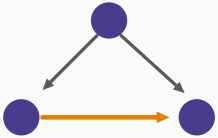
Inférence causale et machine learning

Chapitre 10 : Le double machine learning



Jaouad Madkour

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

Plan du chapitre

Le problème que le machine learning peut résoudre

La procédure

Pourquoi ces deux précautions

Ce que cela permet, et ce que cela ne permet pas

Ce qu'il faut retenir

Le problème que le machine learning peut résoudre

Ce qu'il apporte, et ce qu'il n'apporte pas

Une clarification préalable

L'apprentissage automatique **n'identifie rien**. Le chapitre 3 l'a dit : il ne choisit pas les contrôles, il ne valide aucune hypothèse.

Ce qu'il apporte est ailleurs, et c'est considérable : il permet d'estimer des **fonctions de nuisance** de grande dimension, là où les méthodes paramétriques échouaient.

Fonction de nuisance

Une **nuisance** est une quantité qu'il faut estimer pour atteindre le paramètre d'intérêt, sans qu'elle présente elle-même d'intérêt.

Dans le modèle partiellement linéaire

$$Y = \tau T + g(X) + \varepsilon \qquad\qquad T = m(X) + \nu$$

τ est l'effet cherché ; g et m sont des nuisances.

Pourquoi c'est un vrai progrès

Application en gestion

L'économétrie classique impose une forme à g — souvent linéaire — parce qu'elle ne sait pas faire autrement.

Si la vraie relation ne l'est pas, l'erreur de spécification **contamine** l'estimation de τ. Avec vingt ou cent covariables et des interactions inconnues, cette contamination est certaine.

Le machine learning estime g et m **sans forme imposée**. C'est exactement ce pour quoi il est bon.

La procédure

Trois ingrédients

Étape 1 — orthogonaliser

Régresser Y sur X, et T sur X, par n'importe quelle méthode d'apprentissage. Conserver les **résidus** :

$$\tilde{Y} = Y - \hat{g}(X) \qquad \tilde{T} = T - \hat{m}(X)$$

Étape 2 — régresser les résidus

$$\hat{\tau} = \text{régression de } \tilde{Y} \text{ sur } \tilde{T}$$

C'est le théorème de Frisch-Waugh

Ce résultat est enseigné en économétrie : régresser Y sur T et X donne le même coefficient que régresser les résidus l'un sur l'autre.

Le double machine learning en est la version non paramétrique. L'idée n'est pas neuve ; ce qui l'est, c'est de remplacer les régressions auxiliaires par des méthodes flexibles.

Étape 3 — le fractionnement croisé

Définition

Estimer $\hat{g}$ et $\hat{m}$ sur une partie de l'échantillon, calculer les résidus sur **l'autre**, et alterner les rôles.

Pourquoi ces deux précautions

Elles ne sont pas facultatives

Le rôle de l'orthogonalisation

Les résidus $\tilde{Y}$ et $\tilde{T}$ sont ce qui reste **une fois X retiré des deux côtés**.

Cette construction rend l'estimateur **peu sensible** aux petites erreurs sur $\hat{g}$ et $\hat{m}$: leurs effets se compensent au premier ordre. C'est la propriété de **Neyman-orthogonalité**.

Le rôle du fractionnement

Sans lui, le même échantillon sert à estimer la nuisance **et** l'effet.

Or une méthode flexible sur-ajuste : $\hat{g}$ capte une part du bruit propre à ces observations. Ce sur-ajustement se transmet aux résidus, donc à $\hat{\tau}$, sous forme de **biais**.

Le fractionnement croisé le supprime : chaque observation voit ses résidus calculés par un modèle qui **ne l'a jamais vue**.

Application en gestion

C'est exactement le principe du chapitre 3 d'*Apprentissage statistique* — ne jamais évaluer sur ce qui a servi à estimer.

Ici, l'enjeu n'est plus une mesure de performance honnête, mais l'**absence de biais** du paramètre d'intérêt. La même discipline, pour une raison différente.

Les deux ensemble, et pas l'une sans l'autre

Définition

Précaution	Ce qu'elle traite
Orthogonalisation	la sensibilité aux erreurs de nuisance
Fractionnement croisé	le biais dû au sur-ajustement

Omettre l'une des deux fait perdre les garanties. C'est ce qui distingue le double machine learning d'une simple régression avec des variables prédites.

Ce que cela permet, et ce que cela ne permet pas

Le bilan

Ce qu'on obtient

Un estimateur de τ **convergent**, asymptotiquement normal, avec des écarts-types valides — malgré des nuisances estimées par des méthodes non paramétriques.

C'est un résultat théorique important : l'inférence redevient possible après usage du machine learning.

Les hypothèses n'ont pas changé

Le double machine learning suppose toujours l'**indépendance conditionnelle** du chapitre 2.

Si un confondant est **non observé**, aucune flexibilité n'y remédie. La méthode gère la **dimension** de X, pas son **exhaustivité**.

C'est la mise en garde la plus importante du chapitre. Une procédure sophistiquée sur des contrôles insuffisants reste une estimation biaisée — simplement mieux habillée.

Où il s'applique le mieux

Intuition

Le double machine learning se combine avec les stratégies des chapitres 5 à 9 : il existe des versions pour les variables instrumentales, les doubles différences, les scores de propension.

Dans chaque cas, il prend en charge la partie **estimation** et laisse intacte la partie **identification** — qui reste l'affaire du graphe causal et de l'argument institutionnel.

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. Le machine learning estime les **nuisances**, pas l'identification.
2. **Orthogonaliser** : régresser Y et T sur X, puis les résidus l'un sur l'autre — Frisch-Waugh non paramétrique.
3. Le **fractionnement croisé** supprime le biais dû au sur-ajustement des nuisances.
4. Les deux précautions ensemble donnent un estimateur **convergent** avec inférence valide.
5. L'**indépendance conditionnelle** reste requise : la dimension est gérée, pas l'exhaustivité.

La suite

Jusqu'ici, on a estimé un effet **moyen**. Le chapitre 11 exploite le machine learning pour ce qu'il sait faire de plus utile ici : mesurer comment l'effet **varie** d'un individu à l'autre.