objective: to find** the function f such that
  - for **"most"** $(\mathbf{x}, y)$, $f(\mathbf{x})$ best predicts y

Formalism

- **Data**
  - y: **continuous** outcome ($\sim$ **label**)
  - $\mathbf{x} = (x_1, \ldots, x_p)$: p features (predictors)

- **Objective: to find** the function f such that
  - for **"most"** $(\mathbf{x}, y)$, $f(\mathbf{x})$ best predicts y
  - $\mathbb{E}_{(\mathbf{x},y)}\{[y - f(\mathbf{x})]^2\}$ is minimized.

Formalism

- **Data**
  - y: **continuous** outcome ($\sim$ **label**)
  - $\mathbf{x} = (x_1, \ldots, x_p)$: p features (predictors)

- **Objective: to find** the function f such that
  - for **"most"** $(\mathbf{x}, y)$, $f(\mathbf{x})$ best predicts y
  - $\mathbb{E}_{(\mathbf{x},y)}\{[y - f(\mathbf{x})]^2\}$ is minimized.
  - Solution: $f(\mathbf{x}) = f^*(\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$: the **regression function**

Towards supervised learning

- But f^* unknown

Towards supervised learning

- But f^* unknown
- Solution:
  - Use a **training sample** $\mathcal{T}(n) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$,
  - and the approx. $\frac{1}{n}\sum_i \{y_i - f(\mathbf{x}_i)\}^2 \approx \mathbb{E}_{(\mathbf{x},y)}\{[y - f(\mathbf{x})]^2\}$

Towards supervised learning

- But f^* unknown
- Solution:
  - Use a **training sample** $\mathcal{T}(n) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$,
  - and the approx. $\frac{1}{n}\sum_i \{y_i - f(\mathbf{x}_i)\}^2 \approx \mathbb{E}_{(\mathbf{x},y)}\{[y - f(\mathbf{x})]^2\}$

- **Supervised Learning**

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{C}} \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i)\}^2$$

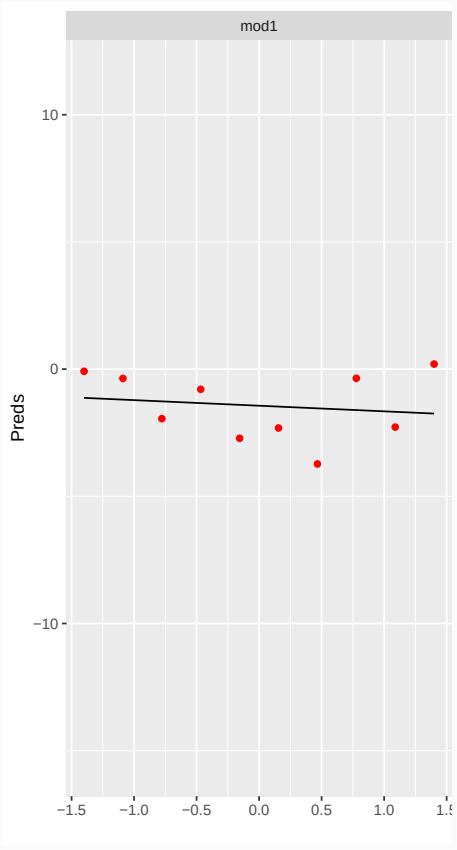
where $\mathcal{C}$ is a **given class of functions**

A first simple example: $\mathbf{X} = X \in \mathbb{R}$

- $\mathcal{C}(r) =$ polynomials of order r

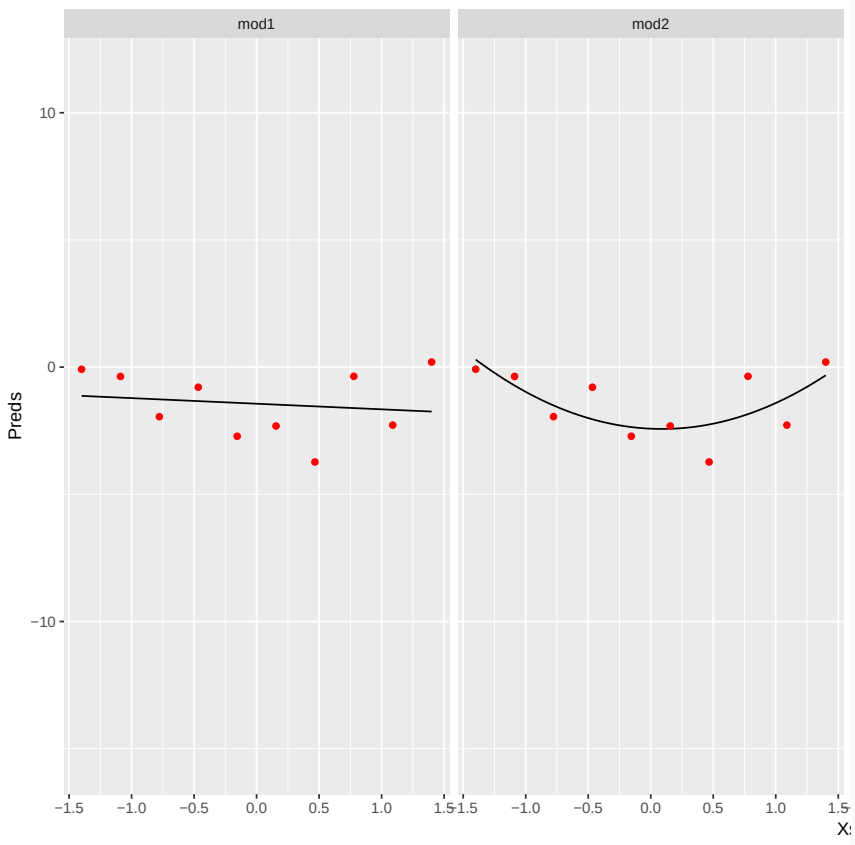
A first simple example: $\mathbf{X} = X \in \mathbb{R}$

- $\mathcal{C}(r) =$ polynomials of order r
- $\mathcal{C}(1) = \{f : f(x) = \beta_0 + \beta_1 x\}$



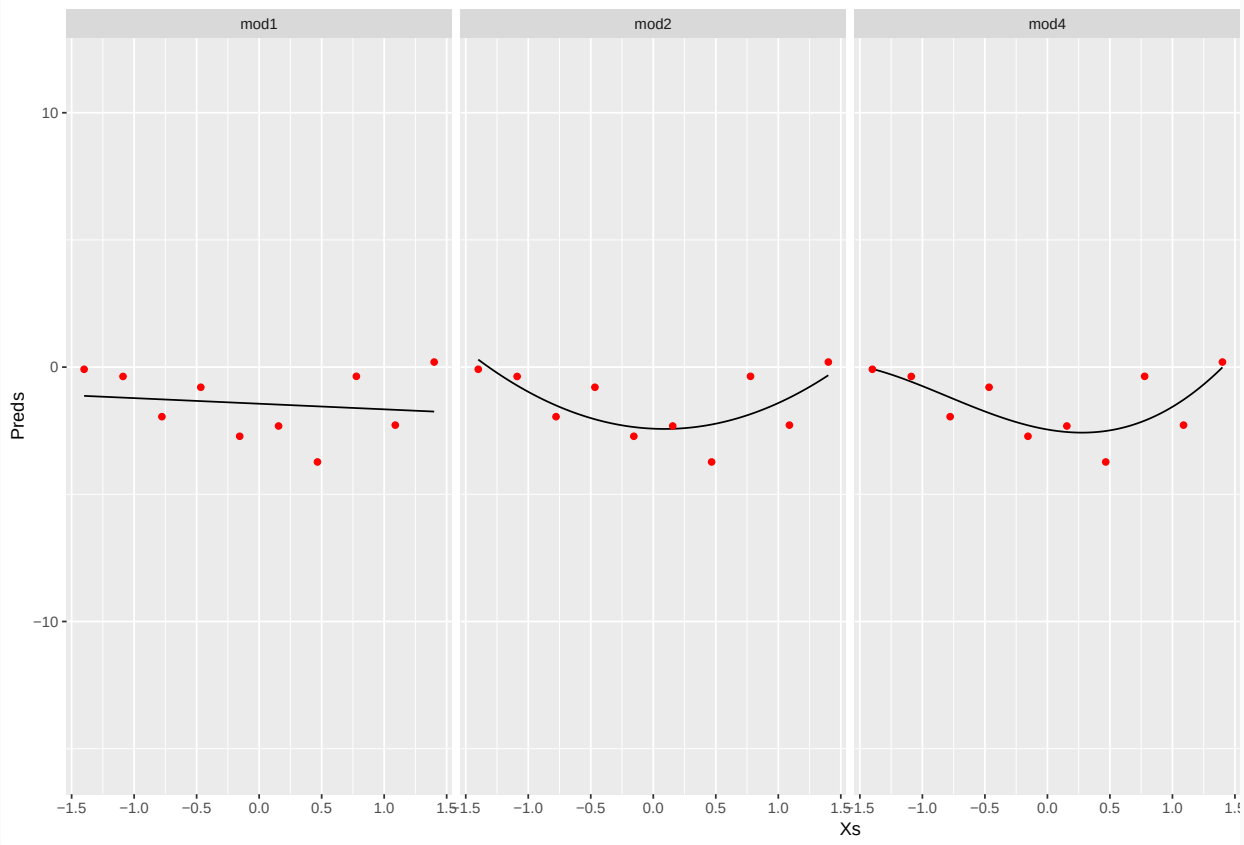
A first simple example: $\mathbf{X} = X \in \mathbb{R}$

- $\mathcal{C}(r) =$ polynomials of order r
- $\mathcal{C}(2) = \{f : f(x) = \beta_0 + \beta_1 x + \beta_2 x^2\}$



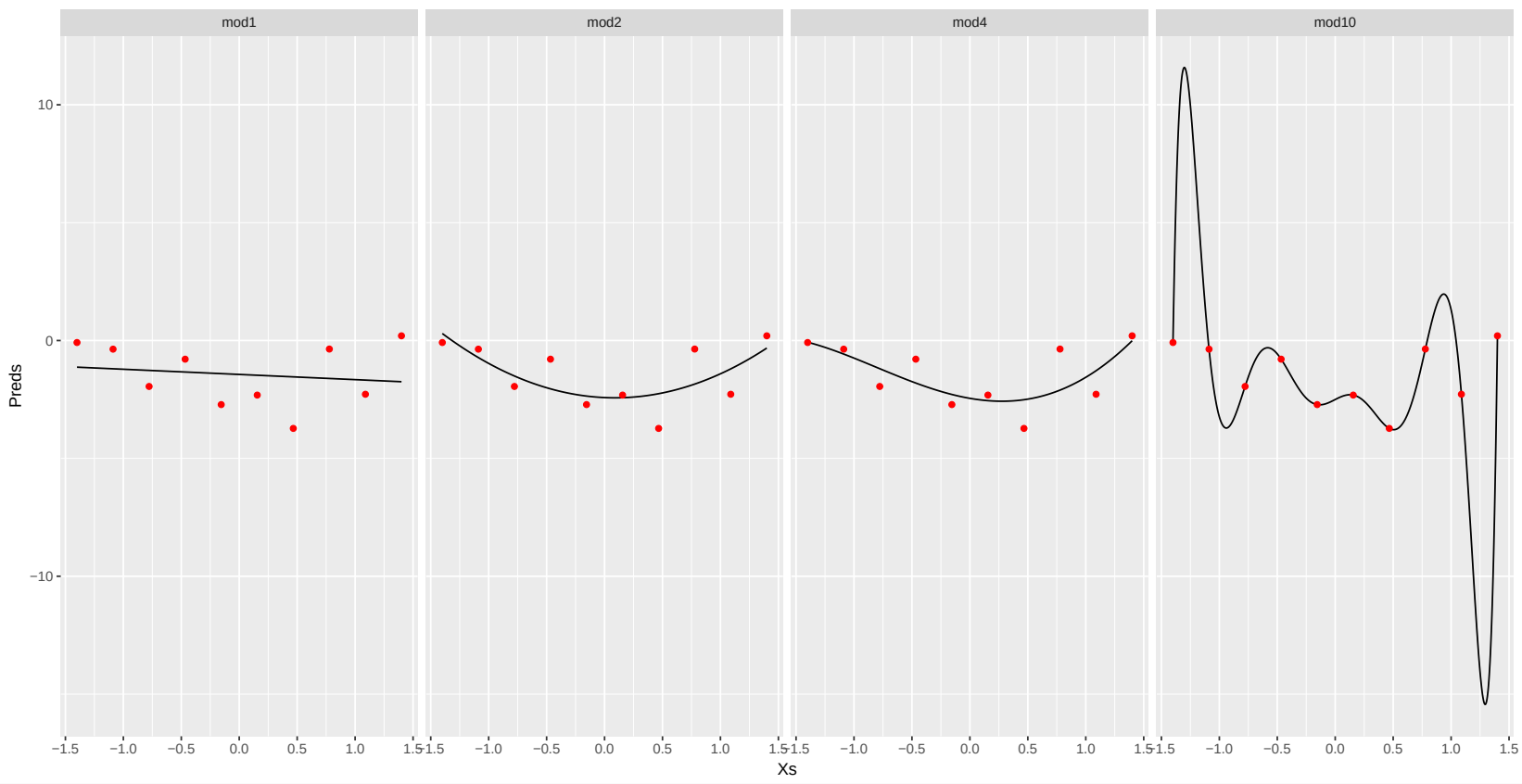
A first simple example: $\mathbf{X} = X \in \mathbb{R}$

- $\mathcal{C}(r) =$ polynomials of order r
- $\mathcal{C}(1) \subset \mathcal{C}(2) \subset \mathcal{C}(4)$



A first simple example: $\mathbf{X} = X \in \mathbb{R}$

- $\mathcal{C}(r) =$ polynomials of order r
- $\mathcal{C}(1) \subset \mathcal{C}(2) \subset \mathcal{C}(4) \subset \mathcal{C}(10)$

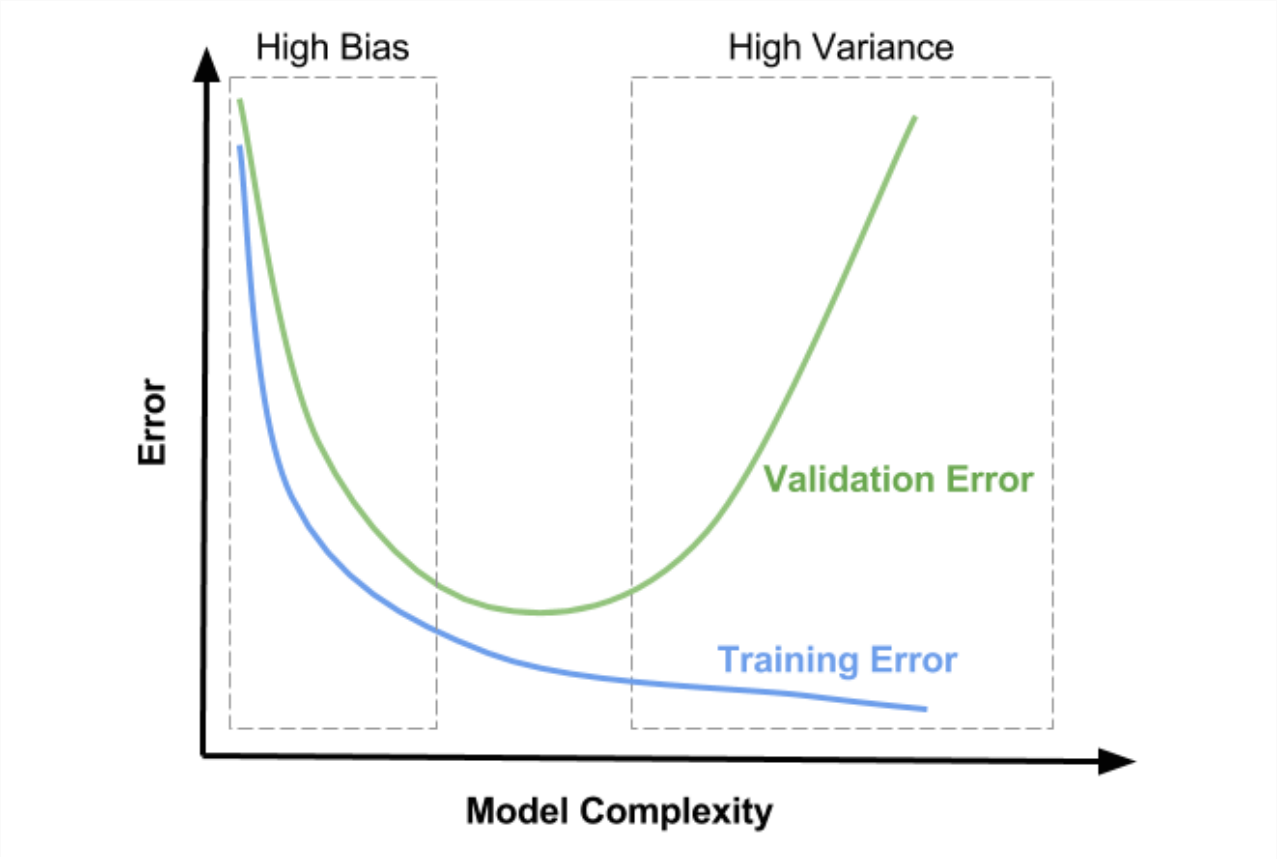


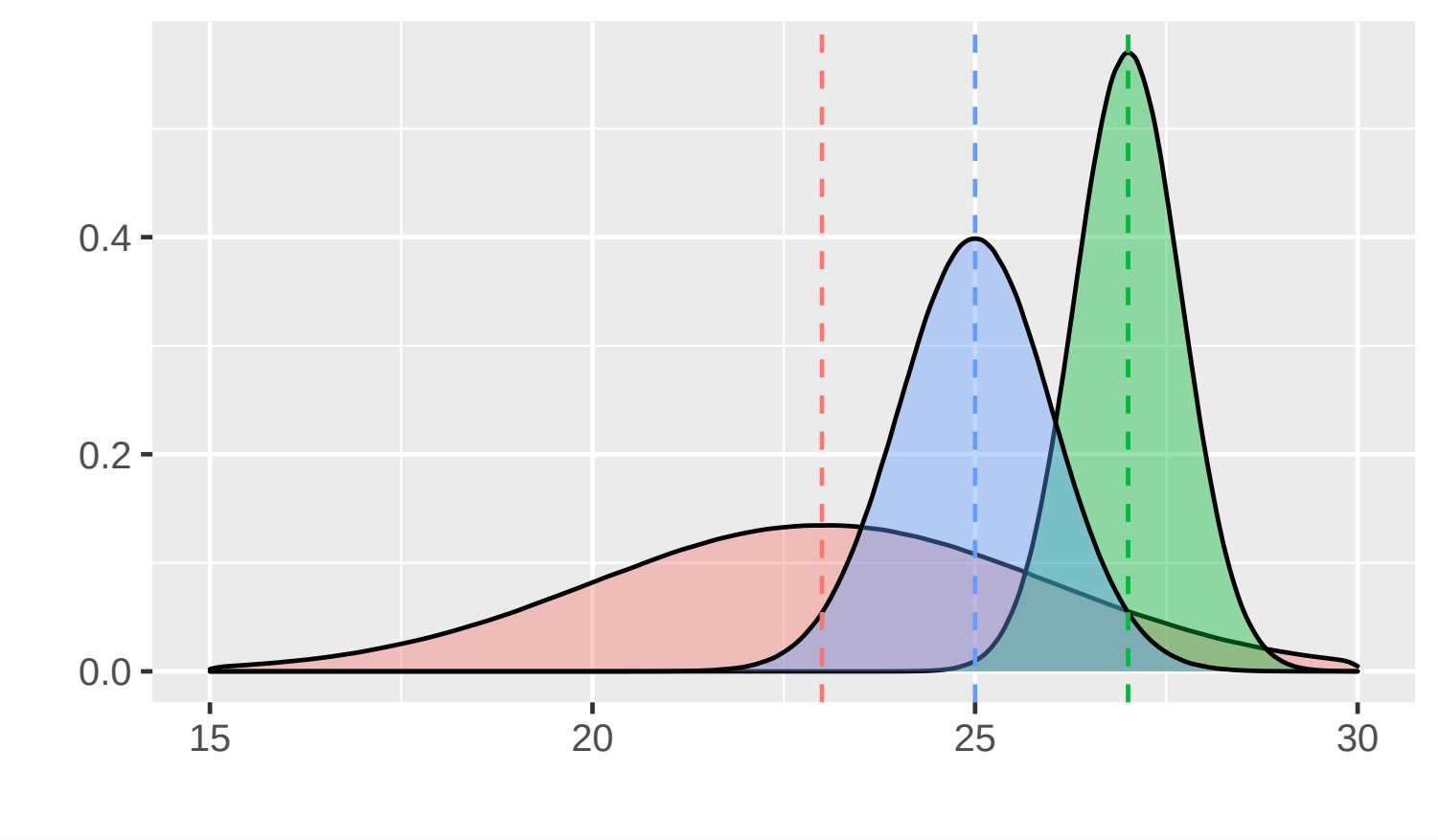
Underfitting (bias) / Overfitting (Variance)

- the more complex $\mathcal{C}$, the better the fit of $\hat{f}$ on $\mathcal{T}(n)$

$\Rightarrow$ the better $\hat{f}$?

- Does $\hat{f}(\mathbf{x})$ really best predict y?





Whole population

High-caloric diet

High-calorie diet + low physical activity + "bad" genes + ...

Summary

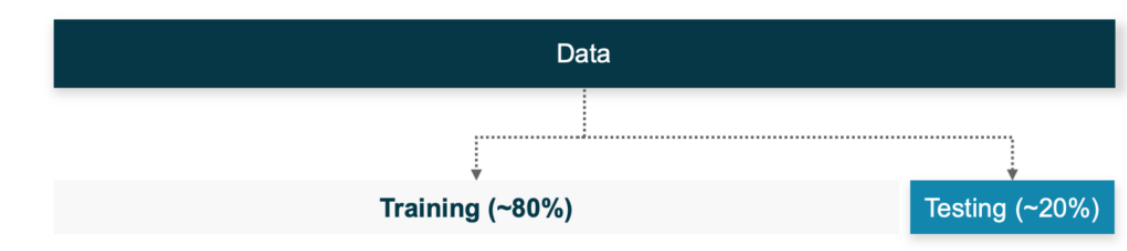
- **Supervised learning**: given $\mathcal{T}(n)$ and a new $\mathbf{x}_0$, predict y_0
- **Bias-variance tradeoff**: the model (class $\mathcal{C}$) should be complex enough to prevent underfitting, but not too complex to prevent overfitting.
- Optimal choice is data-dependent:
  - in particular, the larger n, the more complex the class can be
- Model selection: given $\mathcal{C}(1) \subset \mathcal{C}(2) \subset \ldots \subset \mathcal{C}(K)$, select the best one
  - by
    - fitting $\hat{f}^{(k)}$ (corresponding to model $\mathcal{C}(k)$) on $\mathcal{T}(n)$
    - evaluating each $\hat{f}^{(k)}$ on a validation sample $\mathcal{V}$, where available
  - by using cross-validation otherwise

Cross-validation

- CV is a way to "emulate" validation samples when no independent validation sample is available

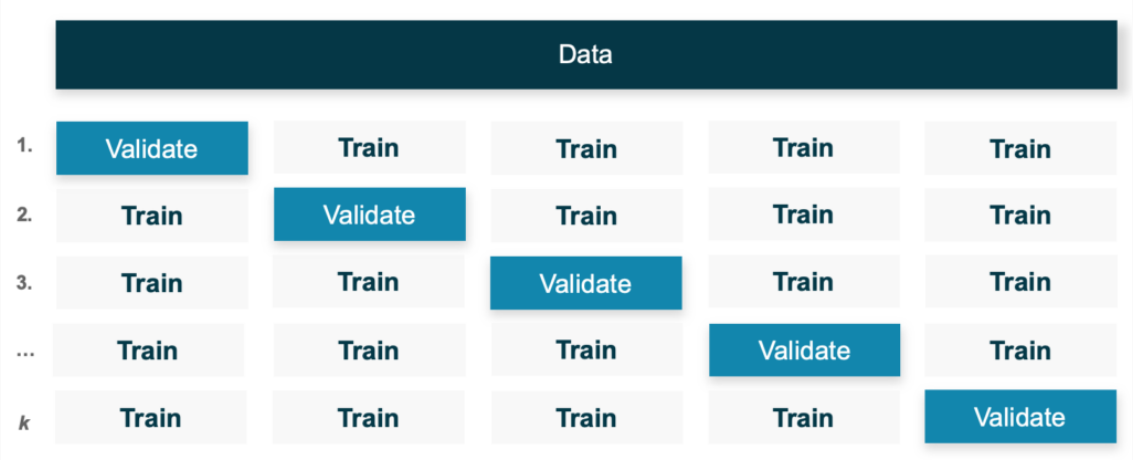
Cross-validation

- CV is a way to "emulate" validation samples when no independent validation sample is available



Cross-validation

- CV is a way to "emulate" validation samples when no independent validation sample is available



Penalized approaches for high-dimensional linear regression models

Linear regression model

- Consider $\mathcal{C} = \mathcal{C}^{(lin)}$ with

$$\mathcal{C}^{(lin)} = \{f : f(\mathbf{x}) = f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p = \sum_j \beta_j x_j = \mathbf{x}^T \boldsymbol{\beta}\}$$

Linear regression model

- Consider $\mathcal{C} = \mathcal{C}^{(lin)}$ with

$$\mathcal{C}^{(lin)} = \{f : f(\mathbf{x}) = f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p = \sum_j \beta_j x_j = \mathbf{x}^T \beta\}$$

  - not as restrictive as it looks: e.g., by augmenting the data

$$x_j = z_1 z_3 + z_2^2 + \sin(z_4) \times \exp(z_5)$$

Linear regression model

- Consider $\mathcal{C} = \mathcal{C}^{(lin)}$ with

$$\mathcal{C}^{(lin)} = \{f : f(\mathbf{x}) = f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p = \sum_j \beta_j x_j = \mathbf{x}^T \beta\}$$

- **Initial objective**: Use $\mathcal{T}(n) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ to find $\hat{f}$ by solving

$$\hat{f} = \underset{f \in \mathcal{C}^{(lin)}}{\text{argmin}} \sum_i \{y_i - f(\mathbf{x}_i)\}^2$$

Linear regression model

- Consider $\mathcal{C} = \mathcal{C}^{(lin)}$ with

$$\mathcal{C}^{(lin)} = \{f : f(\mathbf{x}) = f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p = \sum_j \beta_j x_j = \mathbf{x}^T \beta\}$$

- **Initial objective**: Use $\mathcal{T}(n) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ to find $\hat{\beta}$ by solving

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_i (y_i - \mathbf{x}_i^T \beta)^2$$

Linear regression model

- Consider $\mathcal{C} = \mathcal{C}^{(lin)}$ with

$$\mathcal{C}^{(lin)} = \{f : f(\mathbf{x}) = f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p = \sum_j \beta_j x_j = \mathbf{x}^T \beta\}$$

- **Initial objective**: Use $\mathcal{T}(n) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ to find $\hat{\beta}$ by solving

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_i (y_i - \mathbf{x}_i^T \beta)^2$$

$$\Rightarrow \hat{\beta} = \text{OLS estimator}$$

Overfitting in high-dimensional regression models: the curse of dimensionality

$$\mathbb{E}\{[y - \hat{f}(\mathbf{x})]^2\}$$
$$= \text{Incompressible term} + \text{Bias}(\hat{f}(\mathbf{x}))^2 + \text{Variance}(\hat{f}(\mathbf{x}))$$

- variance of OLS estimates : $\sim \min(p/n, 1)$ ...
- $\mathcal{C}^{(lin)}$ might be a too complex when p is large; $n \not\gg p$

Complexity of linear regression models

- For any given $\tau \geq 0$

$$\mathcal{C}^{(lin)}(\tau) = \{f^{(\beta)} : f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p \\ \text{s.t. } \mathrm{Pen}(\boldsymbol{\beta}) \leq \tau\}$$

- where $\mathrm{Pen}(\boldsymbol{\beta})$ is a measure of the complexity of $\boldsymbol{\beta} \in \mathbb{R}^p$
  - $\mathrm{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$: number of non-zero components of $\boldsymbol{\beta}$ ($\sim$ best subset regression)
  - $\mathrm{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
  - $\mathrm{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} |\beta_j|^2$
  - etc.

Complexity of linear regression models

- For any given $\tau \geq 0$

$$\mathcal{C}^{(lin)}(\tau) = \{f^{(\beta)} : f^{(\beta)}(\mathbf{x}) = \beta_1 x_1 + \ldots \beta_p x_p$$
$$\text{s.t. } \text{Pen}(\beta) \leq \tau\}$$

- where $\text{Pen}(\beta)$ is a measure of the complexity of $\beta \in \mathbb{R}^p$
  - $\text{Pen}(\beta) = \|\beta\|_0$: number of non-zero components of β ($\sim$ best subset regression)
  - $\text{Pen}(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
  - $\text{Pen}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{p} |\beta_j|^2$
  - etc.
- For any given $0 \leq \tau_1 \leq \ldots \leq \tau_R \leq \infty$

$$\mathcal{C}^{(lin)}(0) \subset \mathcal{C}^{(lin)}(\tau_1) \subset \ldots \subset \mathcal{C}^{(lin)}(\tau_R) \subset \mathcal{C}^{(lin)}(\infty) = \mathcal{C}^{(lin)}$$

From constrained optimization to penalized optimization

- Given $0 \leq \tau_1 \leq \ldots \leq \tau_R \leq \infty$;
- for each r, we aim to find

$$\hat{f}^{(r)} = \underset{f \in \mathcal{C}^{(lin)}(\tau_r)}{\operatorname{argmin}} \sum_i \{y_i - f(\mathbf{x}_i)\}^2$$

or equivalently,

$$\hat{\beta}^{(r)} = \underset{\beta \in \mathbb{R}^p : \mathrm{Pen}(\beta) \leq \tau_r}{\operatorname{argmin}} \sum_i (y_i - \mathbf{x}_i^T \beta)^2$$

From constrained optimization to penalized optimization

- Given $0 \leq \tau_1 \leq \ldots \leq \tau_R \leq \infty$;
- for each r, we aim to find

$$\hat{f}^{(r)} = \underset{f \in \mathcal{C}^{(lin)(\tau_r)}}{\operatorname{argmin}} \sum_i \{y_i - f(\mathbf{x}_i)\}^2$$

or equivalently,

$$\hat{\beta}^{(r)} = \underset{\beta \in \mathbb{R}^p : \mathrm{Pen}(\beta) \leq \tau_r}{\operatorname{argmin}} \sum_i (y_i - \mathbf{x}_i^T \beta)^2$$

or equivalently,

$$\hat{\beta}^{(r)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_i (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_r \mathrm{Pen}(\beta) \right\}$$

for some $\lambda_r = \lambda(\tau_r)$: $\infty \geq \lambda_1 \geq \ldots \geq \lambda_R \geq 0$

Penalized regression models

$$\hat{\beta}^{(r)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_i (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_r \mathrm{Pen}(\beta) \right\}$$

- Interpretation
  - goodness-of-fit ($\sim$ bias)
  - complexity of the model ($\sim$ variance)
  - selected, e.g., by cross-validation, etc.

Penalized regression models

$$\hat{\beta}^{(r)} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \sum_i (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_r \text{Pen}(\beta) \right\}$$

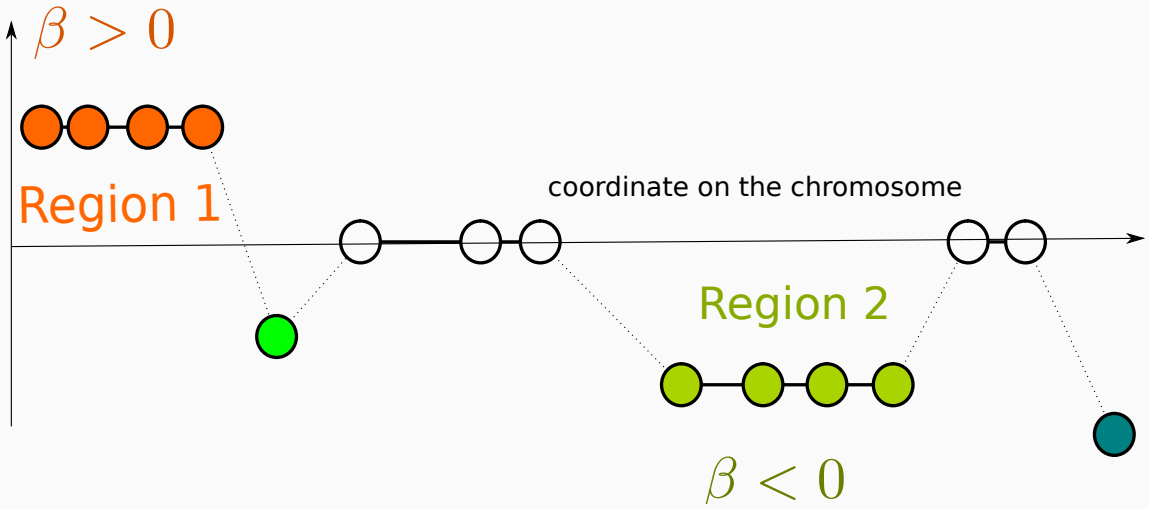
- Interpretation
  - goodness-of-fit ($\sim$ bias)
  - complexity of the model ($\sim$ variance)
  - selected, e.g., by cross-validation, etc.
- Special cases
  - $\text{Pen}(\beta) = \|\beta\|_0 \sim$ BIC: encourages sparsity but computationally impractical.
  - $\text{Pen}(\beta) = \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$: LASSO [Tibshirani, 1996, JRSS-B]: sparsity
  - $\text{Pen}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^p |\beta_j|^2$: RIDGE; no sparsity.

From sparsity to structured sparsity

- Some structure may exist among the predictors

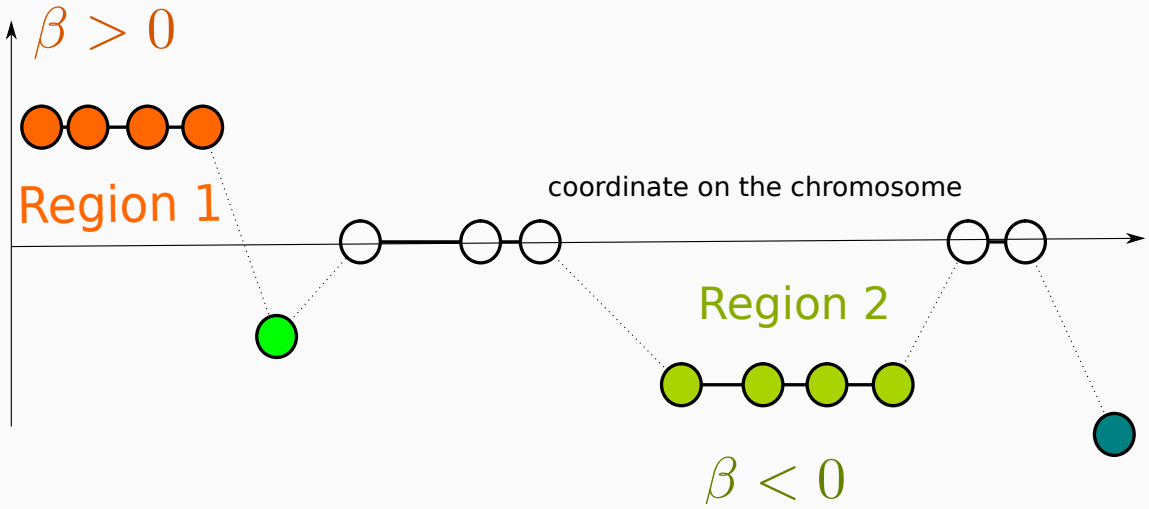
From sparsity to structured sparsity

- Some structure may exist among the predictors
- Epigenetic features are naturally ordered
  - Differentially methylated regions (DMRs) in relation to alcohol intake [Perrier et al., 2019, Clinical Epi.]



From sparsity to structured sparsity

- Some structure may exist among the predictors
- Epigenetic features are naturally ordered
  - Differentially methylated regions (DMRs) in relation to alcohol intake [Perrier et al., 2019, Clinical Epi.]

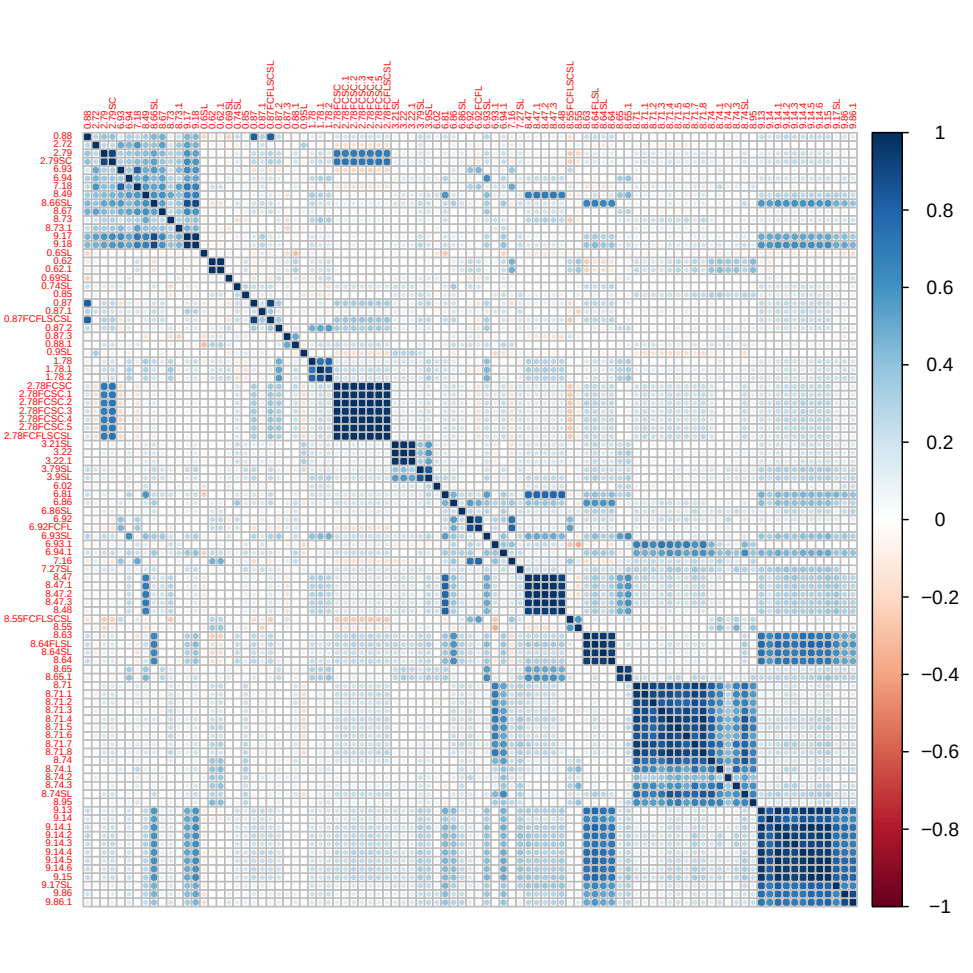


$\Rightarrow$ The fused lasso: selects regions

$$\text{Pen}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \sum_{j>1} |\beta_j - \beta_{j-1}|$$

$\Rightarrow$ **possibly** better interpretability, and better accuracy

Another example: untargeted metabolomics



⇒ Some groups appear:

- several features from the same metabolite (∼ variants of the same metabolite)
- several metabolites from the same nutrient, exposure, etc..

Group sparsity

- predefined groups of variables = "Extra"-information to be accounted for; e.g. via the group-lasso penalty: [Yuan and Lin, 2006, JRSS-B]

$$\boldsymbol{\beta} = (\underbrace{\beta_1, \ldots}_{\boldsymbol{\beta}_1}, \ldots, \underbrace{\ldots \beta_p}_{\boldsymbol{\beta}_G})$$

$$\hat{\boldsymbol{\beta}}(\lambda) \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\beta}) - \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 \right\}.$$

- Selection is performed:
  - at the variable level with the Lasso
  - at the group level with the group Lasso

Multi-task learning and subgroup analysis

Multi-task learning / subgroup analysis

- Subgroup analyses
  - the overall population = K predefined groups (or strata), based on “additional” covariates (e.g., gender, age categories)

- Multi-task learning
  - several "related" outcomes $Y_1, \ldots, Y_k$ (e.g., disease subtypes)

Example 1: Linear regression on stratified data

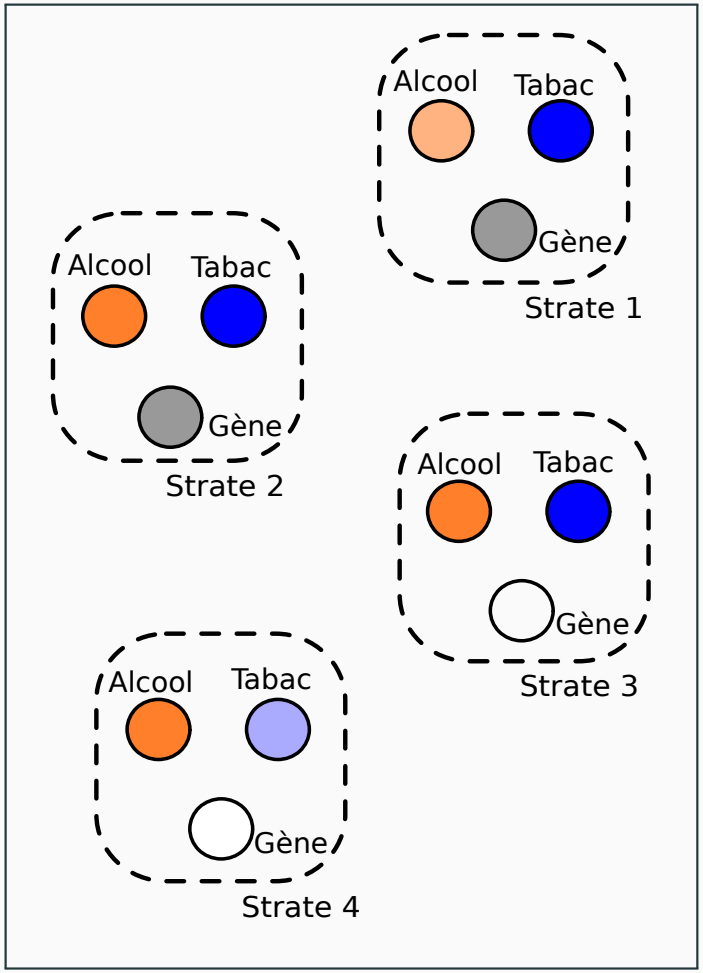
- Association between $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$ on K predefined strata; $Z = 1, \ldots, K$.

- k-th strata, $i = 1, \ldots, n_k$:

$$y_i^{(k)} = \mathbf{x}_i^{(k)^T} \beta_k^* + \xi_i^{(k)}$$

$\Rightarrow$ data shared lasso [Ballout et al., 2020, Biostatistics], or generalized fused lasso [V. et al., 2016, Stat. Comp.]

$$\mathrm{Pen}(\beta_1, \ldots, \beta_K) = \sum_k \|\beta_k\|_1 + \sum_{k<\ell} \|\beta_k - \beta_\ell\|_1$$



Example 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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. $Y_i^2 = 0$

Example 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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
- 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 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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
- The global study: K sub-studies
  1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \beta_1^*$
  2. m_2 pairs: Subtype 2 BC Vs Control
  3. ...
  4. m_K pairs: Subtype K BC Vs Control

Example 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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
- The global study: K sub-studies
  1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \beta_1^*$
  2. m_2 pairs: Subtype 2 BC Vs Control $\Rightarrow \beta_2^*$
  3. ...
  4. m_K pairs: Subtype K BC Vs Control

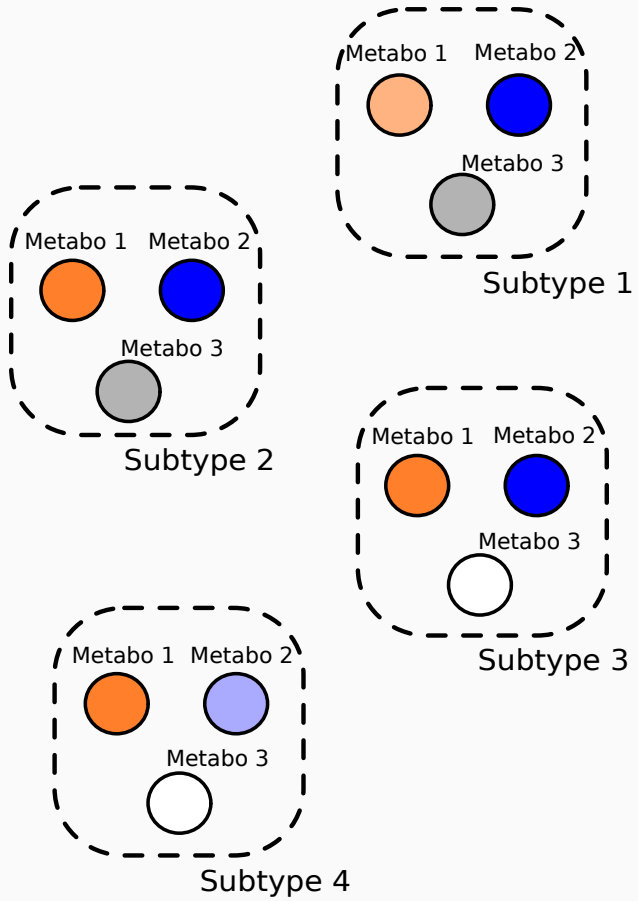
Example 2 : matched case-control studies

[Ballout et al., 2020, Biostatistics]

- $y \in \{0, 1, \ldots, K\}$
  - $y = 0$: control
  - $y = k > 0$: case, of subtype k.
- $m = n/2$ 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
- The global study: K sub-studies
  1. m_1 pairs: Subtype 1 BC Vs Control $\Rightarrow \beta_1^*$
  2. m_2 pairs: Subtype 2 BC Vs Control $\Rightarrow \beta_2^*$
  3. ...
  4. m_K pairs: Subtype K BC Vs Control $\Rightarrow \beta_K^*$

Expected structure in the parameter vector

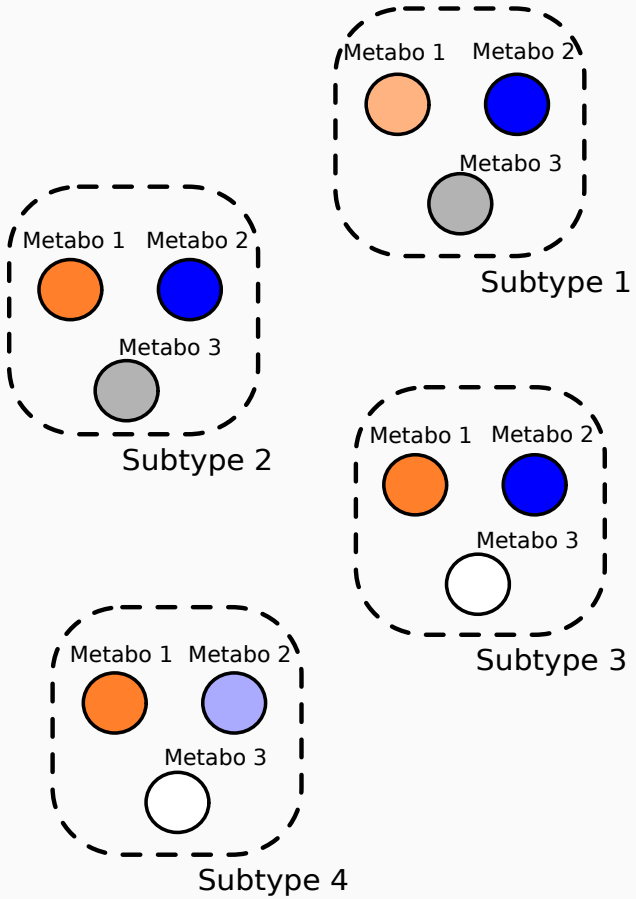
$$\mathbf{B}^* = (\beta_1^*, \ldots, \beta_K^*) \in \mathbb{R}^{Kp}$$

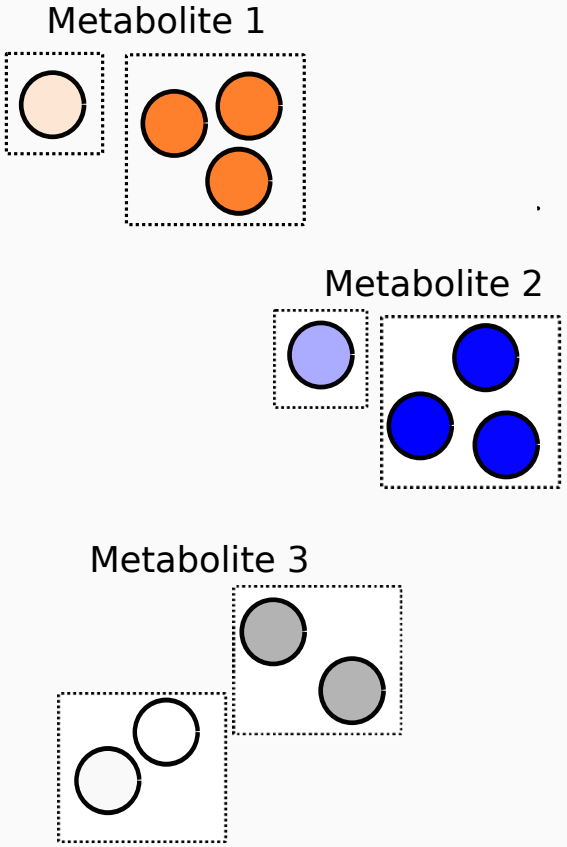


$$\text{Complexity} = \sum_k \|\beta_k^*\|_0 = 10$$

Expected structure in the parameter vector

$$\mathbf{B}^* = (\beta_1^*, \ldots, \beta_K^*) \in \mathbb{R}^{Kp}$$



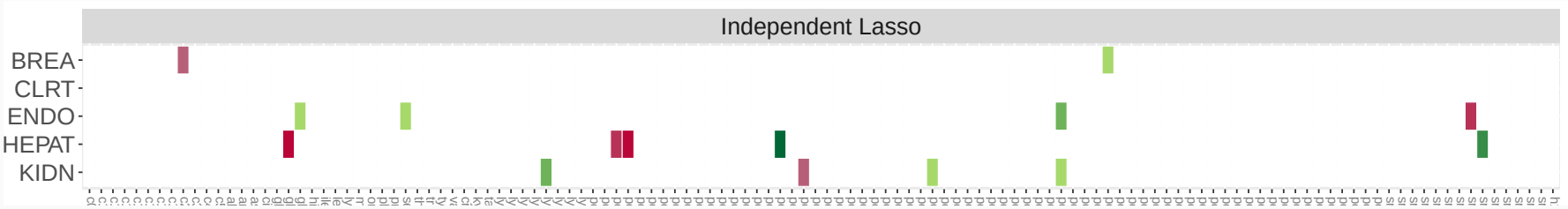


Complexity $= \sum_k \|\beta_k^*\|_0 = 10$

Complexity $= 5$

possibly better interpretability, and better accuracy

Metabolomics and cancer risk (preliminary)



International Agency for Research on Cancer

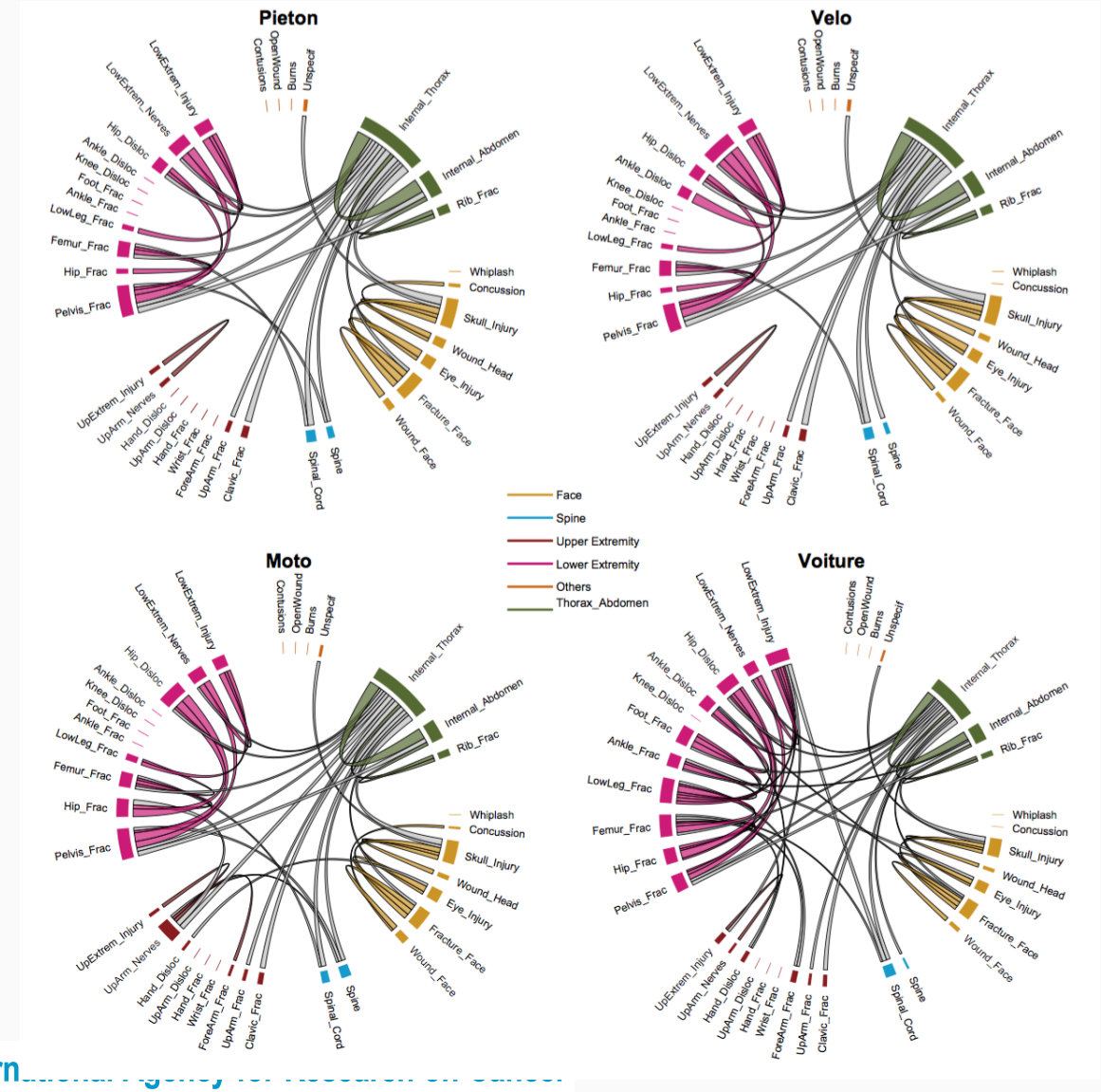
World Health Organization

Metabolomics and cancer risk (preliminary)

- Data shared lasso
  - identification of (potential) common patterns
  - identification of (more interpretable) heterogeneities

Example 3 : K binary graphical models

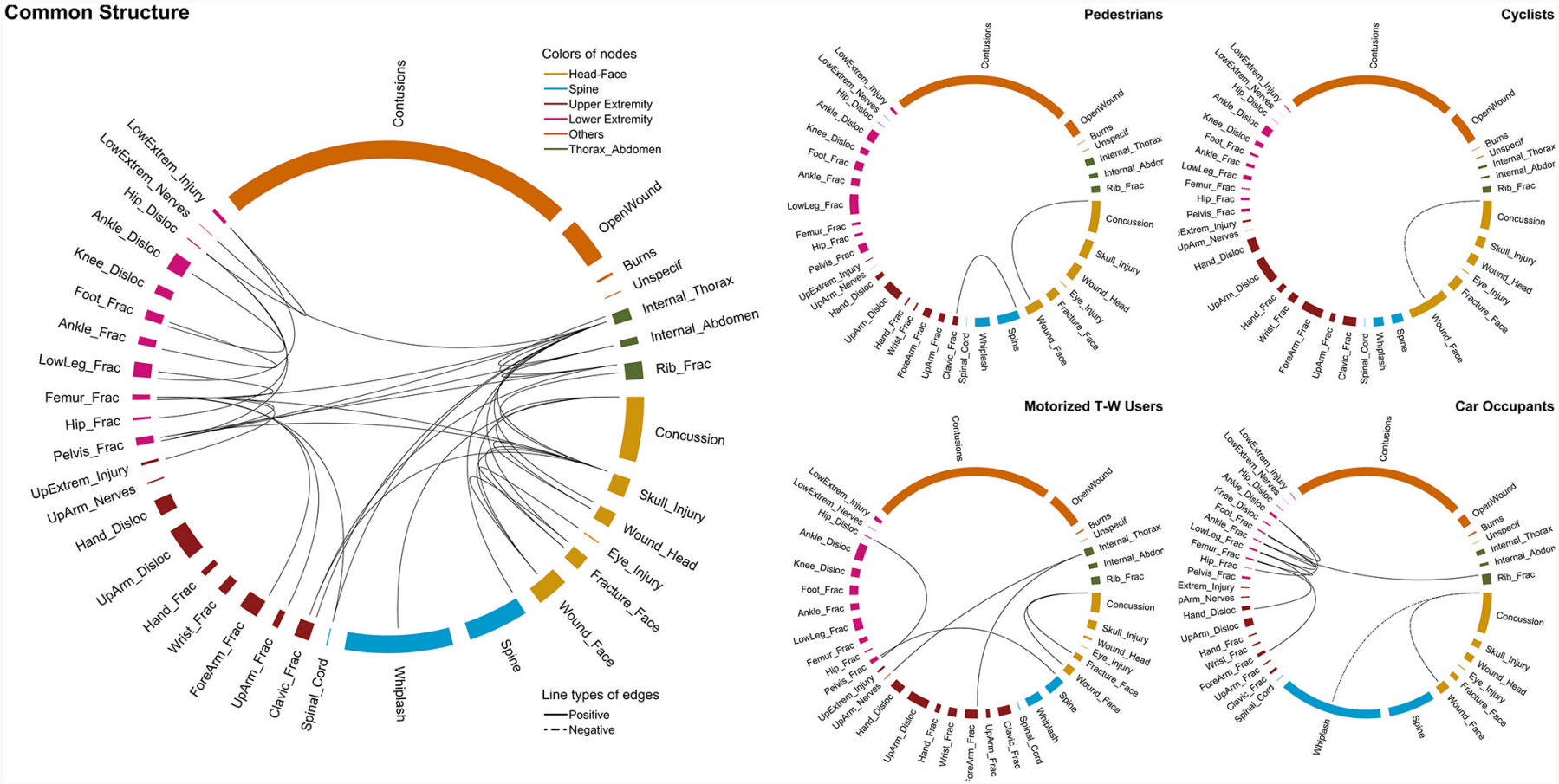
[Ballout and V., 2019, Statist. Med.]



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

Example 3 : K binary graphical models

[Ballout and V., 2019, Statist. Med.]



Discussion

Discussion

- High-dimension supervised learning is a difficult task
  - unless the true model is not too complex, or can be well approximated by not too complex of a model
  - appropriate methods are applied, and design matrices (predictors) are **"well conditioned"**
  - and/or we have large sample size

Discussion

- High-dimension supervised learning is a difficult task
  - unless the true model is not too complex, or can be well approximated by not too complex of a model
  - appropriate methods are applied, and design matrices (predictors) are **"well conditioned"**
  - and/or we have large sample size
- A related, and even more **complicated task: variable selection** ($\sim$ **etiology**)
  - We assumed throughout that $Y = f^*(\mathbf{X}) + \xi$
  - But X_j useful to predict Y
    - $\not\Rightarrow$ X_j is really associated with Y
    - $\not\Rightarrow$ X_j is a cause of Y
  - In particular, the "true" (or a better) model might be $Y = g^*(\boldsymbol{W}, \varepsilon)$.
    - $\boldsymbol{W}$ usually differs from $\mathbf{X}$