Machine Learning pour l'économétrie : éléments de correction des questions

Christophe Gaillac
Université d'Oxford & Nuffield College
CREST

Jérémy L'Hour
Capital Fund Management
CREST

Version : 27 décembre 2023

4 Inférence post-sélection

1. cf. section 4.2.
2. cf. section 4.3.
3. Idéalement, on souhaiterait pouvoir calculer des intervalles de confiance et faire des tests sur ce paramètre d'intérêt. Or, le fait que la distribution asymptotique de l'estimateur Lasso ne soit pas connue complique la construction de ces quantités.
4. Le biais de régularisation est un biais qui se produit en raison du fait de l'utilisation d'outils ML en première étape produit des estimateurs qui ne convergent pas assez rapidement. Dans le cas du Lasso, il s'agit d'un biais de variable omise. Il peut exister dans un cas de petites dimensions si un estimateur non conventionnel est utilisé en première étape ou s'il y a une étape de sélection. (Pensez au modèle de Leeb et Potscher).
5. cf. section 4.5.
6. L'argument de Leeb et Potscher est que l'inférence après une étape de sélection est plus compliquée qu'il n'y paraît, car les estimateurs post-sélection ne bénéficient pas de « bonnes propriétés » (normalité asymptotique par exemple). Le théorème 5.1 n'est pas contradictoire : c'est une solution au problème. Il montre que dans les cas où l'estimateur est immunisé, l'inférence basée sur la distribution normale est toujours possible. Dans de nombreux cas, cela signifie ajouter une autre étape de sélection (d'où le nom de "double sélection").

5 Généralisation et méthodologie

1. Non, la procédure de double sélection est la clé. On note toutefois que le Lasso est « suffisamment bon » pour être utilisé dans les premières étapes.

2. Voir la remarque du chapitre 5. On utilise le partitionnement de l'échantillon.
3. On utilise le Lasso lorsque l'on est prêt à supposer une structure linéaire pour laquelle les coefficients sont parcimonieux (seulement un petit nombre de non-zéros). On préfère utiliser une forêt aléatoire si l'on est prêt à supposer que la fonction de régression est constante par morceaux.
4. Cela supprime le biais de sur-ajustement, mais le coût est la diminution de la taille de l'échantillon et le temps de calcul (secondaire) lorsque l'on utilise l'ajustement croisé.
5. Dans la section 3.4/4.b (étude de simulation du chapitre HD et endogénéité). Expliquez (brièvement) pourquoi.
6. C'est également une procédure de sélection de variables qui ne fait usage que d'une seule équation, elle est donc sujet au problème de l'inférence post-sélection

6 Grande dimension et endogénéité

1. Soit la liste des instruments disponibles et possibles est grande, alors que le chercheur sait que seuls quelques-uns d'entre eux sont pertinents ; mais surtout, même lorsqu'on ne dispose que d'un seul instrument Z, on peut aussi envisager des transformations de l'instrument initial

$$(f_1(Z), \dots, f_p(Z))$$

   par une famille de fonctions $(f_i)_{i=1}^p$, ce qui nous ramène au cas de très nombreux instruments possibles. Dans le cours, nous utilisons un modèle parcimonieux pour les variables instrumentales, en supposant que seuls quelques instruments sont effectivement utiles, et nous fournissons une méthode basée sur Lasso pour estimer le traitement tout en contrôlant l'estimation du paramètre de nuisance.
2. Sans le terme $E[D|X]$ dans (6.10), on peut calculer par exemple que $\partial_\nu M(\tau_0, \eta_0) = \tau_0 \, \mathbb{E}[X(Z'\delta_0 + X'\gamma_0)]$, qui est différent de 0 en général.
3. On utilise la formule de σ_Γ^2 donnée en (6.15), et on obtient que

$$\begin{aligned}\sigma_\Gamma^2 =& \, \mathbb{E}[\varepsilon^2 \, \mathbb{E}[D|Z]^2] / \, \mathbb{E}[\mathbb{E}[D|Z]^2]^2 \\ =& \sigma^2 \, \mathbb{E}[\mathbb{E}[D|Z]^2]^{-1}\end{aligned}$$

   en utilisant l'hypothèse d'homoscedasticité conditionnelle. On a donc que la borne (3.14) d'efficacité semi-paramétrique est atteinte (dans ce cas où $S = D$, $W = Z$ avec les notation de la section 3.4).
4. On vérifie que l'hypothèse (ORT) $\partial_\eta M(\tau_0, \eta_0) = 0$ est satifaite en remarquant qu'avec (6.11) et (6.2), on obtient $\partial_\nu M(\tau_0, \eta_0) = \tau_0 \, \mathbb{E}[X(\zeta'\delta_0)] = 0$, $\partial_\theta M(\tau_0, \eta_0) = -\, \mathbb{E}[X(\zeta'\delta_0)] = 0$. Les deux autres conditions par rapport à γ et δ sont aussi vérifiées en utilisant (6.3) et (6.4).

5. En utilisant les notations de (6.17), on a $u_{i,j} = \delta_j + \varepsilon_{i,j}$ et donc

$$\begin{aligned} P_{i,j} =& P\left(\delta_j + \varepsilon_{i,j} > \delta_{j'} + \varepsilon_{i,j'},\ \forall j' \neq j\right) \\ =& P\left(\delta_j - \delta_{j'} + \varepsilon_{i,j} > \varepsilon_{i,j'},\ \forall j' \neq j\right). \end{aligned}$$

Pour un $\varepsilon_{i,j}$ fixé, cette probabilité est le produit des probabilités des $\delta_j - \delta_{j'} + \varepsilon_{i,j} > \varepsilon_{i,j'}$, qui est donnée par $F(\delta_j - \delta_{j'} + \varepsilon_{i,j}) = \exp(-\exp(-(\delta_j - \delta_{j'} + \varepsilon_{i,j}))$. On a donc

$$P_{i,j} = \int \left(\prod_{j' \neq j} e^{-e^{-(\varepsilon + \delta_j - \delta_{j'})}} \right) e^{-\varepsilon} e^{-e^{-\varepsilon}} d\varepsilon.$$

Après un peu d'algèbre et un changement de variable, on obtient (6.17).

7 Pour aller plus loin

1. Pour prendre en compte les erreurs non gaussiennes et hétéroscédastiques, la pénalité ℓ_1 dans l'estimation Lasso standard est modifiée en utilisant des pénalité spécifiques conçues de manière à pouvoir appliquer les résultats de la théorie des déviations modérées. Cela donne une procédure d'estimation en deux étapes où ces pénalités sont initialisées avec la matrice d'identité, puis mises à jour en utilisant les termes d'erreur estimés de la première étape.
2. Le *sample-splitting* consiste à couper l'échantillon de manière à dé-corréler l'estimation des paramètres de nuisance de celle du paramètre d'intérêt. Un avantage exposé dans la section 7.2 est qu'en théorie on peut autoriser plus nombre de composantes non nulles (condition (7.7)). Un inconvénient du partitionnement est une certaine perte d'efficacité.

8 Inférence sur les effets hétérogènes

1. Le problème fondamental de l'inférence causale est que l'on n'observe jamais la *ground truth*, c'est-à-dire le véritable effet du traitement. Il n'est donc pas possible de réutiliser sans modification des procédures standards telles que la validation croisée, du moins pas avec la même efficacité.
2. L'auteur oppose deux stratégies d'attribution d'un traitement (par exemple une campagne d'e-mailing ou d'appels téléphoniques) : soit cibler ceux qui sont les plus susceptibles de ne pas acheter à nouveau, soit cibler ceux qui répondent le plus lorsqu'ils reçoivent le traitement. Ces deux populations ne sont pas nécessairement constituées des mêmes individus. Cibler ceux qui sont les plus susceptibles de ne pas acheter à nouveau est intuitivement logique, mais c'est aussi inefficace : le traitement devrait être attribué de manière à maximiser son impact, mesuré comme l'augmentation de la probabilité d'achat. L'apprentissage automatique générique pourrait être utilisé pour classer les personnes en groupes.

3. Les splits aléatoires dans la construction des forêts aléatoires imposent d'explorer tout le support des variables explicatives et évitent la concentration des coupures dans une partie limitée de l'espace, ce qui serait obtenu avec des divisions sélectionnées de manière optimale. Ceci est clé pour obtenir la consistance de l'estimateur.
4. On peut estimer le CATE, mais pas de manière convergente. Ainsi, seulement certains caractéristiques du CATE telles que le GATES ou le CLAN peuvent être estimées.
5. Dans ce cas, on a uniquement des intervalles de confiance **conditionnels**

$$\mathbb{P}\left(\theta_A \in [L_A, U_A] | \text{ Data}_A\right) = 1 - \alpha + o_P(1),$$

où $[L_A, U_A] := \left[\widehat{\theta}_A \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}_A\right]$. Cela ne tient pas compte de la variabilité introduite par le fractionnement de l'échantillon, qui empêche toute généralisation à une quelconque distribution de l'ensemble des données.
6. Si le même échantillon était utilisé pour estimer les splits et les valeurs sur les feuilles, l'algorithme aurait tendance à séparer deux feuilles qui ont des effets de traitement hétérogènes (relativement élevés et faibles) dans cet échantillon, conduisant ainsi à une estimation biaisée si nous utilisons l'échantillon pour l'évaluer. Si nous utilisons un autre échantillon pour l'évaluer, cela limite le sur-apprentissage et assure la convergence.
7. L'objectif des forêts aléatoires causales est d'estimer un effet de traitement de manière consistante alors que les forêts aléatoires estiment une fonction de régression et visent à minimiser l'erreur de prédiction (souvent en norme ℓ_2, ou MSE). Cela a des conséquences sur la forme de l'estimateur, les forêts aléatoires causales nécessitant la propriété d'honnêteté pour être consistante.
8. Le meilleur prédicteur linéaire du CATE est la projection linéaire d'un signal sans biais du CATE sur l'espace vectoriel linéaire généré par T. En ce sens BLP dépend donc des performances de T. S'il s'adapte bien au CATE, alors le coefficient de pente du BLP sera proche de un et nous apprendrons des caractéristiques du CATE en regardant T.

9 Apprendre la politique optimale

1. A l'aide de l'hypothèse 9.2 (i), la contrainte est pertinente et donc toujours saturée pour la politique optimale, i.e. $c = \int_{x \in \mathcal{X}} \pi(x) dF_X(x)$. Soit π' une politique optimale différente de π donnée en Proposition 9.1 sur un ensemble de F_X mesure non nulle. Cette politique satisfait aussi la contrainte, et avec l'hypothèse 9.2 (ii), il existe des ensembles Ω' et Ω, tel que

$$\int_{\Omega'} \pi'(x) dF_X(x) = \int_{\Omega} \pi(x) dF_X(x),$$

$\Omega' \subseteq \{x : \tau(x) < \gamma\}$, et $\Omega \subseteq \{x : \tau(x) \geq \gamma\}$. On a donc

$$\begin{aligned}\int_{\Omega'} \pi'(x)\tau(x)dF_X(x) &< \gamma \int_{\Omega'} \pi'(x)dF_X(x) \\ &= \gamma \int_{\Omega} \pi(x)dF_X(x) \\ &\leq \int_{\Omega} \pi(x)\tau(x)dF_X(x),\end{aligned}$$

et on obtient une contradiction.

2. L'objectif utilisé est le contrôle du regret dans le pire des cas (*minimax regret criterion*), voir (9.6). L'avantage est la robustesse des politiques recommandées aux différentes distributions des effets possibles. L'inconvénient est que l'on utilise pas de potentiel *apriori* sur ces effets qui permettraient d'obtenir des meilleurs résultats s'ils sont vérifiés.
3. Il est souhaitable de limiter les classes de politiques considérées de manière à pouvoir les implémenter simplement sur le terrain, mais aussi pour obtenir une borne supérieure sur le regret (Théorème 9.1).
4. On limite la complexité des classes de politiques considérées à l'aide de la dimension de Vapnik-Chervonenkis (VC). On peut autoriser des classes dont la complexité augmente avec la taille d'échantillon, mais moins rapidement que n.
5. cf. Remarque 9.2.
6. Cette formulation fait apparaître l'apprentissage de politiques par maximisation empirique comme un problème d'optimisation pondérée dans le cadre d'une classification. On peut utiliser les outils développés en classification pondérées (voir e.g., Athey et Wager, 2021 ; Zhou et al., 2018 pour plus de détails) pour résoudre ce problème.

10 La méthode du contrôle synthétique

1. Non, ce n'est pas le cas. En effet, il n'y a qu'une seule unité traitée, donc aucun résultat de type loi des grands nombres ne s'applique.
2. Il devrait y avoir trois possibilités parmi : (1) pas d'extrapolation (les poids sont non négatifs et leur somme est égale à un), (2) transparence de l'ajustement (l'ajustement avant le traitement peut être évalué), (3) empêche la recherche de spécification (les poids peuvent être calculés indépendamment du résultat après traitement), (4) sparsité/interprétation qualitative (au plus $p + 1$ sont strictement positifs). Voir le chapitre correspondant. Les réponses qui étaient trop génériques / également vraies pour d'autres estimateurs standard / non expliquées avec un argument précis ont été rejetées.
3. Le vecteur du poids de contrôle synthétique est en général peu dense, ce qui signifie que seules quelques entrées ne sont pas des zéros. Par

conséquent, les unités de contrôle correspondantes ne prennent pas part au contrefactuel. D'une part, il n'utilise pas l'échantillon complet (perte d'efficacité ?), mais d'autre part, il écarte les unités qui n'aident pas à reproduire l'échantillon traité.

4. cf. section 10.4.
5. Cela est impossible car on rejette le test d'absence d'effet du traitement avec une p-value de 0,02. 0 ne peut donc pas se trouver dans l'intervalle de confiance à 0,90.
6. Soit $D^{obs} = (D_1, \ldots, D_n)$ le vecteur observé de l'affectation du traitement et $\hat{\tau}^{obs}$ l'estimateur MCO correspondant. La procédure de Fisher est la suivante :
   (a) Pour $b = 1, \ldots, B$, remanier l'affectation du traitement de manière aléatoire, calculer l'estimateur MCO de τ_0, $\hat{\tau}_b$ et le comparer aux statistiques observées $\hat{\tau}^{obs}$.
   (b) Calculer la p-value de Fisher :

$$\hat{p} := \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ |\hat{\tau}_b| \geq |\hat{\tau}^{obs}| \right\}$$

   (c) Rejeter H_0 si $\hat{p}$ est inférieur à un seuil prédéterminé : l'allocation de traitement observée donne un effet qui est anormalement grand.

11 Prévision en grande dimension

1. On ne peut plus utiliser l'hypothèse d'indépendance des observations, qui sont corrélées dans le temps. Les séries temporelles économiques et financières sont aussi connues pour posséder des queues de distribution épaisses. Enfin, nous devons aussi prendre en compte le fait que les séries ne sont pas échantillonnées à la même fréquence.
2. Certaines variables explicatives ont souvent une structure particulière, qui fait que peu de groupes de variables (macroéconomie, différents secteurs d'activité, variables financières, news) peuvent être utiles pour la prédiction, mais au sein de ces groupes plusieurs variables le sont.
3. La pénalité du Lasso impose la parcimonie et conduit à un biais qu'il est nécessaire de prendre en compte quand il s'agit de faire de l'inférence sur un groupe de coefficients (voir section 7.3).
4. Le risque à utiliser une méthode imposant la parcimonie est un risque de mauvaise spécification, c'est à dire que cette hypothèse peut ne pas être vérifiée, entraînant un biais. Certains exemples empiriques mentionnés en section 11.2.1 montrent que cette hypothèse doit être justifiée avec précaution. L'approche FARM permet de combiner une partie parcimonieuse avec une partie dense, et donc de tester si cette dernière est utile.

12 Travailler avec des données textuelles

1. Le vocabulaire d'un texte peut être très vaste, ce qui nécessite de représenter chaque mot par un vecteur de grande dimension via une représentation one-hot (cf. chapitre 13). Dans les modèles plus avancés, les vecteurs de grande dimension permettent de capturer plus de détails et de nuances dans le texte.
2. Voici deux exemples parmi une infinité. (i) Traitement des erreurs de frappe. L'utilisation de n-grams de caractères permet de contourner ce problème en permettant de représenter le texte à un niveau inférieur, au niveau des séquences de caractères. (ii) Traitement des formes plurielles.
3. Les modèles faisant usage de variables latentes permettent de capturer le contexte général d'un document. Néanmoins, ces modèles sont d'une part complexe à estimer, et ont été, d'autres parts, dépassés par les modèles modernes de langage pré-entrainés en terme de performance.
4. Le premier problème consiste à apparier chaque message du forum avec le ou les actif(s) financier(s) mentionnés. Cette phase dépend largement de la qualité des données. Si les messages mentionnent systématiquement des symboles boursiers (AAPL pour Apple Inc., GME pour GameStop etc.) il suffit de se procurer une liste de ces symboles puis d'en cherche les occurrences dans les messages. En utilisant les dates des messages, on peut ensuite apparier chaque message au(x) prix des actifs mentionnés. Si les messages ne mentionnent pas de symboles boursiers, il faut soit adopter une approche plus sophistiquée (distance aux noms des entreprises etc.), soit adopter une approche plus *data-driven*, par exemple de classification « sac de mots » avec un grand nombre de classes (autant que d'actifs). Une fois l'appariement des données effectué, un grand nombre de stratégies sont possibles (e.g. analyse du sentiment des messages).

13 Représentation distribuée des mots

1. La similarité cosinus pour deux mots distincts est de zéro.
2. La réponse à la question précédente indique que pour deux mots distincts la similarité cosinus est de zéro. Il n'est donc pas possible d'obtenir une notion de « degré de similarité » entre deux mots avec cette approche.
3. Un plongement lexical est une technique utilisée pour représenter les mots d'un texte sous forme de vecteurs de dimension réduite. Le but d'un plongement lexical est de capturer de manière compacte les relations sémantiques entre les mots dans le texte, de manière à pouvoir utiliser ces représentations vectorielles pour effectuer des tâches de traitement du langage naturel. On peut ensuite utiliser ces vecteurs pour représenter un document en les agrégeant (par exemple via une moyenne), puis en utilisant cette représentation par exemple dans une régression logistique.
4. cf. 13.2.

5. cf. section 13.2.
6. cf. section 13.3.
7. On espère que l'on trouvera le vecteur représentant le mot « taureau ».

14 Apprentissage supervisé

1. L'idée est d'avoir une mesure de la qualité d'un produit telle que perçue par le consommateur. Par exemple, on peut penser à inclure dans un modèle la moyenne des embeddings des commentaires d'évaluation d'un produit. Ou bien les mebeddings des photos de ce produit.
2. cf. section 14.2.
3. cf. section 14.2.2.
4. Par l'utilisation de n-grams de mots. Les limites sont qu'il peut être compliqué de prendre en compte des dépendances longues, que le nombre de tokens augmente de manière exponentielle, et que l'algorithme ainsi entrainé ne pourra pas s'adapter à une séquence de mot jamais vue.