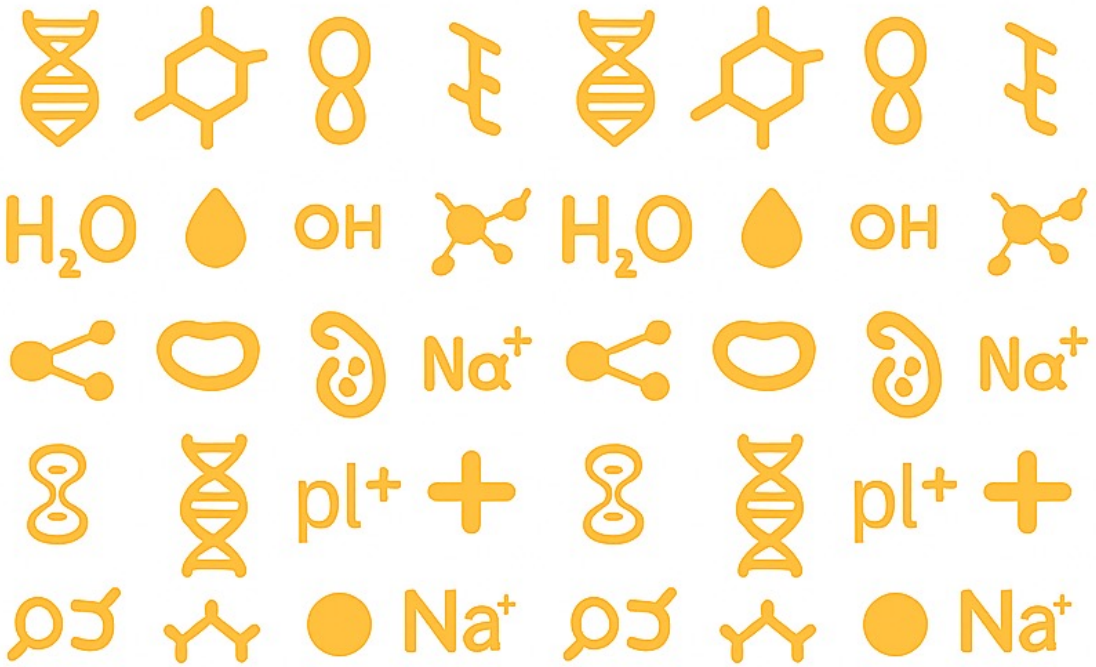
Selecting the most predictive biomarkers with penalized regression

Benedetta Marcozzi

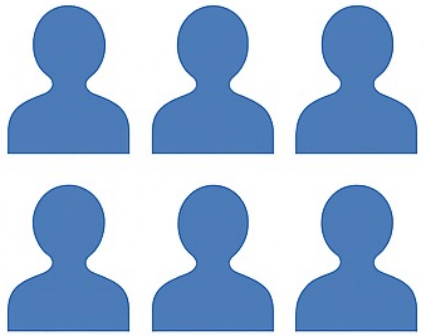
Perché usare la penalizzazione nei modelli di regressione?

Perché usare la penalizzazione nei modelli di regressione?

Molti biomarcatori



Pochi soggetti



Avere molti biomarcatori è un problema?

La regressione lineare tradizionale diventa poco affidabile quando il numero di biomarcatori è molto alto.

1. Variabili non-informative
    - non sono correlate all'outcome
    - riducono la capacità predittiva

2. Variabili ridondanti
    - molte feature «dicono la stessa cosa»
    - modelli più complessi e instabili

3. Instabilità dei coefficienti ($p >> n$)
    - pochi soggetti e molte variabili
    - piccole variazioni nei dati portano a grandi variazioni nei coefficienti
    - il modello diventa poco robusto

Avere molti biomarcatori è un problema?

La regressione lineare tradizionale diventa poco affidabile quando il numero di biomarcatori è molto alto.

1. Variabili non-informative
    - non sono correlate all'outcome
    - riducono la capacità predittiva

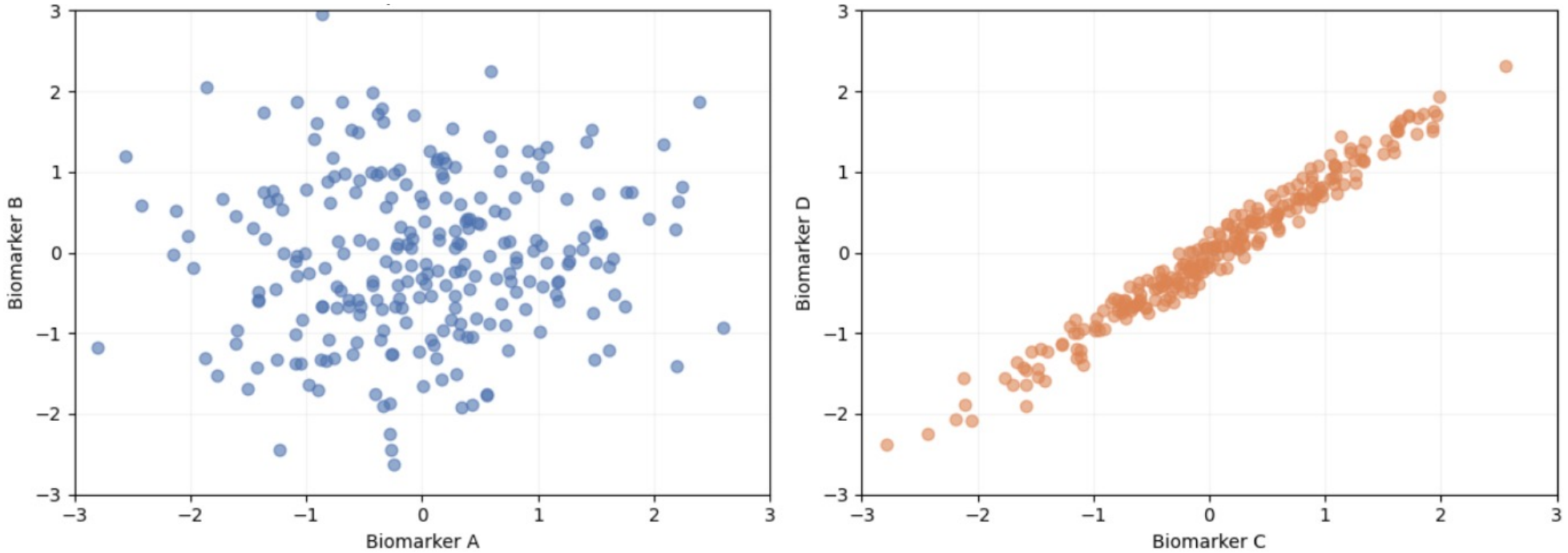
2. Variabili ridondanti
    - molte feature «dicono la stessa cosa»
    - modelli più complessi e instabili

3. Instabilità dei coefficienti (p >> n)
    - pochi soggetti e molte variabili
    - piccole variazioni nei dati portano a grandi variazioni nei coefficienti
    - il modello diventa poco robusto

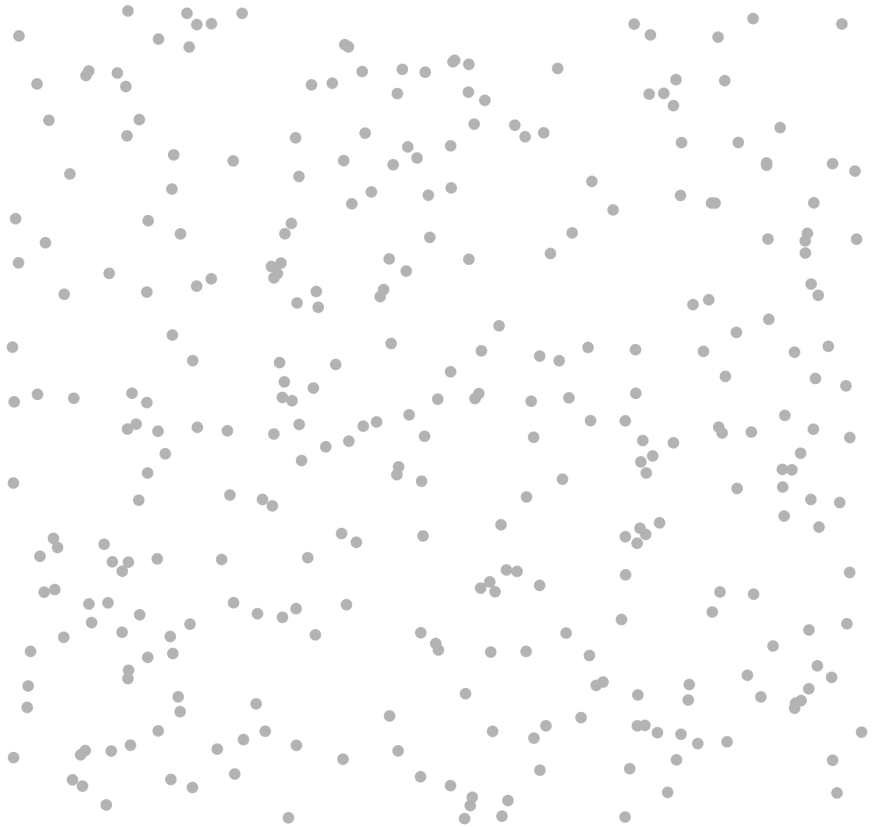
→ **Overfitting e scarsa capacità predittiva**

Avere molti biomarcatori è un problema?

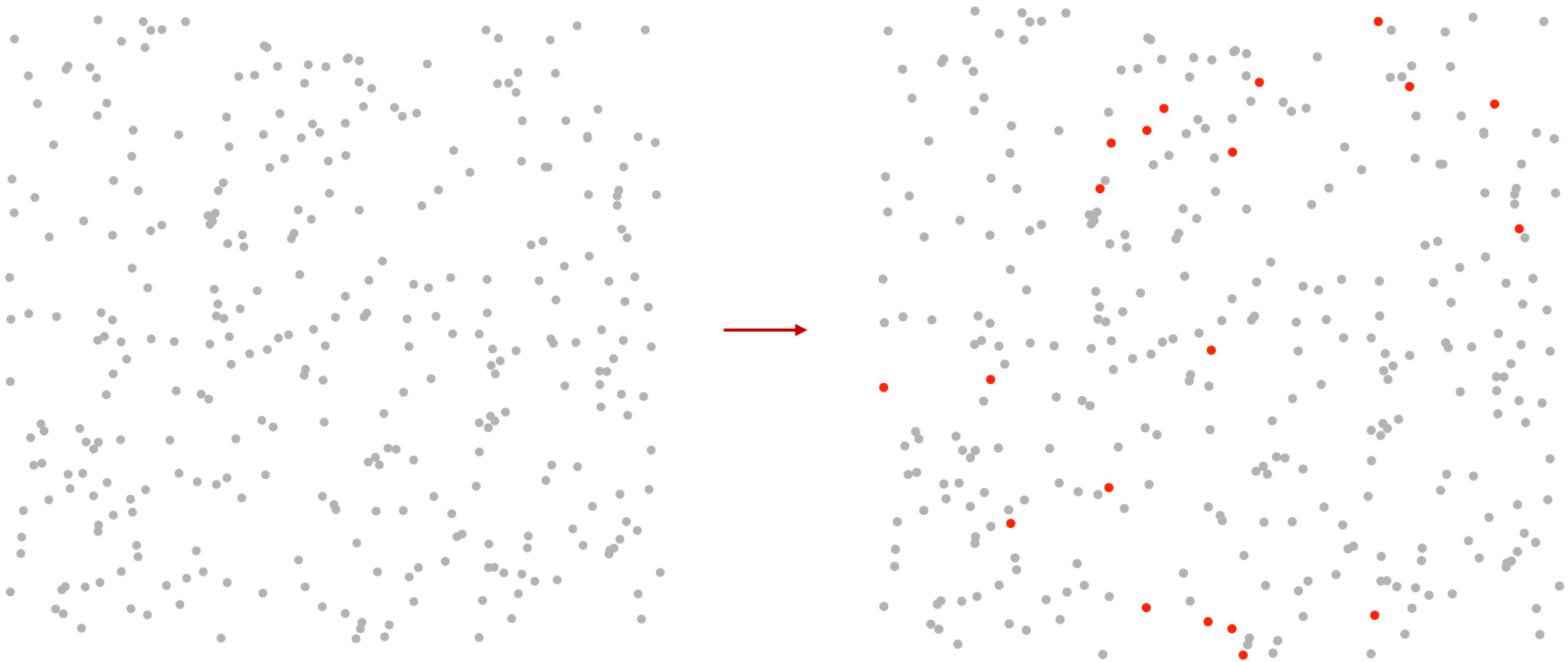
Quando due biomarcatori sono molto correlati portano la stessa informazione.
La regressione lineare non riesce a distinguere quale usare, e i coefficienti diventano instabili.



Molte variabili, poche davvero utili



Molte variabili, poche davvero utili



La penalizzazione

La penalizzazione è una tecnica usata nei modelli predittivi supervisionati per migliorare la capacità di generalizzazione.

In pratica cambiamo il comportamento di apprendimento del modello aggiungendo un “costo” alla complessità: quindi un coefficiente grande deve “pagare”.

$$\text{Loss} = \underbrace{\text{Errore di predizione}}_{\text{Fit}} + \lambda \cdot \underbrace{\text{Penalty sui coefficienti}}_{\text{Compl. modello}}$$

Errore di predizione

Misura **quanto le predizioni del modello sono vicine ai valori reali**.

Nella regressione lineare lo misuriamo con **MSE** (errore quadratico medio):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\underbrace{\text{Errore di predizione}}_{\text{Fit}} \longrightarrow \min_{w} \frac{1}{N}\sum_{i=1}^{N}(y_i - w^T x_i)^2$$

Penalità sulla complessità del modello

La penalità misura quanto il modello è complesso.

- cresce quando i coefficienti diventano grandi
- spinge il modello a usare valori più piccoli e stabili
- riduce rumore, instabilità e rischio di overfitting
- la sua forza è controllata da λ

$$\lambda \cdot \mathrm{Penalty}(w)$$

Effetto della penalità L1 – Lasso (Least Absolute Shrinkage and Selection Operator)

La penalità L1 spinge alcuni coefficienti esattamente a zero.
In pratica il modello sceglie quali variabili tenere e quali eliminare:
restano solo quelle che contribuiscono davvero alla predizione.

$$\min_{w} \ \frac{1}{N}\sum_{i=1}^{N}(y_i - w^T x_i)^2 \ + \ \lambda \sum_{j=1}^{p} |w_j|$$

Effetto della penalità L2 – Ridge

La penalità L2 non annulla i coefficienti, ma li riduce.
Nessuna variabile viene eliminata, ma nessuna può “dominare” con un coefficiente troppo grande.

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \sum_{j=1}^{p} w_j^2$$

Elastic Net combina gli effetti di L1 e L2

- la parte L1 può mettere a zero i coefficienti (selezione di variabili)
- la parte L2 stabilizza il modello quando le variabili sono correlate

Il risultato è un modello che seleziona le variabili più utili, ma allo stesso tempo rimane stabile anche in presenza di biomarcatori fortemente correlati.

$$\min_{w} \frac{1}{N}\sum_{i=1}^{N}(y_i - w^T x_i)^2 + \lambda_1 \sum_{j=1}^{p}|w_j| + \lambda_2 \sum_{j=1}^{p} w_j^2$$

Penalità sulla complessità del modello

Che cosa succede quando aumentiamo λ ?

Penalità sulla complessità del modello

Che cosa succede quando aumentiamo λ ?

λ grande → modello semplice

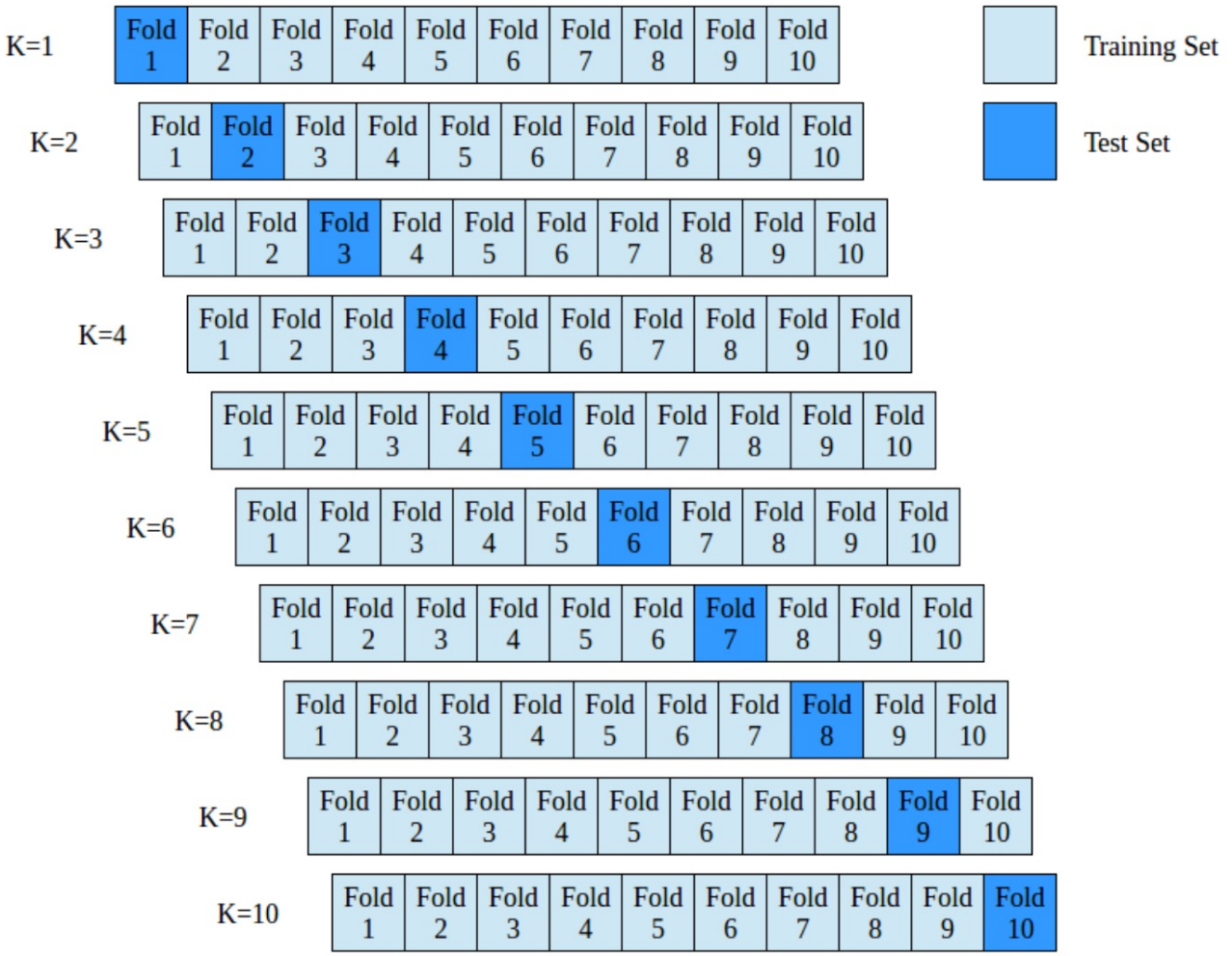
- i coefficienti vengono “schiacciati” verso zero
- riduzione drastica del rumore

λ piccolo → modello complesso

Vediamo un esempio pratico!

Come scegliere il parametro della penalizzazione?

k-fold cross-validation, con **k = 10**.



K-fold cross-validation method diagram (K = 10).

Grazie per l’attenzione!