

## Estimation des paramètres dans les modèles IRT

christophe.lalanne@gmx.net

(20/12/2006)

L'objet de cette note est de décrire les principales méthodes d'estimation des paramètres d'un modèle de réponse à l'item. Nous considérerons successivement différentes classes de modèles, précisément :

**EMV conjointe.** Commençons dans un premier temps par la méthode JMLE [1].

$$L_{JML}(\beta, \theta) = \prod_{p=1}^P \prod_{i=1}^I \Pr(Y_{pi} = y_{pi})$$

Le principal désavantage de cette méthode est que les estimateurs des paramètres (des items) ne sont pas consistants, car le nombre de paramètres augmente avec la taille de l'échantillon<sup>1</sup>. Cet inconvénient est flagrant dans un contexte descriptif où l'on cherche à calibrer un instrument de mesure.

C'est la méthode utilisée par **Bigsteps**.

**EMV conditionnelle.** Dans le cas du modèle de Rasch, on dérive comme statistique "suffisante" pour l'effet spécifique de l'individu ( $\theta_p$ ) le score total  $s_p = \sum_{i=1}^I y_{pi}$  [2]. Après conditionnement, la probabilité d'observer un certain profil de réponse ne dépend pas de l'effet lié à l'individu, mais seulement de cette statistique suffisante. Par conséquent, l'effet spécifique lié à l'individu disparaît de la vraisemblance dite *conditionnelle* :

$$L_{CML}(\beta) = \prod_{p=1}^P \Pr(Y_{p1} = y_{p1}, \dots, Y_{pI} = y_{pI} | s_p)$$

La vraisemblance conditionnelle est maximisée par rapport à  $\beta$ .

Les estimateurs obtenus par cette méthode sont consistants [3], mais cette technique présente tout de même quelques désavantages, notamment dans un contexte de mesure. En effet, aucune inférence n'est possible sur la variable individu<sup>2</sup>.

Cette méthode ne peut pas être utilisée avec des modèles à 2 paramètres puisque ceux-ci n'incluent pas de statistique suffisante pour les paramètres liés aux individus (i.e. ce ne sont pas des GLMM).

Cette méthode d'estimation est disponible dans le package **eRm** de R.

**EMV marginale.** Dans cette approche, on considère les effets liés aux individus comme des tirages aléatoires effectués dans une densité de probabilité définie sur la population des individus. Cette densité, notée  $g(\theta_p | \psi)$ , est caractérisée par un vecteur de paramètres de population inconnus,  $\psi$ , qui doit être estimé comme les paramètres des effets fixes  $\beta_i$ . La vraisemblance à maximiser s'exprime sous la forme :

$$L_{MML}(\beta, \psi) = \prod_{p=1}^P \int_{-\infty}^{+\infty} \prod_{i=1}^I \Pr(Y_{pi} = y_{pi} | \theta_p) g(\theta_p | \psi) d\theta_p$$

Si la densité est discrète, l'intégrale doit être remplacée par une somme.

En fonction des hypothèses sur la densité de probabilité théoriques (non observée) des effets aléatoires, on distingue trois cas de figure :

- (a) l'approche non-paramétrique,
- (b) l'approche semi-paramétrique,

---

<sup>1</sup> chaque nouvel individu apporte un paramètre supplémentaire

<sup>2</sup> Une solution possible consisterait à considérer les paramètres liés à la variable item (après estimation) comme des valeurs connues, et de les utiliser dans la vraisemblance conjointe

(c) l'approche paramétrique.

(a) Dans le cas le plus général, l'EMV non-paramétrique ou entièrement semi-paramétrique [4] ne suppose aucune hypothèse sur  $g(\theta_p | \psi)$  — celle-ci n'est tout simplement pas spécifiée. Dans ce cas, il a été montré que l'estimée de la fonction de distribution  $G(\theta_p | \psi)$  est une fonction en escalier avec un nombre fini de pas. (b) Dans la méthode d'estimation semi-paramétrique, la position des pas est supposée connue mais les masses de probabilité en ces points définis doivent être estimés. (c) Dans la méthode d'estimation paramétrique, la densité de probabilité  $g(\theta_p | \psi)$  est choisie comme étant une densité paramétrique dont les paramètres sont à estimer. Dans la plupart des modèles, on supposera  $g(\theta_p | \psi) \sim \mathcal{N}(0; \sigma^2)$  ( $\sigma$  inconnu).

On notera qu'en assumant que les paramètres spécifiques des individus sont échantillonnés aléatoirement, le modèle initial dispose d'un paramètre supplémentaire. Si le modèle n'ajuste pas les données de manière satisfaisante, cet écart peut être dû au fait que la distribution postulée ne décrit pas correctement la vraie distribution des effets aléatoires. Dans ce cas, il est possible d'utiliser un mélange de lois normales comme distribution théorique pour les effets aléatoires.

Considérons l'EMV marginale avec une distribution normale des effets aléatoires. Soit  $\phi(\theta_p | \mu_\theta, \sigma_\theta^2)$ , où  $\mu_\theta$  désigne la moyenne (fixée à 0 par convention) et  $\sigma_\theta^2$  la variance inconnue. La probabilité d'un profil de réponse  $\mathbf{y}_p$  généré par la personne  $p$  sur l'ensemble des  $I$  items, conditionnellement à  $\theta_p$  est notée  $P(\mathbf{y}_p | \beta, \theta_p)$ , où  $\beta$  est un vecteur de dimension  $I$  contenant les effets fixes (un par item). Pour le modèle de Rasch,  $\Pr(y_p | \beta, \theta_p) = \prod_{i=1}^I \Pr(y_{pi} | \beta, \theta_p)$  (avec  $\Pr(y_{pi} | \beta, \theta_p) = \pi_i$  si  $y_{pi} = 1$ ), mais ce n'est pas vrai si l'hypothèse d'indépendance n'est pas vérifiée.

La vraisemblance marginale à optimiser s'écrit :

$$L(\beta, \sigma_\theta^2) = \prod_{p=1}^P L_p(\beta, \sigma_\theta^2) = \prod_{p=1}^P \int \Pr(y_p | \beta, \theta_p) \phi(\theta_p | 0, \sigma_\theta^2) d\theta_p,$$

où  $L_p(\beta, \sigma_\theta^2)$  désigne la contribution de la personne  $p$  à la vraisemblance marginale. On utilise généralement le logarithme pour faciliter la maximisation de la fonction. Notons que pour la plupart des modèles rencontrés, ce type d'intégrale ne possède pas de solution analytique (à la différence de celle apparaissant dans les modèles mixtes).

Il y a deux types d'approches à ce problème : la première consiste à approximer l'intégrale avec des méthodes d'intégration numérique, tandis que la seconde consiste à approximer l'intégrande de sorte que l'intégrale possède une solution analytique [5].

La première technique utilise une approximation de la vraisemblance, et on peut distinguer quatre manières de procéder, comme indiqué ci-dessous :

(a)	directe	(b)	indirecte
(c)	déterministe	(d)	stochastique

**Méthode directe.** Avec la méthode directe (a), on utilise une règle d'intégration numérique spécifique (approche déterministe, c). Dans le cas unidimensionnel, l'intégrale est remplacée par une somme finie d'aires rectangulaires qui permet d'approximer l'aire sous la courbe de l'intégrande. Comme les effets aléatoires sont supposés être distribués selon une loi normale, la méthode de Gauss-Hermite est la plus communément adoptée. L'approximation selon cette méthode donne :

$$L_p(\beta, \sigma_\theta^2) = \int \Pr(y_p | \beta, \theta_p) \phi(\theta_p | 0, \sigma_\theta^2) d\theta_p \\ \approx \sum_{m=1}^M \Pr(y_p | \beta, \sqrt{2}\sigma_\theta q_m) \frac{\omega_m}{\sqrt{\pi}},$$

où  $q_m$  et  $\omega_m$  sont les  $m$ -ième abscisses et poids de quadrature, respectivement. Les abscisses pour une quadrature gaussienne sont distribués et pondérés de manière optimale de sorte qu'avec  $M$  points l'approximation

est exacte si la fonction  $\Pr(y_p | \beta, \theta_p)$  est polynomiale de degré  $2M - 1$  ou moins. Les abscisses et les poids peuvent être trouvés dans [6] (cf. également [7], § 4.5).

Dans l'intégration numérique par la méthode de Gauss standard, les abscisses sont normalisées (et recentrés, mais ce dernier point n'a pas d'influence car la moyenne de population est 0) de sorte qu'ils couvrent l'ensemble du domaine de la distribution de la population. Mais cette normalisation est identique pour chaque individu  $p$ , ce qui n'est pas toujours le plus pertinent. En revenant à la forme de l'intégré,  $\Pr(y_p | \beta, \theta_p)\phi(\theta_p | 0, \sigma_\theta^2)$ , on constate qu'il s'agit de la distribution *a posteriori* (non normalisée) de  $\theta_p$  fonction des données et des paramètres des effets fixes. Si la valeur pour l'individu  $p$  est extrême (e.g. presque que des 1 ou des 0), la distribution *a posteriori* de  $\theta_p$  le sera également et déviara fortement de la distribution de la population, qui concentre plus de masse dans la région où sont localisées les valeurs modérées de  $\theta_p$ .

En conséquence, il serait plus approprié d'effectuer une normalisation (et un recentrage) individuelle. C'est en substance l'idée de l'*intégration numérique gaussienne adaptative* [8]. Pour chaque individu, on calcule une estimation empirique bayésienne de  $\theta_p$  (i.e.  $\hat{\theta}_p$ ) ainsi que la variance asymptotique de cet estimateur. Ces deux quantités sont calculés à partir de l'estimation actuelle des effets fixes en fonction des données. Ensuite, on peut réécrire la contribution de l'individu  $p$  à la vraisemblance marginale sous la forme :

$$\begin{aligned} L_p(\beta, \sigma_\theta^2) &= \int \Pr(y_p | \beta, \theta_p)\phi(\theta_p | 0, \sigma_\theta^2)d\theta_p \\ &= \int \frac{\Pr(y_p | \beta, \theta_p)\phi(\theta_p | 0, \sigma_\theta^2)}{\phi(\theta_p | \hat{\theta}_p, \hat{\tau}_p^2)}\phi(\theta_p | \hat{\theta}_p, \hat{\tau}_p^2)d\theta_p, \end{aligned}$$

où  $\hat{\tau}_p^2$  est la variance asymptotique de l'estimateur empirique de Bayes. Dans ce cas,  $\phi(\theta_p | \hat{\theta}_p, \hat{\tau}_p^2)$  est la distribution qui détermine la position et les poids des points de quadrature au lieu de  $\phi(\theta_p | 0, \sigma_\theta^2)$ . Ceci signifie que l'estimateur empirique de Bayes  $\hat{\theta}_p$  doit être ajouté aux points  $q_m$ , et les points doivent être multipliés par  $\sqrt{2}\hat{\tau}_p$ .

L'intégration gaussienne adaptative nécessite moins de points de quadrature car elle se concentre dans la région d'intérêt du continuum. Le prix à payer est que l'estimateur empirique de Bayes doit être évalué à chaque étape de l'algorithme, ce qui augmente bien évidemment le temps de calcul. On préférera donc souvent une règle d'intégration numérique régulière <sup>3</sup>.

Les deux principaux algorithmes permettant de maximiser la fonction de vraisemblance approximée (obtenue à partir d'une intégration adaptative ou non) sont l'*algorithme de Newton-Raphson* et le *score de Fisher*.

Une méthode alternative consiste à utiliser une *intégration de type Monte Carlo*. L'intégrale sur la distribution des effets aléatoires peut être vue comme l'espérance de la fonction  $\Pr(y_p | \beta, \theta_p)$  sur la variable aléatoire  $\theta_p$  distribuée normalement :

$$L_p(\beta, \sigma_\theta^2) = \int \Pr(y_p | \beta, \theta_p)\phi(\theta_p | 0, \sigma_\theta^2)d\theta_p = E(\Pr(y_p | \beta, \theta_p)).$$

Une espérance peut être estimée en tirant un échantillon aléatoire et en calculant sa moyenne empirique. Cela signifie que  $M$  valeurs de  $\theta_p$  sont tirés de la population, et par conséquent la quantité suivante est calculée :

$$L_p(\beta, \sigma_\theta^2) \approx \frac{1}{M} \sum_{m=1}^M \Pr(y_p | \beta, \theta_p^{(m)}),$$

avec  $\theta_p^{(m)}$  la valeur de  $\theta_p$  au point  $m$ . Cette procédure est l'équivalent stochastique (*d*) de l'intégration gaussienne avec recentrage et normalisation (i.e. non adaptative). L'intégration gaussienne adaptative possède également sa contrepartie stochastique [8], mais dans ce cas les tirages de  $\theta_p$  s'effectue dans la distribution  $\phi(\theta_p | \hat{\theta}_p, \hat{\tau}_p^2)$ .

**Méthode indirecte.** Dans le cadre des méthodes indirectes, l'optimisation de la (log)vraisemblance est transférée à une autre fonction pour laquelle on peut montrer que sa maximisation conduit à une augmentation de la vraisemblance marginale initiale. L'algorithme de maximisation indirecte le plus courant et qui

<sup>3</sup> Les deux méthodes donnent en général des résultats comparables dans le cas du modèle de Rasch [1], § 2

est appliqué dans le cadre des modèles à effets aléatoires est l’*algorithme EM* (“Expectation-Maximization”) [9].

Dans l’algorithme EM, l’ensemble des effets aléatoires de tous les individus  $\theta = (\theta_1, \dots, \theta_P)$  sont considérés comme des données manquantes et, avec les données observées  $\mathbf{y} = (y'_1, \dots, y'_P)'$ , ils forment les données complètes. Les effets aléatoires sont manquants et donc ne sont pas observés, de sorte qu’à chaque étape de l’algorithme, on commence par calculer la valeur attendue de la vraisemblance des données complètes, étant données les valeurs observées et les estimations des effets fixes  $\beta^{old}$  et  $\sigma_\theta^2^{old}$  obtenues à l’étape précédente, et des données observées. Il s’agit de l’étape E. Ensuite, la log-vraisemblance attendue est maximisée : c’est l’étape M. Chaque itération de l’algorithme EM consiste donc en une étape E, suivie d’une étape M, et ce cycle se poursuit jusqu’à la convergence.

L’espérance de la log-vraisemblance des données complètes,  $\ell_C(\beta, \sigma_\theta^2)$ , est définie comme :

$$\begin{aligned} E(\ell_C(\beta, \sigma_\theta^2) \mid y, \sigma_\theta^2^{old}) &= E(\log \prod_{p=1}^P (\Pr(y_p \mid \beta, \theta_p) \phi(\theta_p \mid 0, \sigma_\theta^2)) \mid y, \sigma_\theta^2^{old}, \beta^{old}) \\ &= \sum_{p=1}^P E(\log(\Pr(y_p \mid \beta, \theta_p) \phi(\theta_p \mid 0, \sigma_\theta^2)) \mid y, \sigma_\theta^2^{old}, \beta^{old}) \\ &= \sum_{p=1}^P \int (\log(\Pr(y_p \mid \beta, \theta_p)) + \log(\phi(\theta_p \mid 0, \sigma_\theta^2))) h(\theta_p \mid y, \sigma_\theta^2^{old}, \beta^{old}) d\theta_p, \end{aligned} \quad (a)$$

où  $h(\theta_p \mid y, \sigma_\theta^2^{old}, \beta^{old})$  est la densité conditionnelle des effets aléatoires connaissant les données observées, les estimations actualisées des paramètres fixes et la variance de la distribution des effets aléatoires. Après avoir calculé la log-vraisemblance attendue avec les données complètes (étape E), celle-ci est maximisée par rapport à  $\beta$  et  $\sigma_\theta^2$  (étape M).

Notons que l’intégrale impropre n’a pas disparu de la log-vraisemblance attendue des données complètes [eq. (a)]. Ainsi, l’intégrale doit-elle être approchée par une technique d’intégration gaussienne ou de type Monte Carlo.

Pourquoi utiliser l’algorithme EM dans ce cas ? Celui-ci offre trois avantages. Premièrement, cet algorithme garantit qu’à chaque itération la log-vraisemblance marginale augmente, bien que l’algorithme ne la maximise pas directement [5]. Cela rend l’algorithme numériquement très stable. Ce n’est pas garanti lorsque l’intégrale n’est qu’une approximation. Deuxièmement, la log-vraisemblance attendue de l’équation [a] est écrite sous la forme d’une somme d’une composante décrivant les paramètres des effets fixes et d’une composante décrivant le paramètre de variance. Cela signifie que l’estimation de ces deux ensembles de paramètres peut être effectuée séparément durant l’étape M, ce qui réduit la dimension du problème d’optimisation. En dernier lieu, l’étape M dans l’algorithme EM donne des solutions admissibles pour certains paramètres. Pour les composantes de variance sous une hypothèse de normalité, une telle solution existe. Pour les paramètres qui ne possèdent pas de solutions admissibles, il est nécessaire de recourir lors de l’étape M à une méthode d’optimisation itérative, par exemple la méthode de Newton-Raphson.

Un désavantage de l’algorithme EM est que la convergence vers le maximum n’est généralement pas très rapide, en particulier au voisinage du maximum de la vraisemblance marginale. Il existe des variantes de l’algorithme EM qui permettent d’accélérer la convergence ou de faciliter le calcul de l’étape de maximisation.

Dans le contexte de la modélisation traditionnelle de réponse à l’item en considérant seulement les indicatrices des items comme variables prédictrices, l’application de l’algorithme EM présente un autre avantage. Si l’on considère le modèle de Rasch (mais cela reste vrai avec un modèle 2PL), le vecteur des paramètres des items  $\beta$  peut être subdivisé en  $I$  sous-ensembles disjoints de paramètres (dans ce cas, des paramètres individuels),  $\beta_1, \dots, \beta_I$ , chacun étant associé à un item. Étant donné l’effet aléatoire  $\theta_p$ , il y a indépendance conditionnelle, et par conséquent, la log-vraisemblance attendue peut s’écrire comme une somme de termes indépendants — un pour chaque item — et chacun peut être maximisé séparément. Cette propriété permet d’analyser des jeux de données avec un grand nombre d’items (e.g. 50 ou plus), ce qui serait autrement impossible. La même propriété s’applique pour les paramètres des individus dans le modèle. La composante liée aux individus dans le modèle de régression peut être vue comme la moyenne non nulle d’une distribution normale, et de ce fait les coefficients de régression peuvent être estimés séparément de la difficulté des items. Ceci explique la popularité de l’estimation MML avec EM dans le domaine de la psychométrie.

L'objet de l'approximation de l'intégrande est d'obtenir une expression telle que l'intégrale de l'approximation possède une solution admissible. Deux types de techniques sont présentées dans les paragraphes qui suivent : la méthode de Laplace et une classe de méthodes appelée méthodes de pseudo-vraisemblance.

## Références

- [1] P. De Boeck & M. Wilson, *Explanatory Item Response Models* (New York: Springer-Verlag, 2004).
- [2] E. B. Andersen, *Discrete Statistical Models with Social Science Applications* (Amsterdam: North-Holland, 1980).
- [3] E. B. Andersen, Asymtotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society, Series B*, 32 (283–301, 1970).
- [4] T. Heinen, *Latent Class and Discrete Latent Trait Models: Similarities and Differences* (Thousand Oaks, CA: Sage, 1996).
- [5] K. Lange, *Numerical Methods for Statisticians* (New York: Wiley, 1999).
- [6] M. Abramowitz & I. Stegun, *Handbook of Mathematical Functions* (New York: Dover Publications, 1974).
- [7] K. Lange, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, 1992).
- [8] P. C. Pinheiro & D. M. Bates, *Mixed-Effects Models in S and S-PLUS* (New York: Springer, 2000).
- [9] A. P. Dempster, N. M. Laird & D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion) *Journal of the Royal Statistical Society, Series B*, 39 (1–38, 1977).