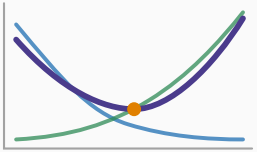
Apprentissage statistique

Chapitre 5 : Sélection de variables et grande dimension



Jaouad Madkour

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

Plan du chapitre

Sélectionner un sous-ensemble

Trois familles de méthodes

Les approches classiques

Méthode	Principe	Coût
Exhaustive	essayer tous les sous-ensembles	2^p modèles
Ascendante	partir de rien, ajouter la meilleure variable	$\approx p^2/2$
Descendante	partir de tout, retirer la moins utile	$\approx p^2/2$

Pourquoi l'exhaustive est impraticable

Exemple

p	Nombre de modèles
10	$1\,024$
20	$1\,048\,576$
40	plus de mille milliards

Vingt variables sont déjà à la limite ; quarante sont hors de portée.

Le défaut plus grave que le coût

Intuition

Explorer un grand nombre de modèles et retenir le meilleur revient à **choisir le plus chanceux**.

Avec assez de candidats, un sous-ensemble paraîtra excellent par pur hasard. C'est une forme de sur-ajustement qui porte non sur les paramètres mais sur le **choix du modèle** — et elle est invisible si l'on ne réévalue pas sur des données à part.

La sélection fait partie de l'apprentissage : elle doit être refaite à l'intérieur de chaque bloc de validation croisée, jamais une fois pour toutes en amont.

Pénaliser la complexité

Critère	Pénalité par paramètre	Tendance
R^2 ajusté	faible	modèles larges
AIC	2	orienté prévision
BIC	$\ln(n)$	plus sévère, modèles parcimonieux

Ces critères vous sont déjà connus

Le cours *Modélisation ARMA* les emploie au chapitre 13 pour choisir les ordres p et q.

L'idée est identique : l'ajustement seul favorise toujours le modèle le plus large, il faut donc lui opposer un prix. La **validation croisée** fait la même chose de manière directe, sans hypothèse de forme.

La grande dimension

Quand p dépasse n

Le régime p > n

Lorsque les variables sont plus nombreuses que les observations, les moindres carrés admettent une **infinité** de solutions ajustant parfaitement les données.

L'ajustement parfait n'a alors aucune valeur : il est atteint par construction, non par découverte.

Des situations fréquentes

- Texte : chaque mot est une variable, des dizaines de milliers.
- Génomique : des milliers de gènes, quelques centaines de patients.
- Finance : indicateurs techniques et fondamentaux démultipliés.
- Toute base enrichie d'interactions et de transformations.

Ce qui rend le problème soluble

L'hypothèse de **parcimonie** : seul un petit nombre de variables importe réellement.

Sans elle, rien n'est estimable. Avec elle, le **lasso** retrouve les bonnes variables sous certaines conditions — c'est le résultat qui a fait son succès.

Trois pièges de la grande dimension

Ce qui cesse d'être fiable

1. Les **p-values** et intervalles de confiance classiques, calculés après sélection, sont invalides — le modèle a été choisi sur les mêmes données.
2. La **multicolinéarité** est quasi certaine : avec p grand, des variables sont mécaniquement corrélées.
3. Le R^2 perd tout sens : il peut valoir 1 sur des données pures bruit.

L'illusion la plus simple à produire

Avec 100 observations et 200 variables **tirées au hasard**, sans aucun lien avec Y, on obtient un ajustement parfait.

Un praticien qui ne regarderait que le R^2 conclurait à un modèle excellent. L'erreur de test le détromperait immédiatement.

L'inférence après sélection

Intuition

Publier un coefficient et son écart-type après avoir choisi les variables sur les mêmes données produit des résultats sur-optimistes.

Des méthodes existent — inférence post-sélection, division de l'échantillon — mais la solution la plus sûre reste de **séparer les données** : sélectionner sur une moitié, estimer sur l'autre.

Réduire plutôt que sélectionner

Une autre voie

Régression sur composantes principales

Remplacer les p variables par un petit nombre de **composantes principales**, puis régresser sur celles-ci.

Vous savez déjà le faire

C'est l'ACP du cours d'**Analyse des données**, employée ici comme prétraitement.

Son défaut : les composantes sont construites pour résumer X, **sans regarder** Y. Rien ne garantit que les axes de plus forte variance soient les plus prédictifs. La **régression des moindres carrés partiels** corrige ce point en tenant compte de Y dans la construction des axes.

Sélection ou réduction

La **sélection** garde des variables d'origine : le modèle reste lisible.

La **réduction** crée des combinaisons : elle exploite toute l'information, au prix de l'interprétabilité.

Le choix dépend de ce qu'on fera du modèle — le présenter, ou seulement s'en servir.

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. La sélection **exhaustive** coûte 2^p modèles : impraticable au-delà de vingt variables.
2. Explorer beaucoup de modèles et garder le meilleur est une forme de **sur-ajustement**.
3. La sélection doit être **incluse dans la validation croisée**, jamais faite en amont.
4. Si $p > n$, l'ajustement parfait est automatique ; seule la **parcimonie** rend le problème soluble.
5. Après sélection, **p-values et R^2 ne sont plus fiables**.

La suite

Tout ce qui précède supposait une cible quantitative. Le chapitre 6 passe à la classification — et c'est la situation de notre portefeuille de crédit.