Estimation et inférence sur données "multi-indicées" (ou en cas de "multiway clustering")

Laurent Davezies
CREST

Xavier D'Haultfœuille
CREST

Yannick Guyonvarch
INRAE

JMS
Mars 2022

Plan

Le cadre

Cadre usuel de la statistique inférentielle et de l'économétrie:

- Données iid $(Y_i, X_i)_{i=1,\dots,n}$.
- Etude de l'asymptotique des estimateurs

$$\widehat{\theta} = g(Y_1, ..., Y_n, X_1, ..., X_n)$$

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

- Inférence basée sur une estimation de Σ ou par bootstrap.
- Exemple: MCO, Logit, Probit, quantiles, Méthode des moments (généralisée), etc...

Le cadre

Question :Comment adapter les résultats quand on observe des données indiciées par deux indices (correspondant à deux populations différentes $\mathcal{P}$ et $\mathcal{P}'$ ou à une paire issue d'une seule population $\mathcal{P} = \mathcal{P}'$) ?

Exemples $\mathcal{P} = \mathcal{P}'$:

- commerce international : $i \in \{1, ..., n\} \subset \mathcal{P}$ = population des pays, $Y_{i,j}$ = exportation de i vers j = importation de j d'origine i, $X_{i,j}$ = distance ou accords commerciaux entre i et j ou caractéristiques de i, caractéristique de j...
- données de championnat sportif : $i \in \{1, ..., n\} \subset \mathcal{P}$: population des joueurs (ou des équipes si sport collectif), $Y_{i,j}$ = résultats sportifs, $X_{i,j}$ conditions du match, caractéristiques de i ou caractéristiques de j...

Le cadre

Question : Comment adapter les résultats quand on observe des données indiciées par deux indices (correspondant à deux populations différentes $\mathcal{P}$ et $\mathcal{P}'$ ou à une paire issue d'une seule population $\mathcal{P} = \mathcal{P}'$) ?

Exemples $\mathcal{P} \neq \mathcal{P}'$:

- Marché : $i \in \{1, ..., n_1\} \subset \mathcal{P}$ = population des offreurs, $j \in \{1, ..., n_2\} \subset \mathcal{P}'$ = population des demandeurs, $Y_{i,j}$ quantité échangée entre (i, j), $X_{i,j}$ caractéristiques de l'offreur i , du demandeur j, ou "distance" entre i et j...
- "Multiway-clustering" : $i \in \{1, ..., n_1\} \subset \mathcal{P}$ = population de secteurs économiques, $j \in \{1, ..., n_2\} \subset \mathcal{P}'$ population de marchés du travail locaux, $Y_{i,j} = \sum_{\ell=1}^{N_{i,j}} W_{\ell,i,j}$ somme des salaires dans le secteur industriel i et le secteur géographique j, $N_{i,j}$ = nombre de salariés observés sur le marché du travail (i, j), $X_{i,j}$ caractéristiques des secteurs industriels i, des secteurs géographiques j ou du marché du travail (i, j), ou $X_{i,j} = \sum_{\ell=1}^{N_{i,j}} U_{\ell,i,j}$ avec $U_{\ell,i,j}$ des caractéristiques des salariés.

Le cadre

Question : Comment adapter les résultats quand on observe des données indiciées par deux indices (correspondant à deux populations différentes $\mathcal{P}$ et $\mathcal{P}'$ ou à une paire issue d'une seule population $\mathcal{P} = \mathcal{P}'$) ?

Données non indépendantes ! :

- Si $\mathcal{P} = \mathcal{P}'$: $(Y_{i,j}, X_{i,j})$ corrélées à $(Y_{i',j'}, X_{i',j'})$ si $\{i,j\} \cap \{i',j'\} \neq \emptyset$.
- Si $\mathcal{P} \neq \mathcal{P}'$: $(Y_{i,j}, X_{i,j})$ corrélées à $(Y_{i',j'}, X_{i',j'})$ si $i = i'$ ou $j = j'$.

Le cadre

Question : Comment adapter les résultats quand on observe des données indiciées par deux populations $\mathcal{P}$, $\mathcal{P}'$ (avec éventuellement $\mathcal{P} = \mathcal{P}'$) ?

Données échangeables : La distribution des données est invariante par permutation des individus dans la population $\mathcal{P}$ et $\mathcal{P}'$.

- Si $\mathcal{P} = \mathcal{P}'$, **tableau jointement échangeable**: pour toute permutation π de $\mathcal{P}$:

$$(Y_{i,j}, X_{i,j})_{i,j \in \mathcal{P}^2} \overset{loi}{=} (Y_{\pi(i),\pi(j)}, X_{\pi(i),\pi(j)})_{i,j \in \mathcal{P}^2}$$

- Si $\mathcal{P} \neq \mathcal{P}'$, **tableau séparablement échangeable**: pour toutes permutations π_1 de $\mathcal{P}$ et π_2 de $\mathcal{P}'$:

$$(Y_{i,j}, X_{i,j})_{i,j \in \mathcal{P} \times \mathcal{P}'} \overset{loi}{=} (Y_{\pi_1(i),\pi_2(j)}, X_{\pi_1(i),\pi_2(j)})_{i,j \in \mathcal{P} \times \mathcal{P}'}$$

Le cadre

Question : Comment adapter les résultats quand on observe des données indiciées par deux populations $\mathcal{P}$, $\mathcal{P}'$ (avec éventuellement $\mathcal{P} = \mathcal{P}'$) ?

Données dissociées : Sans perte de généralité, on peut alors supposer l'indépendance entre des données qui ne concernent pas "les mêmes individus".

- Si $\mathcal{P} = \mathcal{P}'$: pour $\mathcal{A}, \mathcal{B} \subset \mathcal{P}^2$, $\mathcal{A} \cap \mathcal{B} = \emptyset$

$$(Y_{i,j}, X_{i,j})_{i,j \in \mathcal{A}^2} \perp\!\!\!\perp (Y_{i,j}, X_{i,j})_{i,j \in \mathcal{B}^2}$$

- Si $\mathcal{P} \neq \mathcal{P}'$: pour $\mathcal{A}, \mathcal{B} \subset \mathcal{P}^2$, $\mathcal{A} \cap \mathcal{B} = \emptyset$ et $\mathcal{A}', \mathcal{B}' \subset \mathcal{P}'^2$, $\mathcal{A}' \cap \mathcal{B}' = \emptyset$

$$(Y_{i,j}, X_{i,j})_{i,j \in \mathcal{A} \times \mathcal{A}'} \perp\!\!\!\perp (Y_{i,j}, X_{i,j})_{i,j \in \mathcal{B} \times \mathcal{B}'}$$

Le cadre

Question : Comment adapter les résultats quand on observe des données indiciées par deux populations $\mathcal{P}$, $\mathcal{P}'$ (avec éventuellement $\mathcal{P} = \mathcal{P}'$) ?

Cadre asymptotique : l'économètre observe $(Y_{i,j}, X_{i,j})$ pour $i = 1, ..., n_1$ et $j = 1, ..., n_2$. Que peut-on dire des estimateurs usuels quand $\underline{n} := \min(n_1, n_2) \to \infty$?

- Si $\mathcal{P} = \mathcal{P}'$: on suppose $\underline{n} = n_1 = n_2 := n$ tend vers l'infini.
- Si $\mathcal{P} \neq \mathcal{P}'$: on suppose que $\underline{n}$ tend vers l'infini et que $\lim_{\underline{n}} \frac{\underline{n}}{n_1} = \lambda_1 \in [0; 1]$, $\lim_{\underline{n}} \frac{\underline{n}}{n_2} = \lambda_2 \in [0; 1]$

Plan

LGN et TCL uniformes

Dans le cadre usuel de l'économétrie, de nombreux estimateurs sont définis comme solution d'une maximisation et/ou comme solution d'équations du type:

$$\sum_{i=1}^{n} m(Y_i, X_i, \widehat{\theta}) = \sum_{i=1}^{n} m_{i,\widehat{\theta}} = 0,$$

qui est motivé par $\mathbb{E}(m(Y_1, X_1, \theta)) = \mathbb{E}(m_{1,\theta}) = 0$ si et seulement si $\theta = \theta_0$.

- MCO: $m_{i,\theta} = X_i (Y_i - X_i'\theta)$
- VI (avec $\dim(Z) = \dim(X)$): $m_{i,\theta} = Z_i (Y_i - X_i'\theta)$
- Logit: $m_{i,\theta} = X_i (Y_i - \Lambda(X_i'\theta))$, avec $\Lambda(u) = \frac{\exp u}{1+\exp u}$
- Probit: $m_{i,\theta} = X_i \frac{\varphi(X_i'\theta)}{\Phi(X_i'\theta)\left(1-\Phi(X_i'\theta)\right)} (Y_i - \Phi(X_i'\theta))$ avec Φ, φ fonction de répartition et densité de $\mathcal{N}(0,1)$
- Poisson : $m_{i,\theta} = X_i (Y_i - \exp(X_i'\theta))$

Asymptotique

Si le modèle statistique n'est pas "trop irrégulier" alors:

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}(0, H^{-1}VH^{-1})$$

$$\widehat{H} := \frac{1}{n}\sum_i \nabla'_\theta m_{i,j,\widehat{\theta}} \xrightarrow{\mathbb{P}} H = \mathbb{E}\left(\nabla'_\theta m_{1\theta_0}\right)$$

$$\widehat{V} := \frac{1}{n}\sum_{i=1}^{n} m_{i,\widehat{\theta}} m'_{i,\widehat{\theta}} \xrightarrow{\mathbb{P}} V := \mathbb{V}\left(m_{1,\theta_0}\right)$$

LGN et TCL uniformes

Quelles conditions de régularité ?

Si les moments d'ordre 1 et/ou 2 existent, la loi des grands nombres et le théorème central limite assurent respectivement que pour toute valeur de θ:

$$\frac{1}{n}\sum_{i=1}^{n} m_{i,\theta} - \mathbb{E}\left(m_{1,\theta}\right) \xrightarrow{\mathbb{P}} 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta m_{i,\theta} - \mathbb{E}\left(\nabla_\theta m_{1,\theta}\right) \xrightarrow{\mathbb{P}} 0$$

$$\frac{1}{n}\sum_{i=1}^{n} m_{i,\theta} m'_{i,\theta} - \mathbb{E}\left(m_{1,\theta} m'_{1,\theta}\right) \xrightarrow{\mathbb{P}} 0$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_{i,\theta} - \mathbb{E}\left(m_{1,\theta}\right)\right) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}(m_{1,\theta})).$$

LGN et TCL uniformes

Mais il faut plus que cela pour assurer...

- la convergence de $\widehat{\theta}$ et de son estimateur de variance $\rightarrow$ loi des grands nombres uniforme:

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m_{i,\theta} - \mathbb{E}\left(m_{1,\theta}\right) \right| \xrightarrow{\mathbb{P}} 0$$

$$\sup_{\theta \in \mathcal{V}(\theta_0)} \left| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta m_{i,\theta} - \mathbb{E}\left(\nabla_\theta m_{1,\theta}\right) \right| \xrightarrow{\mathbb{P}} 0$$

$$\sup_{\theta \in \mathcal{V}(\theta_0)} \left| \frac{1}{n} \sum_{i=1}^{n} m_{i,\theta} m'_{i,\theta} - \mathbb{E}\left(m_{1,\theta} m'_{1,\theta}\right) \right| \xrightarrow{\mathbb{P}} 0$$

- Implique la loi des grands nombres pour chaque θ (mais pas équivalent)

LGN et TCL uniformes

Mais il faut plus que cela pour assurer...

- ... la normalité asymptotique de $\widehat{\theta} \rightarrow$ théorème central limite uniforme:

$$\mathbb{G}_n(.) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_{i,.} - \mathbb{E}\left(m_{1,.}\right)\right) \xrightarrow{d} \mathbb{G}(.),$$

  où $\mathbb{G}$ est un processus gaussien de $\ell^\infty(\mathcal{V}(\theta_0))$ de covariance $\mathbb{C}ov(m_{1,.}, m_{1,.})$.
  C'est à dire que $\mathbb{G}_n, \mathbb{G} \in (\ell^\infty(\mathcal{V}(\theta_0)), ||.||_\infty) =$ ensemble des fonctions aléatoires bornées définies sur un voisinage $\mathcal{V}(\theta_0)$ de θ_0. $\mathbb{G}_n$ converge en loi vers $\mathbb{G}$ comme éléments de $(\ell^\infty(\mathcal{V}(\theta_0)), ||.||_\infty)$.

- Implique le TCL multivarié pour chaque $\theta \in \mathcal{V}(\theta_0)$ (mais pas équivalent)

LGN et TCL uniformes

Les conditions de régularité dans le cadre iid qui assurent les convergences uniformes : le modèle statistique est "couvrable" par un nombre de boules qui ne croit pas trop vite quand le rayon des boules dimininue (pour une introduction à ces concepts: chapitres 18 et 19 de van der Vaart, *Asymptotic Statistics*, 2000). Conditions vérifiées par tous les modèles statistiques utilisés en pratique.

Contribution : Sous des conditions de régularité similaires au cas iid, c'est aussi le cas pour les tableaux échangeables !

Donc tout va aussi "fonctionner" pour des données multi-indicées... mais :

- les vitesses de convergence sont généralement en $\sqrt{n}$ pour au moins n^2 observations
- les variances asymptotiques sont modifiées

Convergence du bootstrap

Le bootstrap est une méthode populaire pour estimer des écarts-types, des intervalles de confiance ou des p-value, par re-échantillonage (avec remise). La validité du bootstrap (pour les estimateurs précédents) suppose également d'établir des théorèmes de convergence uniforme.

Contribution : Les hypothèses de régularité impliquent également la convergence uniforme du bootstrap dans notre cadre.

Mais il faut échantillonner les "individus" i et j mais pas les paires (i, j) :

- Si $\mathcal{P} = \mathcal{P}'$, il faut échantilloner avec remise les "individus" $i_1^*, ..., i_n^*$ et estimer les quantités d'intérêt sur les "paires" $(Y_{i_i^*, i_{i'}^*}, X_{i_i^*, i_{i'}^*})_{i,i'=1,...,n, i_i^* \neq i_{i'}^*}$
- Si $\mathcal{P} \neq \mathcal{P}'$, il faut échantilloner avec remise n_1 individus $i_1^*, ..., i_{n_1}^*$ dans $\mathcal{P}$, n_2 individus $j_1^*, ..., j_{n_2}^*$ dans $\mathcal{P}'$ et estimer les quantités d'intérêt sur les "paires" $(Y_{i_i^*, j_j^*}, X_{i_i^*, j_j^*})_{i=1,...,n_1, j=1,...,n_2}$

Plan

Le cadre

Pour une regression, un Logit, un Probit, ...

$$\sum_{i,j} m_{i,j,\widehat{\theta}} = 0$$

avec:

- MCO: $m_{i,j,\theta} = X_{i,j}\left(Y_{i,j} - X_{i,j}'\theta\right)$
- VI (avec $\dim(Z) = \dim(X)$): $m_{i,j,\theta} = Z_{i,j}\left(Y_{i,j} - X_{i,j}'\theta\right)$
- Logit: $m_{i,j,\theta} = X_{i,j}\left(Y_{i,j} - \Lambda(X_{i,j}'\theta)\right)$, avec $\Lambda(u) = \frac{\exp u}{1+\exp u}$
- Probit: $m_{i,j,\theta} = X_{i,j}\frac{\varphi(X_{i,j}'\theta)}{\Phi(X_{i,j}'\theta)\left(1-\Phi(X_{i,j}'\theta)\right)}\left(Y_{i,j} - \Phi(X_{i,j}'\theta)\right)$ avec Φ, φ fonction de répartition et densité de $\mathcal{N}(0,1)$
- Poisson: $m_{i,j,\theta} = X_{i,j}\left(Y_{i,j} - \exp(X_{i,j}'\theta)\right)$

Cas particulier : "Multiway Clustering"

Le cadre :

- $((Y_{\ell,i,j})_{\ell\geq 1}, (X_{\ell,i,j})_{\ell\geq 1}, N_{i,j})_{i,j\geq 1}$ est un array séparablement échangeable (donc $\mathcal{P} \neq \mathcal{P}'$)
- Exemple: $\ell \geq 1$: population d'individus ; $i \in \mathcal{P}$: population des marchés locaux d'emploi (secteur géographique) ; $j \in \mathcal{P}'$: population des secteurs industriels
- L'économètre observe les $N_{i,j}$ premières observations $(Y_{\ell,i,j}, X_{\ell,i,j})_{\ell=1,\ldots,N_{i,j}}$ dans la "cellule" (i,j), et cela pour n_1 éléments $i \in \mathcal{P}$ et n_2 éléments $j \in \mathcal{P}'$.

Cas particulier : "Multiway Clustering"

Pour une regression, un Logit, un Probit, ...:

$$\sum_{i,j} m_{i,j,\widehat{\theta}} = 0$$

avec:

- MCO : $m_{i,j,\theta} = \sum_{\ell=1}^{N_{i,j}} X_{\ell,i,j} \left(Y_{\ell,i,j} - X'_{\ell,i,j}\theta \right)$
- VI (avec $\dim(Z) = \dim(X)$) : $m_{i,j,\theta} = \sum_{\ell=1}^{N_{i,j}} Z_{\ell,i,j} \left(Y_{\ell,i,j} - X'_{\ell,i,j}\theta \right)$
- Logit : $m_{i,j,\theta} = \sum_{\ell=1}^{N_{i,j}} X_{\ell,i,j} \left(Y_{\ell,i,j} - \Lambda(X'_{\ell,i,j}\theta) \right)$, avec $\Lambda(u) = \frac{\exp u}{1+\exp u}$
- Probit : $m_{i,j,\theta} = \sum_{\ell=1}^{N_{i,j}} X_{\ell,i,j} \frac{\varphi(X'_{\ell,i,j}\theta)}{\Phi(X'_{\ell,i,j}\theta)\left(1-\Phi(X'_{\ell,i,j}\theta)\right)} \left(Y_{\ell,i,j} - \Phi(X'_{\ell,i,j}\theta) \right)$ avec Φ, φ fonction de répartition et densité de $\mathcal{N}(0,1)$
- Poisson : $m_{i,j,\theta} = \sum_{\ell=1}^{N_{i,j}} X_{\ell,i,j} \left(Y_{\ell,i,j} - \exp(X'_{\ell,i,j}\theta) \right)$

Asymptotique

Si $\mathcal{P} = \mathcal{P}'$

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, H^{-1} V H^{-1})$$

$$\widehat{H} := \frac{1}{n(n-1)} \sum_{i,j} \nabla'_\theta m_{i,j,\widehat{\theta}} \xrightarrow{\mathbb{P}} H = \mathbb{E}\left(\nabla'_\theta m_{1,2,\theta_0}\right)$$

$$\widehat{V} := \frac{1}{n(n-1)^2} \sum_{i=1}^{n} \left(\sum_{j \neq i} \rho_{i,j,\widehat{\theta}}\right) \left(\sum_{j \neq i} \rho_{i,j,\widehat{\theta}}\right)' \xrightarrow{\mathbb{P}} V := \mathbb{C}ov\left(\rho_{1,2,\theta_0}, \rho_{1,3,\theta_0}\right)$$

$$\rho_{i,j,\theta} = m_{i,j,\theta} + m_{j,i,\theta}$$

Asymptotique

Si $\mathcal{P} \neq \mathcal{P}'$ pour $\underline{n} = \min(n_1, n_2)$, $(\frac{\underline{n}}{n_1}, \frac{\underline{n}}{n_2}) \to (\lambda_1, \lambda_2) \in [0;1]^2$:

$$\sqrt{\underline{n}}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, H^{-1} V H^{-1})$$

$$\widehat{H} := \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \nabla'_\theta m_{i,j,\widehat{\theta}} \xrightarrow{\mathbb{P}} H = \mathbb{E}\left(\nabla'_\theta m_{1,1,\theta_0}\right)$$

$$\widehat{V} := \frac{\underline{n}}{n_1^2 n_2^2} \left[\sum_{i=1}^{n_1} \left(\sum_{j=1}^{n_2} m_{i,j,\widehat{\theta}} \right) \left(\sum_{j=1}^{n_2} m_{i,j,\widehat{\theta}} \right)' + \sum_{j=1}^{n_2} \left(\sum_{i=1}^{n_1} m_{i,j,\widehat{\theta}} \right) \left(\sum_{i=1}^{n_1} m_{i,j,\widehat{\theta}} \right)' \right]$$

$$\xrightarrow{\mathbb{P}} V := \lambda_1 \mathbb{C}ov\left(m_{1,1,\theta_0}, m_{1,2,\theta_0}\right) + \lambda_2 \mathbb{C}ov\left(m_{1,1,\theta_0}, m_{2,1,\theta_0}\right)$$

Inférence

Si V est inversible alors les tests de Student, Fisher, les intervalles de confiance basés sur la normalité asymptotique et les estimateurs $\widehat{H}$ et $\widehat{V}$ sont convergents.

Si V n'est pas inversible:

1. la vitesse de convergence n'est pas $\sqrt{n}$ mais $n = \sqrt{n^2}$ pour $\mathcal{P} = \mathcal{P}'$,
   la vitesse de convergence n'est pas $\sqrt{\underline{n}}$ mais $\sqrt{\underline{n^2}}$ pour $\mathcal{P} \neq \mathcal{P}'$,
2. la limite n'est pas forcément Gaussienne,
3. mais les tests de Student, Fisher, les intervalles de confiance, les p-values basés sur la vitesse $\sqrt{n}$ ou $\sqrt{\underline{n}}$ et en utilisant $\widehat{H}$ et $\widehat{V}$ sont conservateurs.

Inférence

Implication pratique pour le multiway-clustering:

1. Estimation des matrices *cluster-robust* de variance-covariance pour chaque "dimension" des indices (2 dimensions dans cette présentation mais plus de dimensions possibles).
   Ces estimateurs sont généralement disponibles dans les logiciels statistiques/économétriques.
2. En sommant de ces estimateurs on obtient exactement l'estimateur $\widehat{H}^{-1}\widehat{V}\widehat{H}^{-1}$.
3. On peut alors calculer les intervalles de confiance, les p-values ou les tests.

Plan

Bibliographie

Communication basée sur une partie des résultats de:

- Empirical Process Results for Exchangeable Arrays : arxiv.org/abs/1906.11293
- Asymptotic Results under Multiway Clustering : arxiv.org/abs/1807.07925

On trouvera dans ces documents d'autres résultats (nombre quelconque d'indices, "Delta-méthode fonctionnelle", "multiplier bootstrap", etc.)