Variables latentes dans les modèles linéaires généralisés

BoRdeaux, 2 et 3 juillet 2012

DJENEBA B. THIAM

Université Paris Descartes
MAP5 UMR CNRS 8145
IRD UMR 216

1ères $\mathbb{R}$encontres 

Intêret des variables latentes

Définition

- Variable hypothétique (dont on fait l'hypothèse)
- Variable non-mesurable
- Variable que l'on a pas observée dans notre échantillon

Interêt

- Synthétiser les données
- S'approcher au plus près des concepts théoriques sous-jacents à ce qui a été mesuré
- Comparer la théorie à l'échantillon

Exemple de variables latentes

- Variable latente continue: données non observées en raison de censures
- Variable latente discrète: classe latente

Gestion des variables latentes dans les modèles Glm

Modèles à variables latentes

- Variables latentes continues: modèles de réponse à l'item, modèles équations structurales, modèles Tobit.
- Variables latentes discrètes: modèles de profils latents, modèles de classes latentes.

Objectifs

- Gestion des variables latentes dans les modèles GLM par l'algorithme EM
- Utilisation des packages R disponibles (`lm`, `glm`, `lmer`)
- Idée: exploiter l'option `weights` de ces packages

Gestion des variables latentes dans les modèles Glm

Packages existant sous le logiciel R

- `mmlcr`
  - Estimation par algoritme EM
  - Première version (1.3.2) 2003; dernière version (1.3.5) avril 2006
- `lcmm` (latent class mixture model)
  - Proust, Jacqmin-Gadda (2005)
  - Estimation par algorithme de Marquardt
  - Première version (1.0) en 2010; dernière version (1.5.2) avril 2012
- `flexmix`
  - Leisch F (2004)
  - Estimation par algoritme EM
  - Première version (0.7-0) 2003; dernière version (2.3-8) mai 2012

Modèle et Notation

Rappel: Algorithme EM

Notation

- Y: Observations
- S: Variables latentes
- X: Covariable observée
- Θ: Paramètre à estimer

Algorithme

- Initialisation arbitraire de $\Theta^{(0)}$
- Étape E: calcul de la fonction auxiliaire $Q(\Theta|\Theta^{(t)})$
- Étape M: Calcul de $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(t)})$
- Répétition des étapes E puis M jusqu'à convergence des paramètres

Fonction auxiliaire

$$Q(\Theta'|\Theta) = \mathbb{E}\left[\log \mathbb{P}(Y, S; \Theta')|Y; \Theta\right]$$

Modèle et Notation

Rappel: modèle GLM sous R

Le modèle

$$\mathtt{y} \sim \mathtt{x} + \mathtt{s} \Leftrightarrow Y = X\beta + S\gamma + \varepsilon$$

Fonction utilisées sous R

- `lm(formula, data, subset, weights, ...)`
- `glm(formula, family = gaussian, data, weights, ...)`
- `lmer(formula, data,...weights,...)`
- Option `weights`: attribution de poids aux variables

Exemple: variable latente discrète binaire

Le modèle

- s variable latente binaire: $\mathtt{s} \in \{0,1\}^n$
- Probabilité à posteriori: $\mathtt{w} = \mathbb{P}(\mathtt{s}=1|\mathtt{x},\mathtt{y};\theta)$ $\qquad \mathtt{y} \sim \mathtt{x} + \mathtt{s}$

Algorithme EM

$$M(\theta) = \arg\max_{\theta'} \underbrace{\sum_{\mathtt{s}} \mathbb{P}(\mathtt{s}|\mathtt{x},\mathtt{y};\theta) \log \mathbb{P}(\mathtt{y},\mathtt{s}|\mathtt{x};\theta')}_{Q(\theta'|\theta)}$$

$$\begin{pmatrix} \mathtt{y} \\ \mathtt{y} \end{pmatrix} \sim \begin{pmatrix} \mathtt{x} \\ \mathtt{x} \end{pmatrix} + \begin{pmatrix} \mathtt{s}=1 \\ \mathtt{s}=0 \end{pmatrix} \quad \text{avec} \quad \mathtt{weights} = \begin{pmatrix} \mathtt{w} \\ 1-\mathtt{w} \end{pmatrix}$$

Sous R

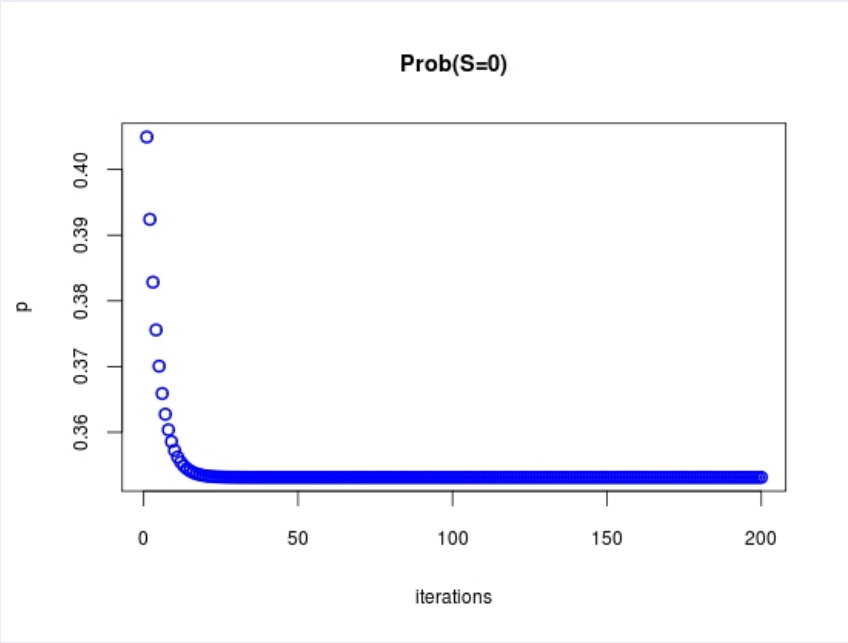
$$\mathtt{fit} = \mathtt{lm}(\mathtt{c}(\mathtt{y},\mathtt{y}) \sim \mathtt{c}(\mathtt{x},\mathtt{x}) + \mathtt{c}(\mathtt{s}=1,\mathtt{s}=0), \mathtt{weights} = \mathtt{c}(\mathtt{w}, 1-\mathtt{w}))$$

Code R

```
yy=c(y,y);
xx=c(x,x);
ss=c(rep(1,n),rep(0,n));
w=sample(0:1,size=n,replace=TRUE);
for (iter in 1:200) {
  fit=lm(y~x+s,data=data.frame(y=yy,x=xx,s=ss),
      weights=c(w,1.0-w))
  sigma=summary(fit)$sigma;
  f1=dnorm(y-fitted(fit)[1:n],sd=sigma);
  f0=dnorm(y-fitted(fit)[n+(1:n)],sd=sigma);
  p=mean(w); w=p*f1/(p*f1+(1-p)*f0);
}
```

$$\mathtt{w} = \frac{\mathbb{P}(\mathtt{s}=1)\mathbb{P}(\mathtt{y}|\mathtt{x},\mathtt{s}=1;\theta)}{\mathbb{P}(\mathtt{s}=1)\mathbb{P}(\mathtt{y}|\mathtt{x},\mathtt{s}=1;\theta)+\mathbb{P}(\mathtt{s}=0)\mathbb{P}(\mathtt{y}|\mathtt{x},\mathtt{s}=0;\theta)}$$

Exemple d'une variable latente binaire



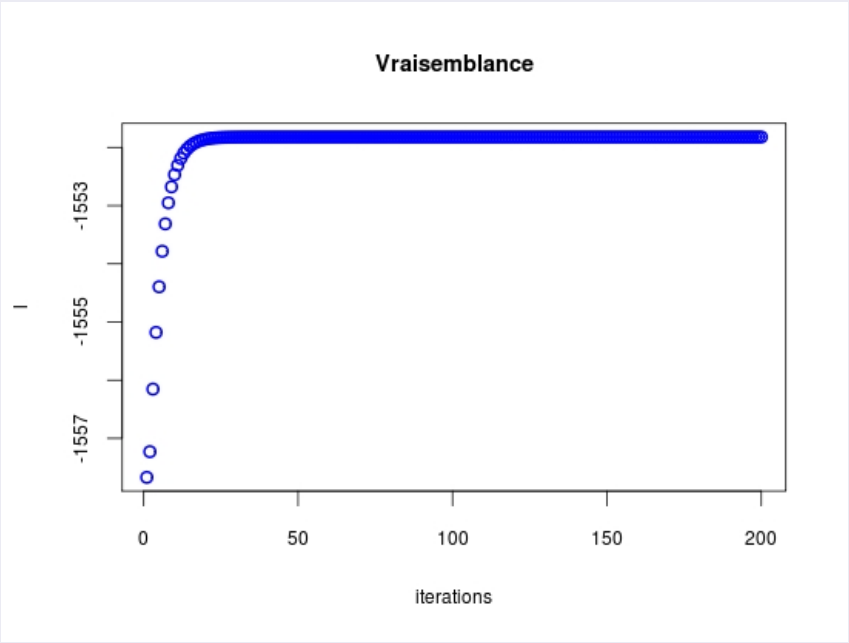


Figure: Convergence des paramètres du modèle et croissance de la fonction de vraisemblance $p = 0.25$; $n = 1000$; $a = 3.2$; $b = 1.5$; $\sigma = 0.7$. Le modèle:$Y \sim S$ avec $Y \sim \mathcal{N}(a + b * s, \sigma)$. $\hat{p} = 0.353$, $\hat{a} = 3.027$, $\hat{b} = 1.565$, $\hat{\sigma} = 0.410$

Conclusion

Avantages

- Méthodologie rapide à implémenter
- Utilisation des packages glm usuels
- Méthodologie flexible

limites

- Duplication des données dans le cadre continu
- UNU.RAN lent

Perspectives

- Dans le cadre continu, comment dupliquer les données.
- Méthode optimale pour découper une densité continue en n quantiles.
- Application a la validation d'essais cliniques

For Further Reading . . .

Cecile Proust-Lima, Benoit Liquet

lcmm: an R package for estimation of latent class mixed models and joint latent class models
cran.r-project.org/web/packages/lcmm/index.html, 2009.

Bettina Grün, Friedrich Leisch

FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters
Journal of Statistical Software, 28(4), 2008.

Kenneth A. Bollen

Latent variables in psychology and social sciences.
Annual Review of Psychology ,**53**, 605-634 2002.

A.P Dempster, N.M. Laird and D.B. Rubin

Maximum Likelihood from Incomplete Data via the EM Algorithm
Journal of the Royal Statistical Society, 1–38, 1977.