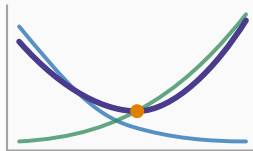


Apprentissage statistique

Chapitre 4 : La régression pénalisée : ridge et lasso



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Pourquoi pénaliser

La régression ridge

Le lasso

Ce qu'il faut retenir

Pourquoi pénaliser

Ce que font les MCO

Les moindres carrés minimisent la somme des carrés des résidus :

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Ils fournissent l'estimateur **sans biais** de variance minimale parmi les estimateurs linéaires — c'est le théorème de Gauss-Markov.

Sans biais n'est pas ce qu'on veut ici

Gauss-Markov garantit la variance minimale **parmi les estimateurs sans biais**. Rien n'interdit qu'un estimateur légèrement biaisé ait une variance bien plus faible, et donc une erreur totale inférieure.

C'est exactement ce que fait la pénalisation. Le chapitre 2 l'avait annoncé : **accepter du biais pour retrancher de la variance**.

Définition

1. **La multicolinéarité** — des variables très corrélées rendent les coefficients instables et énormes, de signes parfois absurdes.
2. **La grande dimension** — si $p > n$, la solution des MCO n'est même pas unique.

La régression ridge

Le critère ridge

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Le second terme **pénalise** les coefficients trop grands. $\lambda \geq 0$ règle la sévérité.

Définition	
λ	Effet
0	on retrouve les MCO
croissant	les coefficients rétrécissent vers zéro
$\rightarrow \infty$	tous tendent vers zéro, le modèle devient une constante
Le biais augmente avec λ , la variance diminue . Il existe donc un λ optimal, qu'on choisit par validation croisée .	

Intuition

Deux variables très corrélées peuvent recevoir des coefficients de $+500$ et -498 : leur somme est stable, chacun ne l'est pas.

Ridge interdit cette compensation en punissant le carré des coefficients. Il les **répartit** entre variables corrélées plutôt que d'en gonfler un.

Application en gestion

1. **Standardiser** les variables. La pénalité porte sur l'échelle : une variable en dirhams et une en milliers de dirhams seraient traitées différemment sans cela.
2. **Ne pas pénaliser la constante**. Elle ne mesure aucun effet, et la pénaliser rendrait le modèle dépendant du centrage de Y .

Ces deux points sont la source d'erreur la plus fréquente dans les premières mises en œuvre.

Le lasso

Une pénalité qui sélectionne

Le critère lasso

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

La seule différence est la **valeur absolue** au lieu du carré. Sa conséquence est considérable.

La différence décisive

	Pénalité	Coefficients
Ridge	$\sum \beta_j^2$	rétrécissent, jamais nuls
Lasso	$\sum \beta_j $	certain deviennent exactement nuls

Le lasso fait donc de la sélection de **variables**, ridge non.

Intuition

Géométriquement, la contrainte du lasso a des **coins** sur les axes — un losange en dimension deux — tandis que celle de ridge est un cercle, sans angle.

La solution touche la contrainte, et il est bien plus probable de toucher un losange à un coin qu'un cercle en un point particulier. Or être sur un coin signifie qu'une coordonnée est nulle.

Choisir entre les deux

Quand préférer l'un ou l'autre

Situation	Méthode
Beaucoup de variables faiblement utiles	ridge
Peu de variables vraiment utiles	lasso
Variables corrélées par blocs	elastic net
Modèle à interpréter et présenter	lasso

L'elastic net

Il combine les deux pénalités. Il sélectionne comme le lasso, tout en répartissant entre variables corrélées comme ridge.

Le lasso, entre deux variables très corrélées, en garde **une au hasard** et annule l'autre — ce qui rend son résultat instable d'un échantillon à l'autre. L'elastic net corrige ce défaut.

Application en gestion

Qu'une variable soit retenue par le lasso ne prouve **pas** qu'elle cause quoi que ce soit. Elle peut être retenue parce qu'elle est corrélée à la vraie cause, ou parce qu'elle est une conséquence.

Le lasso sélectionne pour **prévoir**, pas pour expliquer. Lire ses variables retenues comme des facteurs causaux est l'erreur que le chapitre 12, puis tout le cours d'*Inférence causale*, s'emploieront à écarter.

Ce qu'il faut retenir

Cinq points

1. Les MCO sont optimaux **parmi les sans-biais** — ce qui n'est pas le critère en prévision.
2. **Ridge** pénalise $\sum \beta_j^2$: les coefficients rétrécissent sans jamais s'annuler.
3. **Lasso** pénalise $\sum |\beta_j|$: certains deviennent exactement nuls, d'où une **sélection**.
4. λ se choisit par **validation croisée**; il faut **standardiser** et ne pas pénaliser la constante.
5. La sélection du lasso est **prédictive**, jamais causale.

La suite

Le lasso a introduit la sélection de variables. Le chapitre 5 examine ce que devient ce problème quand les variables sont plus nombreuses que les observations.