MATH 60604
Modélisation statistique
§ 5g - Comparaison de structures de covariance

Léo Belzile

HEC Montréal
Département de sciences de la décision

Résumé des modèles de covariance traités

- Nous avons vu comment ajuster cinq types de structure de covariance sur les résidus du modèle de régression:
  - Une structure d'équicorrélation (`cs`) (échangeable): toutes les paires d'observations ont la même corrélation.
  - Une structure AR(1): la corrélation entre deux observations décroit géométriquement avec le temps écoulé entre les deux.
  - Une structure ARH(1): une extension hétérogène du modèle AR(1), avec la même corrélation mais des variances différentes pour chaque temps.
  - Une covariance non structurée, qui permet une covariance différente pour chaque paire d'observations dans le temps.

Sélection de la structure de covariance

Modèle	$-2\ell_{\text{reml}}$	**AIC**	**BIC**
Indépendance	776,7	778,7	782,6
Équicorrélation	709,4	713,4	718,2
AR(1)	681,8	685,8	690,5
ARH(1)	675,3	687,3	701,6
Non structurée	659,3	689,3	725,0

- Les critères d'information AIC et BIC mène au choix du modèle avec la structure AR(1).
- Il s'agit d'un modèle parcimonieux (deux paramètres pour la structure de covariance) qui semble bien tenir compte de la corrélation intra-sujet.
- Ce qui est rassurant en plus, c'est que peu importe la structure utilisée, les conclusions quant aux effets des variables explicatives sont toujours les mêmes.

Choix de la structure de covariance

- Petit détail technique: si on veut comparer des modèles avec le AIC ou le BIC et que la méthode d'estimation REML est utilisée, il faut absolument que les modèles comparés contiennent les mêmes variables explicatives (effets fixes).
- Les critères d'information provenant de deux modèles ayant des effets fixes différents, estimés par REML, ne sont **pas** comparables. Par contre, ils sont comparables si on utilise la méthode du maximum de vraisemblance.

Remarques importantes sur les tests d'hypothèses sur la structure de covariance

- La plupart du temps, les questions de recherche font référence à des tests d'hypothèses concernant les paramètres de la partie "moyenne" du modèle.
- Dans le cas de données corrélées, nous savons maintenant qu'il faut que la partie covariance soit modélisée adéquatement afin que l'inférence sur la partie moyenne soit valide.
- Procéder en choisissant la structure de covariance selon des critères comme le AIC et le BIC est raisonnable. Mais il est aussi possible de faire des tests formels sur les paramètres de la structure de covariance.

Rappel sur les tests de rapport de vraisemblance (REML)

- Le test compare la log-vraisemblance du modèle complet (sous $\mathscr{H}_1$) à celle du modèle restreint emboîté (sous $\mathscr{H}_0$).
- L'hypothèse nulle est que le modèle restreint est une simplification adéquate du modèle complet.
- La statistique du test de rapport de vraisemblance est

$$D = 2\{\ell_{\text{reml}}(\widehat{\boldsymbol{\theta}}) - \ell_{\text{reml}}(\widehat{\boldsymbol{\theta}}_0)\}$$

- Sous $\mathscr{H}_0$, $D \dot{\sim} \chi^2_k$, où le nombre de degrés de liberté k est la différence du nombre de paramètres des deux modèles.
- On calcule la valeur-p du test à partir de la loi χ^2_k.

Tests du rapport de vraisemblance pour structures de covariance

- Il est possible de tester des hypothèses compliquées à l'aide du test de rapport de vraisemblance.
- Par exemple, on peut vérifier s'il est nécessaire de spécifier des variances différentes avec une structure autorégressive AR(1).
- Dans ce cas, on veut tester si le modèle AR(1) est adéquat ou si le modèle ARH(1) est préférable.
- L'hypothèse nulle est $\mathscr{H}_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_5^2$ contre l'alternative qu'au moins deux des variances sont différentes.
- Le modèle complet est le modèle ARH(1) ($\mathscr{H}_1$) et le modèle restreint le modèle AR(1) ($\mathscr{H}_0$).

Statistique du rapport de vraisemblance: AR(1) versus ARH(1)

- Sur la base des sorties précédentes, la différence de log-vraisemblance restreinte $-2\ell_{\text{reml}}$ pour ces deux modèles est $681.8 - 675.3 = 6.5$.
- Il y a **quatre** paramètres additionnels dans le modèle complet. On compare la valeur de la statistique 95% quantile d'une variable χ^2_4, soit 9.48.

Code SAS pour calculer la valeur-*p* avec la loi nulle χ^2_4

```
data pval;
pval=1-CDF('CHISQ', 6.5, 4);
run;
proc print data=pval;
run;
```

- On obtient une valeur-*p* de 0.165. On ne rejette pas $\mathscr{H}_0$ — le modèle AR(1) est une simplification adéquate du modèle ARH(1).

Remarques finales

- On a utilisé la méthode du maximum de vraisemblance restreinte (REML) pour estimer les paramètres de variance (option par défaut de `proc mixed`).
- Plusieurs modèles considérés sont emboîtés, donc on peut faire des tests pour établir des comparaisons.
  - par exemple, l'indépendance $\prec$ AR(1)$\prec$ ARH(1) $\prec$ non structurée.
- L'utilisation des critères AIC et BIC pour comparer ces modèles est valide **tant que** la structure de la moyenne inclut les **même** variables explicatives, comme ce fut le cas dans tous les modèles ajustés dans ces diapositives.
- Si on veut comparer des modèles qui incluent des variables explicatives différentes, il faudrait utiliser la méthode du maximum de vraisemblance et non pas la méthode REML.