Mixed Logit

Fabian Bastin

Avril 2011

Formulation

Dans une forme populaire du modèle logit mélangé (mixed logit) multinomial, l'utilité de l'alternative j pour l'individu q est

$$U_{q,j} = \boldsymbol{\beta}_q^{\mathrm{t}} \boldsymbol{x}_{q,j} + \epsilon_{q,j} = \sum_{\ell=1}^{s} \beta_{q,\ell} x_{q,j,\ell} + \epsilon_{q,j}$$

où $\boldsymbol{\beta}_q = (\beta_{q,1}, \ldots, \beta_{q,s})^{\mathrm{t}}$ est un vecteur aléatoire non-observé de paramètres de goût (ou coefficients) pour chaque individu q,

$\boldsymbol{x}_{q,j} = (x_{q,j,1}, \ldots, x_{q,j,s})^{\mathrm{t}}$: attributs observés pour le choix j et l'individu q

$\epsilon_{q,j}$: variables de Gumbel indépendantes, de facteur d'échelle égaux à 1.

Probabilité individuelle

Pour un individu aléatoire q, β_q a une densité multivariée f_θ qui dépend d'un vecteur de paramètres θ. L'individu q sélectionne toujours l'alternative j ayant la plus grande utilité $U_{q,j}$.

Conditionnellement à β_q, q sélectionne j avec la probabilité

$$L_q(j, \beta_q) = \frac{e^{\beta_q^t \boldsymbol{x}_{q,j}}}{\sum_{a \in \mathcal{A}(q)} e^{\beta_q^t \boldsymbol{x}_{q,a}}},$$

indépendemment des autres individus, où $\mathcal{A}(q)$ est l'ensemble de ses alternatives.

La probabilité non conditionnelle qu'un individu sélectionné aléatoirement sélectionne l'alternative j est

$$p_q(j, \theta) = E[L_q(j, \beta_q)] = \int_{\mathcal{R}^s} L_q(j, \beta) f_\theta(\beta)\, d\beta.$$

Probabilité individuelle (suite)

Nous supposons que β_q peut être écrit comme

$$\beta_q = h(\boldsymbol{\theta}, \boldsymbol{U})$$

pour une certaine fonction explicite h, où $\boldsymbol{U}$ est un vecteur de variables aléatoires indépendantes uniformes sur $(0, 1)$.

Cette hypothèse est standard, et requiert de simuler des réalisations de β_q à partir de nombres aléatoires uniformes indépendants.

Si $\boldsymbol{U}$ est de dimension s,

$$p_q(j, \boldsymbol{\theta}) = E\left[L_q(j, h(\boldsymbol{\theta}, \boldsymbol{U}))\right] = \int_{(0,1)^s} L_q(j, h(\boldsymbol{\theta}, \boldsymbol{u}))\, \mathrm{d}\boldsymbol{u}.$$

Estimation MC des probabilités de choix

Afin d'estimer ces intégrales par Monte Carlo, pour θ et q donnés, nous générons n_q points aléatoires indépendants $\boldsymbol{U}_q^{(1)}, \ldots, \boldsymbol{U}_q^{(n_q)}$ dans $(0,1)^s$, prenons $\beta_q^{(i)}(\theta) = h(\theta, \boldsymbol{U}_q^{(i)})$ for $i = 1, \ldots, n_q$, et calculons l'estimateur non biaisé

$$\hat{p}_q^{n_q}(j, \theta) = \frac{1}{n_q} \sum_{i=1}^{n_q} L_q(j, \beta_q^{(i)}(\theta)) = \frac{1}{n_q} \sum_{i=1}^{n_q} L_q(j, h(\theta, \boldsymbol{U}_q^{(i)})).$$

Panel

Quand le même individu (avec le même β_q) délivre $T_q > 1$ observations et sélectionne l'alternative j_t pour sa t^e décision, sous l'hypothèse que les sélections sont indépendantes, conditionnellement à β_q, la vraisemblance jointe de cette séquence de choix est

$$L_q^{T_q}(j_1, \ldots, j_{T_q}, \beta_q) = \prod_{t=1}^{T_q} L_q(j_t, \beta_q),$$

où chaque $L_q(j_t, \beta_q)$ est calculé avec la formule logit et sa probabilité non-conditionnelle est

$$p_q(j_1, \ldots, j_{T_q}, \theta) = E\left[L_q^{T_q}(j_1, \ldots, j_{T_q}, h(\theta, \boldsymbol{U}))\right].$$

Cette espérance peut être estimé par Monte-Carlo de façon similaire au cas cross-sectional, en remplaçant chaque $L_q(\cdot)$ par un produit $L_q^{T_q}(\cdot)$.

Biais et variance

Nous venons de produire un estimateur non-biaisé de la fonction de vraisemblance pour un individu.

Pour estimer θ à partir d'un jeu de données, habituellement, nous calculons le maximum de la log-vraisemblance $\ln L(\theta)$ (the logarithm of the likelihood). Mais nous n'avons pas un estimateur sans biais de cette fonction.

Spécifiquement, supposons que nous avons un jeu de données avec une observation par individu, pour m individus, dans lequel chaque individu q fait face au vecteur d'attributs $\boldsymbol{x}_{q,j}$ pour chaque alternative j et fait le choix y_q, for $q = 1, \ldots, m$.

Maximum de vraisemblance

La log-vraisemblance, que nous divisons par m pour obtenir une moyenne sur les individus et éviter que l'expression n'explose quand $m \to \infty$, est

$$\frac{\ln L(\boldsymbol{\theta})}{m} = \frac{1}{m} \ln \prod_{q=1}^{m} p_q(y_q, \boldsymbol{\theta}) = \frac{1}{m} \sum_{q=1}^{m} \ln p_q(y_q, \boldsymbol{\theta}).$$

En remplaçant les $p_q(y_q, \boldsymbol{\theta})$ par leur estimateur $\hat{p}_q^{n_q}(y_q, \boldsymbol{\theta})$, nous obtenons l'estimateur de $\ln L(\boldsymbol{\theta})/m$:

$$\begin{aligned} \frac{\ln(\hat{L}(\boldsymbol{\theta}))}{m} &= \frac{1}{m} \sum_{q=1}^{m} \ln \left(\hat{p}_q^{n_q}(y_q, \boldsymbol{\theta}) \right) \\ &= \frac{1}{m} \sum_{q=1}^{m} \ln \left(\frac{1}{n_q} \sum_{i=1}^{n_q} L_q \left(y_q, h(\boldsymbol{\theta}, \boldsymbol{U}_q^{(i)}) \right) \right). \end{aligned}$$

Quand $T_q > 1$, le $L_q(\cdot)$ à l'intérieur de la somme est remplacé par le produit $L_q^{T_q}(\cdot)$.

Biais et variance

Cet estimateur est biaisé comme ln n'est pas une fonction linéaire, et le biais est négatif vu que ln est concave. Soit

$$R_q = \frac{\hat{p}_q^{n_q}(y_q, \theta) - p_q(y_q, \theta)}{p_q(y_q, \theta)},$$

l'erreur relative d'estimation de $p_q(y_q, \theta)$.

Le développement de Taylor de $\ln(\hat{p}_q^{n_q}(y_q, \theta))$ autour $\ln(p_q(y_q, \theta))$ donne

$$\ln(\hat{p}_q^{n_q}(y_q, \theta)) - \ln(p_q(y_q, \theta)) \quad = \quad R_q - R_q^2/2 + \mathcal{O}(|R_q|^3). \quad (1)$$

Biais et variance (suite)

Le biais total peut alors s'écrire, comme $E[R_q] = 0$,

$$E\left[\frac{\ln(\hat{L}(\theta)) - \ln(L(\theta))}{m}\right] = \frac{1}{m}\sum_{q=1}^{m} E\left[-\frac{R_q^2}{2} + \mathcal{O}(|R_q|^3)\right]$$
$$\approx -\frac{1}{2m}\sum_{q=1}^{m} E[R_q^2],$$

et, puisque $\mathrm{Var}[R_q] = E[R_q^2]$ et que les R_q's sont indépendants, la variance est

$$\mathrm{Var}\left[\frac{\ln(\hat{L}(\theta))}{m}\right] \approx \mathrm{Var}\left[\frac{1}{m}\sum_{q=1}^{m} R_q\right] = \frac{1}{m^2}\sum_{q=1}^{m} E[R_q^2]. \qquad (2)$$

Biais et variance (suite)

Avec MC, $E[R_q^2] = \mathcal{O}(1/n_q)$, aussi si $n_q = n$ pour tout q, le biais est en $\mathcal{O}(1/n)$, et la variance en $\mathcal{O}(1/(mn))$.

Pour m fixé, la contribution du carré du biais à l'erreur quadratique moyenne (MSE: mean square error) devient négligeable en comparaison de la variance quand n est assez grand: $\mathcal{O}(n^{-2})$ comparé à $\mathcal{O}((mn)^{-1})$.

En pratique, n n'est pas toujours très grand, aussi le biais peut être significatif, et n'est pas réduit quand nous augmentons m, au contraire de la variance.

Biais et variance: observations

- On pourrait s'attendre à ce que $E[R_q^2]$ soit plus grand quand p_q est plus petit, et ceci est typiquement ce qui peut être observé empiriquement.
- Afin de réduire le biais, nous pourrions soustraire son approximation à l'estimateurs $\ln(\hat{L}(\theta))/m$. Bastin and Cirillo (2010) ont examiné cette stratégie dans le contexte d'estimation MC, et ont empiriquement observé que cette correction élimine la majeure part du biais sans significativement augmenter la variance.
- Il existe cependant une autre source de biais, de signe opposé, venant de l'optimisation, et mitigant les gains de cette approche de correction.

Optimisation

Si nous fixons les réalisations de $\boldsymbol{U}_q^{(i)}$'s, la fonction de log-vraisemblance approchée devient une fonction déterministe de θ, appelée l'*approximation d'échantillonnage moyen* (SAA) de $\ln L(\theta)/m$.

L'algorithme d'optimisation SAA calcule le $\hat{\theta}$ qui maximise la log-vraisemblance approchée et le prend comme approximation du θ^* maximisant $\ln L(\theta)/m$.

Sa convergence dans le contexte du modèle mixed logit est étudiée dans Bastin, Cirillo, and Toint (2006).

Si nous supposons que le problème SAA est résolu exactement, le comportement de l'erreur $\hat{\theta} - \theta^*$ dépend seulement de l'erreur de la SAA comme approximation de la véritable fonction de log-vraisemblance.

Matrice d'information

Notons aussi que l'hypothèse **R7** considérée lors de l'étude des propriétés asymptotiques est violée lorsque nous considérons les modèles mixed logit, comme souvent nous avons plus d'une solution au problème de maximum de vraisemblamce. Par exemple, pour une paramètre β_j normalement distribué, un programme d'optimisation peut délivrer un écart-type négatif si aucune contrainte n'est imposéem alors qu'il faut interpréter cette écart-type en valeur absolue.

Nous pouvons forcer la première partie de **R7** en réduisant l'ensemble réalisable (p.e. en imposant des contraintes de non-négativité pour les écarts-type) Des méthodes de correction peuvent être appliquées, mais ceci n'affecte le biais d'optimisation. Ceci rend l'hypothèse de consistance douteuse.

Toutefois, pour un nombre fixé de tirs, l'estimateur demeure non consistant comme il est biaisé.

(R)QMC et mixed logit

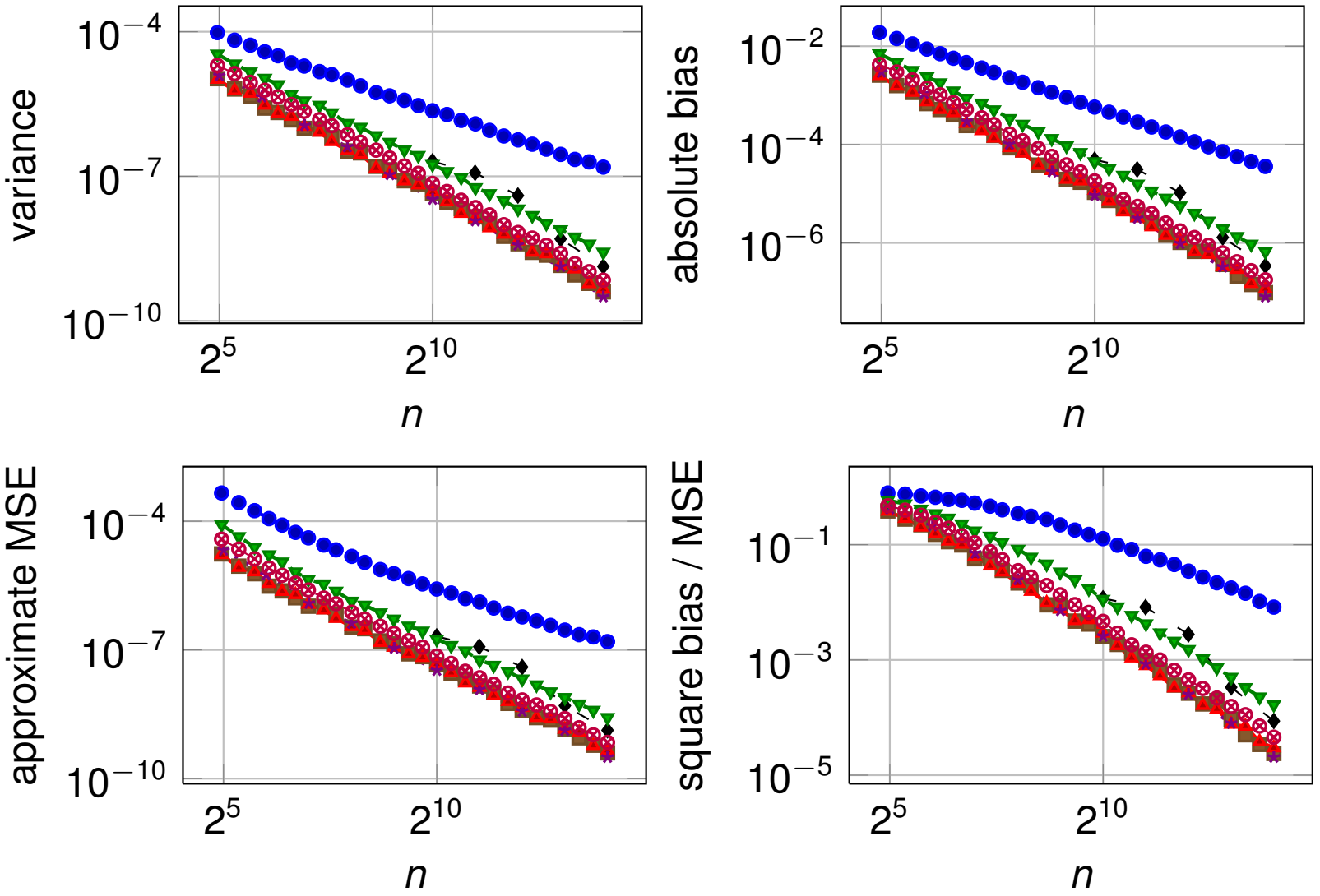
Afin de réduire la variance, il est d'usage d'utiliser des tirs QMC Halton. La grande majorité des études publiées aujourd'hui le font.

Mais...

1. implémentation de base pour la plupart des logiciels, et des performances inférieures aux autres approches (R)QMC;
2. très peu utilisent la randomisation, du coup il n'y a pas d'estimation de l'erreur;
3. nombre de tirs utilisé sans se soucier de la taille de la population et pire, du nombre d'observations par individus.

Pour les données avec un grand nombre de répétitions par individus, c'est souvent une très mauvaise idée!

Comparaison de méthodes RQMC



Cas ind. ($s = 5$, $T_q = 1$): MC (· ·● ·), lattice-γ_u (—■—), lattice-0.1 (—▲—), lattice-$M32$ (- ♦ -), Sobol' nets (···★···) Halton-shift (- ▼ -), Halton-FL (- ⊗ -).

Facteur de réduction de MSE

$n = 10^4$.

Cas indépendant

s	5			10		15	
T_q	1	3	10	1	3	1	3
Halton-shift	43	9.0	10	7.5	2.6	3.4	1.5
Halton-FL	150	21	9.8	10	3.5	5.5	1.8
Sobol' nets	**300**	**32**	**15**	14	3.4	5.0	1.9
lattice-0.1	230	30	11	**18**	**4.4**	**7.9**	**2.4**

Cas corrélé (PCA)

s	5			10		15	
T_q	1	3	10	1	3	1	3
Halton-shift	59	15	12	13	3.1	6.2	2.1
Halton-FL	200	32	13	18	4.2	9.1	2.6
Sobol' nets	**400**	**60**	**20**	24	4.9	9.9	2.8
lattice-0.1	350	47	13	**27**	**5.4**	**11**	**3.3**