Effects of Ordered Scores in Rasch Analysis

Effetti di Ordinamenti dei Punteggi in Analisi di Rasch

Lucio Bertoli-Barsotti
Department of Mathematics, Statistics, Computer Science and Applications
University of Bergamo, I-24127 Bergamo, lucio.bertoli-barsotti@unibg.it

Riassunto. Si considera la matrice dei punteggi -con risposte politomiche su scala Likert- ordinata in base ai totali di riga e colonna. Si presentano due risultati concernenti le corrispondenti (stime di massima verosimiglianza delle) misure di Rasch di item e di soggetto. 1) La piena ordinabilità dei patterns di risposta è la nota condizione di struttura di Guttman *perfetta* (PGS); si introduce la nozione di struttura di Guttman *in senso debole* (WGS; una PGS è, a maggior ragione, una WGS) e si mostra che essa identifica -per certi versi paradossalmente- la condizione di *non esistenza* delle misure di Rasch. 2) Punteggi totali su scala Likert e misure di Rasch sono messi a confronto da un punto di vista dell'ordinamento. Mentre il Rating Scale Model mantiene una relazione di co-monotonicità, ciò non accade per il Partial Credit Model.

Keywords: Rasch analysis; Likert scoring; Guttman scaling; rating scale model; partial credit model; ordering; maximum likelihood estimate.

1. Introduction

Unobservable latent traits such as consciousness, dependence, achievement, interest, aptitude, needs, preferences, etc., - here referred to, generically, as 'ability', from now on - are hypothetical constructions that have to be defined operationally by the methods used to measure them. Typically, the measurement of the 'ability' may be deduced from the responses of the subject to a suitable set of stimuli (items). Following a very common testing *practice*, numbers are used to 'score' the item responses; in particular, in this paper we shall consider the case of polytomous ordered item scores, with a fixed number of answer categories. Let the ordered categories $\mathrm{C}_0, \mathrm{C}_1, \ldots, \mathrm{C}_h, \ldots, \mathrm{C}_m$ be scored, respectively, with the numbers $0, 1, \ldots, h, \ldots, m$. Then a *total score* (or *simple unweighted sum score* or *raw score*, or *Likert score*) is obtained by adding up these numbers, for each respondent and each item. Obviously, simply adding integer scores does not guarantee a proper scale. The transformation of the total scores into Rasch 'measures' will be considered in this paper. In particular, we shall refer to the two most popular polytomous Rasch models: the Rating Scale Model (RSM) and the Partial Credit Model (PCM).

It is known that Item Response Theory (IRT) models have a number of advantages over the classical test theory models. In particular, for the Rasch models these include the following remarkable facts: 1) interval-level measurements (moreover, under weak conditions Fischer, 1995, proved that the *only* IRT model that will produce interval-level measurement is the Rasch model); 2) test-free/sample-free measurements (also called person *invariance* of item parameters and item *invariance* of person parameters), i.e. the measure of the ability of the respondents does not depend upon the set of items

selected for the test and, symmetrically, the measure of the –let us say- ‘easiness’ of the items does not depend upon the particular set of respondents; 3) estimate of a standard error of each item/person measurement and a control over the item (and the test) information function; 4) Rasch models can be falsified.
Let the items be ordered according to their total scores. Then, as the individual response pattern is consistent with the ordering of the items, the person total score can be used to serve as a means of measuring the construct (for a recent comparative study considering the efficiency of the total scores -in the dichotomous case- see Cox and Wermuth, 2002). What difference is to be expected between raw scores and Rasch-based measures from a rank order point of view? As proved in section 3, the answer depends on the model chosen (and perhaps on the method of estimation too) and unlike the dichotomous case, there may be a lack of co-monotonicity between the two methods of scoring, when we consider the case of polytomous Rasch models.
Regarding the above mentioned principle of ‘consistence’, the lack of order in the response pattern is a cause of misfit. By a standard procedure of Rasch Analysis, disordered response patterns -e.g. patterns of respondents that tend to obtain low scores on easy items and vice versa- are considered ‘odd’ (unexpected) and are excluded from the sample, for the sake fitting in with model. In this sense, Theorem 1 in the next section describes the somewhat surprising situation of a trouble due, paradoxically, to a pathological ‘excess’ of ordering in response patterns.

2. Guttman Scaling and ill-conditioned data matrices

Typically, IRT assumes an *unidimensional* latent trait -a single dominant construct triggering a response behavior pattern. The Guttman scaling is also based on the idea of a unidimensional scale. If items are dichotomously scored, in a *perfect* (deterministic) unidimensional scale the pattern of the responses of a subject can be exactly reproduced simply by knowing his total score – if the items are ordered according to the total scores.

Perfect Guttman Structure (PGS). The response matrix satisfies the PGS if, for *each* pair of respondents, A and B, A endorses all items endorsed by B.

As a consequence of the PGS, the total score of A is higher (not lower) than the total score of B (note that ties are allowed). This principle may be easily extended to the case of polytomously scored items by dichotomizing the data at each possible item response level - i.e. by merging the categories from C_0 to C_{h-1} into a new category with score 0, and the categories from C_h to C_m into a new category with score 1 $(h = 1,...,m)$. The condition of PGS for a response matrix with polytomously scored items is said to be satisfied if and only if, for each pair of respondents, A and B, *and for every h*, $h = 1,...,m$, A endorses all items endorsed by B.
A slightly weak condition may be introduced (obviously, the PGS is seldom found in practice). The following notion refers to binary data - but analogous extension to the case of polytomously scored items is possible, as above.
Weak Guttman Structure (WGS). The response matrix satisfies the WGS if a partition exists (that may not be unique) of the set of the respondents into two non-empty subsets,

G_1 and G_2, so that for *each* pair of respondents, A and B, A endorses all items endorsed by B, whenever A belongs to G_1 and B belongs to G_2.

Clearly, the PGS implies the WGS. Interestingly enough, the Guttman scaling is strictly related to the structure of an *ill-conditioned* response matrix, i.e. a matrix for which the Rasch measures do *not exist*. Here Rasch measures is intended to denote the maximum likelihood (ML) estimates of the item/person parameters, when PCM is used. In this sense, in theory, the Rasch measures are not always well-defined. The following theorem is a straightforward consequence of the WGS notion and a recent result (Bertoli-Barsotti, 2004).

Theorem 1. Let the response matrix be *admissible* (i.e. without null categories). Then the Rasch measures are *not* defined if and only if the response matrix satisfies the WGS.

3. Rasch measures vs total scores

In the last few years several papers have been devoted to the comparison of Rasch measures and total scores - from the point of view of precision, sensitivity to change, discriminating power, and so on (see, for instance, Raczek et al., 1998; Cook et al., 2001, White et al., 2002). Authors are often confident of the somewhat minimal relation of co-monotonicity between the two scoring methods. This co-monotonicity has been recently proved by Bertoli-Barsotti (2003) for the case of the dichotomous Rasch model. The result remains valid both for conditional and unconditional ML estimation approaches. Nevertheless, an extension to the case of polytomously scored items is problematic: it may be proved that the property continues to hold for the RSM but not for the PCM.

Let $P_{\upsilon ih}$ the probability of person υ, with person parameter ϑ_υ, scoring *h* on item *i* - characterized by the *m* threshold parameters δ_{ih} $(h = 1,...,m)$- the PCM is defined by the log odds $log(P_{\upsilon ih} / P_{\upsilon ih-1}) = \vartheta_\upsilon + \delta_{ih}$. Let $\delta_{ih} = \alpha_i + \tau_{ih}$, with $\sum_{h=1}^{m} \tau_{ih} = 0$; α is the item parameter (representing the easiness of the item); the δ 's and the τ 's are called, respectively, uncentralized and centralized threshold parameters. A *third* parameterization of the model is possible, by taking $\beta_{ih} = \sum_{t=0}^{h} \delta_{it}$, with $\delta_{i0} = \beta_{i0} = 0$.

With the latter parameterization, one finds $P_{\upsilon ih} = exp(h\vartheta_\upsilon + \beta_{ih})[\sum_{h=0}^{m} exp(h\vartheta_\upsilon + \beta_{ih})]^{-1}$.

When *m*=1 the dichotomous Rasch model is obtained; when $\tau_{it} = \pi_t$, $\forall i \forall t$, the PCM gives the RSM.

Theorem 2. (i) The ML estimates of person and item parameters of the RSM are co-monotone with the corresponding total scores. (ii) The ML estimates of person parameters of the PCM are co-monotone with the corresponding total scores.

Sketch of proof. Let $\boldsymbol{x}'_{\upsilon i} = (x_{\upsilon i0}, x_{\upsilon i1}, ..., x_{\upsilon ih}, ..., x_{\upsilon im})$ be the response vector, where $x_{\upsilon ih} = 1$ and $x_{\upsilon ij} = 0$, $j \neq h$, if the response is in category C_h. Moreover let $r_\upsilon = \sum_{ih} h x_{\upsilon ih}$ and $c_i = \sum_{\upsilon h} h x_{\upsilon ih}$. The log-likelihood function for the PCM is $l = l(\boldsymbol{\vartheta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \; = A + B - C$, where $A = A(\boldsymbol{\vartheta}) = \sum_\upsilon \vartheta_\upsilon r_\upsilon$, $B = B(\boldsymbol{\alpha}, \boldsymbol{\tau}) = \sum_i \alpha_i c_i + \sum_{\upsilon ih} \sum_t^h \tau_{it} x_{\upsilon ih}$ and $C = C(\boldsymbol{\vartheta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) =$ $\sum_{\upsilon i} \log \sum_h \exp \sum_t^h (\vartheta_\upsilon + \alpha_i + \tau_{it})$. For the case of the RSM, the new terms $B_1 =$ $\sum_i \alpha_i c_i + \sum_{\upsilon ih} \sum_t^h \pi_t x_{\upsilon ih}$ and $C_1 = \sum_{\upsilon i} \log \sum_h \exp \sum_t^h (\vartheta_\upsilon + \alpha_i + \pi_t)$ substitute, respectively, B and C. For the sake of convenience, let the data matrix be arranged so that $\boldsymbol{r} = \boldsymbol{r}_\uparrow$, $\boldsymbol{c} = \boldsymbol{c}_\uparrow$ (where $\mathbf{v}_\uparrow$ denotes the vector $\mathbf{v}$ with its components in increasing order). Hence the proof of (i)-(ii) follows (see Bertoli-Barsotti, 2003) from the inequalities $A(\boldsymbol{\vartheta}_\uparrow) \geq A(\boldsymbol{\vartheta})$, $B_1(\boldsymbol{\alpha}_\uparrow, \boldsymbol{\pi}) \geq B_1(\boldsymbol{\alpha}, \boldsymbol{\pi})$, by the invariance of C_1 with respect to rearrangements of persons and items and the invariance of C with respect to rearrangements of persons.

Counter-examples may be given to prove that the co-monotonicity does not hold for the item parameters of the PCM.

References

Bertoli-Barsotti L. (2003) An order-preserving property of the maximum likelihood estimates for the Rasch model, *Statistics & Probability Letters*, 61, 91-96.

Bertoli-Barsotti L. (2004) On the existence and uniqueness of JML estimates for the partial credit model, *Psychometrika*, to appear.

Cook K.F., Gartsman, G.M., Roddey, T.S., Olson, S.L. (2001) The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation, *Archives of Physical Medicine and Rehabilitation*, 82, 1558-1565.

Cox D.R., Wermuth, N. (2002) On some models for multivariate binary variables parallel in complexity with the multivariate Gaussian distribution, *Biometrika*, 89, 462-469.

Fischer G. H. (1995) Derivations of the Rasch Model, in *Rasch Models. Foundations, recent developements, and applications*, G. H. Fischer & I. Molenaar (Eds.), Springer-Verlag, New York, 15-38.

Raczek A.E., Ware, J.E., Bjorner, J.B., Gandek, B., Haley, S.M., Aaronson, N.K., Apolone, G., Bech, P., Brazier, J.E., Bullinger, M., Sullivan, M. (1998) Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project, *Journal of Clinical Epidemiology*, 51, 1203-1214.

White L.J., Velozo, C.A. (2002) The use of Rasch measurement to improve the Owestry classification scheme, *Archives of Physical Medicine and Rehabilitation*, 83, 822-831.