JOURNAL OF APPLIED MEASUREMENT, *11*(1),

Predicting Responses from Rasch Measures

John M. Linacre
University of Sydney, Australia

There is a growing family of Rasch models for polytomous observations. Selecting a suitable model for an existing dataset, estimating its parameters and evaluating its fit is now routine. Problems arise when the model parameters are to be estimated from the current data, but used to predict future data. In particular, ambiguities in the nature of the current data, or overfit of the model to the current dataset, may mean that better fit to the current data may lead to worse fit to future data. The predictive power of several Rasch and Rasch-related models are discussed in the context of the Netflix Prize. Rasch-related models are proposed based on Singular Value Decomposition (SVD) and Boltzmann Machines.

Requests for reprints should be sent to

Introduction

The estimation of Rasch measures from ordinal data has been exhaustively investigated for 40 years, but the accuracy of prediction of polytomous ratings from those measures has not attracted as much attention.

Georg Rasch (1961) proposes a multi-dimensional logistic model for multinomial data. Erling Andersen (1977) implies that equidistant scoring of categories is required for the marginal scores to be sufficient statistics for measures to be located on the real line, i.e., to be unidimensional. Andrich (1977) makes this explicit by formulating a Rasch model for polytomous ordinal response categories. This was extended by Masters (1982) to model a different response structure for each test item, and the multinomial Rasch model continues to be yet further extended by numerous researchers (Rost, 2001).

But, given a dataset of polytomous data, which model best predicts future ratings?

The Basic Andrich Model

From a Rasch perspective, the simplest polytomous model is the Andrich (1978) model, expressed here in logit-linear form:

$$\log (P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_j \qquad (1)$$

where P_{nij} is the probability that person n encountering item i is observed in category j of a set of ordered response categories $j = s + 1, s + m$. So that the rating scale categories are numbered $s, s + m,$ a consecutive ascending sequence of ordinal numbers. For algebraic convenience, $s = 0$ here.

$P_{ni(j-1)}$ is the probability that person n encountering item i is observed in category j–1.

B_n is the ability of person n.

D_i is the difficulty of item i.

F_j is the Rasch-Andrich threshold located at the point of equal probability of categories j–1 and j. The set of $\{F_j\}$ is termed here the "rating scale structure".

It is conventional to set $\Sigma\{F_j\} = 0$ for $j = 1, m,$ so that the item difficulty is the point on the latent variable at which the lowest and highest categories are modeled to be equally probable. In manipulating (1), the term F_0 may arise. This can be set to $F_0 = 0$ or any convenient value, because it cancels out algebraically. The $\{F_j\}$ are modeled to be as independent as possible of both the items and the persons. So they can be conceptualized as a rating-scale structure shared by all items, or as a response-style structure shared by all persons.

The Andrich model is simple to express as Newton-Raphson estimation equations (Wright and Masters, 1982). Its estimates are relatively straightforward to communicate and utilize, so it would be beneficial if this model could be implemented for all data sets, but it cannot be, as we shall discuss.

Category Widths

The depiction of category widths is central to communicating how the rating scale relates to the latent variable. Since the latent variable is infinite, the extreme categories, 0 and $m,$ of the rating scale are always infinitely wide. For the intermediate categories, there are several ways of expressing their widths, depending on the purpose for which the width is being estimated.

The $\{F_j\}$ indicate the placements of the categories on the latent variable. The $\{F_j\}$ themselves are the points of equal-probability of adjacent categories so, if the $\{F_j\}$ are a series of ascending values, they indicate the modal region for each category and so the ends of the modal intervals, $\{M_j\}$, on the latent variable for which each category is more probable to be observed than any other category. When the $\{F_j\}$ exhibit disorder, then some categories will not be modal, and so the $\{M_j\}$ will differ from the $\{F_j\}$ and not depict all categories. Segmenting the latent variable into modal category intervals is useful for inference, but requires careful communication. When the non-specialist is told "category x is the most probable category," this statement may be understood to mean "category x is more probable than all other categories combined". But this may

not be the case. In fact, the highest probability of a modal intermediate "most probable" category of a 4-category rating scale may be $p = 0.3$ or less.

The highest probability of the extreme categories of a rating scale is always modeled to be 1.0 at the infinite extremes of the latent variable. This is an extrapolation beyond the data that can be misleading for inference. It implies that the rating scale definitely functions in that manner at its extremes. However, an extreme category may have only a few observations, and those observations may provide scant evidence of its functioning. An example of such a rating scale is "Excellent, Good, Acceptable, Needs improvement." A rater could perceive "Needs improvement" to be the appropriate category for a "Good"-rated performance by a star performer who should be in the "Excellent" category. Category-level fit statistics could flag this problem, but they are rarely provided to the end-user of an instrument.

The end-user may wish to compare performance on an item to a criterion level on the rating scale. Is the performance "above or below a category threshold" or would we predict it to be? The most direct way to answer this is to conceptualize the width of the intermediate categories on the latent variable in terms of Rasch-Thurstone thresholds, the points of equal cumulative category probability above and below the threshold. Thus, at Rasch-Thurstone threshold T_t relative to D_i, where the person ability is B_t,

$$\sum_{k=0}^{j} P_{tik} = 0.5 = \sum_{k=j+1}^{m} P_{tik} . \qquad (2)$$

Expressing the Andrich model in terms of Rasch-Thurstone thresholds is awkward. For instance, consider a three-category rating scale with Rasch-Andrich thresholds, F_1 and F_2 and with the conventional constraints so that $F_1 = -F_2$. Then the equivalent Rasch-Thurstone thresholds are T_1 and T_2, so that $T_1 = -T_2$, and T_1 is given by

$$\begin{aligned} \log(P_{ni1}/P_{ni0}) &= B_n - D_i - F_1 \\ &= B_n - D_i - D_i - T_1 + \log\left(e^{2T1} - 1\right). \end{aligned} \qquad (3)$$

It is seen that the T_1 and T_m, the extreme thresholds, are always more extreme than F_1 and F_m, and that the $\{T_j\}$ are always ascending or equal in value, unlike the $\{F_j\}$ which can be disordered.

In another conceptualization of category widths, it is the "average" performance on the rating scale that is of interest. "What is the predicted average rating on the rating scale of a person of such-and-such ability?" To answer, this, let us also express the Andrich model in terms of Rasch-half-point thresholds $\{H_h\}$, where $h = 1,m$. The $\{H_h\}$ are the points at which the expected item score on the model item characteristic curve (ICC) is $\{h - 0.5\}$. Then, if B_h is the person ability at point H_h relative to item difficulty D_i,

$$\sum_{k=0}^{m} kP_{hik} = h - 0.5 . \qquad (4)$$

It is seen that the rating-score interval $h - 0.5$ to $(h + 1) - 0.5$ contains performances which average within half a score-point of the category value, *h.* Since only discrete category values can be observed, performances in this interval can be considered to round to the category value, *h.* Thus H_h to H_{h+1} is the category interval on the latent variable for category *h* according to this conceptualization. The $\{H_j\}$ are always strictly ascending in value. This formulation has proven effective for communicating rating scale functioning because it can be explained with only one curve, the ICC. Its explanation can be presented in frequentist terms: "If there were 1,000 people at this point on the latent variable there average performance would be in the region of category *h*." This one-curve explanation contrasts with the set of probability curves required to explain the Rasch-Andrich and Rasch-Thurstone Thresholds. For non-technical audiences, understanding interactions between curves can prove daunting.

For our 3-category rating scale,

$$\begin{aligned} \log(P_{ni1}/P_{ni0}) &= B_n - D_i - F_1 \\ &= B_n - D_i - H_1 + \log\left(2e^{2H1} - 0.5\right). \end{aligned} \qquad (5)$$

It is seen that H_1 and H_m must be yet more extreme than T_1 and T_m.

The relationship between these three sets of thresholds, $\{M_j\}$, $\{T_j\}$, $\{H_j\}$, is shown in Figure

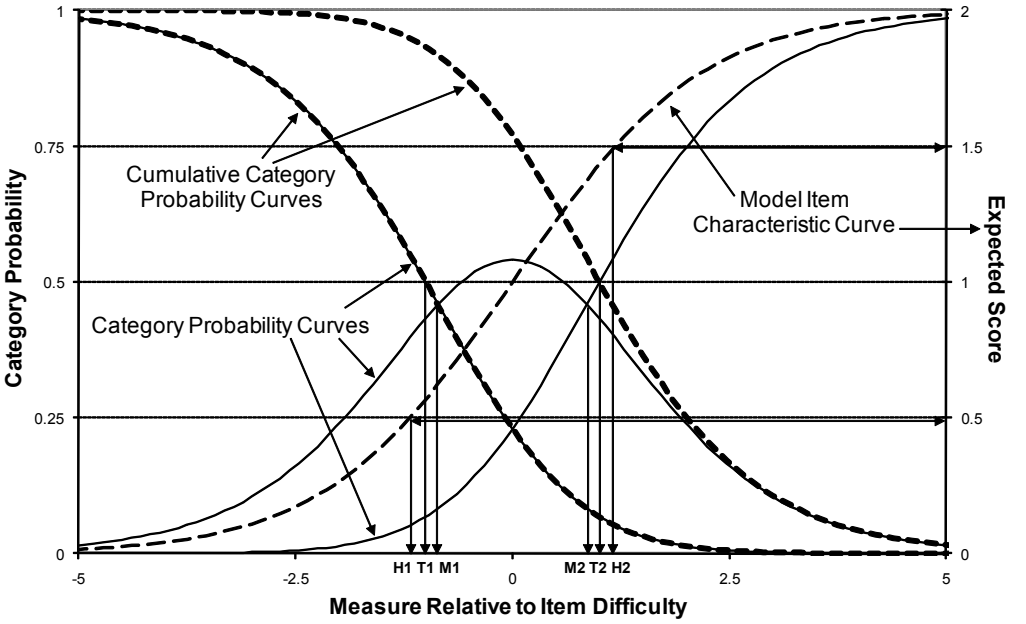

Figure 1. Category Intervals on the Latent Variable

1 for the "Liking for Science" data (Wright and Masters, 1982).

Structural Considerations in Rating Prediction

Items on a test instrument may not share the same rating scale structure. Most conspicuously, items may have different numbers of response categories, for instance, 3 categories, *yes – may be – no, or* 4 categories, *never – sometimes – often – always*. Modeling items in groups according to the lengths of their rating scales addresses this:

$$\log\left(P_{ngj} / P_{ng(j-1)}\right) = B_n - D_{gi} - F_{gj}, \quad (6)$$

where *g* indicates the group of items sharing the same rating-scale length, so that in group *g,* the categories range from 0 *to* m_g.

Items may share the same number of categories, but the categories themselves may have different meanings. Accordingly the grouping may also be particularized to groups of items considered to share the same substantive rating scale. At their most diverse, each item may become its own group, so that each item is modeled to with a unique rating scale structure. This is the Masters (1982) Partial Credit model, originally intended for giving partial credit when partially-correct distractors are selected in response to multiple-choice questions. Here is Masters' Partial Credit model:

$$\begin{aligned}\log\left(P_{nij} / P_{ni(j-1)}\right) &= B_n - D_i - F_{ij} \\ &= B_n - D_{ij}, \end{aligned} \quad (7)$$

where D_{ij} is the Rasch-Andrich threshold *j* for item *i* relative to the latent variable.

Similarly, each person can be model to have a unique perception of a supposedly shared rating scale. Ben Wright called this the person-style model:

$$\begin{aligned}\log\left(P_{nij} / P_{ni(j-1)}\right) &= B_n - D_i - F_{nj} \\ &= B_{nj} - D_i, \end{aligned} \quad (8)$$

Many rating scales can be conceptualized as progressive. To be observed in category 2, the person must have succeeded on category 1. This suggests that such a scale should produce better

fit to a model with those characteristics. Such a model is the Steps-success model (Verhelst et al., 1997) which has been implemented in software since 1991 (Glas and Verhelst, 1991). The model can be written:

$$\log\left[\sum_{k=j}^{m} P_{nik} \Big/ P_{ni(j-1)}\right] = B_n - D_i - F_{ij}. \quad (9)$$

It is seen that the category threshold is the log-odds of passing the threshold, i.e., being observed in a category at or above the threshold, and of failing the threshold, i.e., being observed in the category immediately below the threshold. This model suggests a Steps-failure (Linacre, 1991) model for rating scales modeling degrading performance. Here, being observed in a lower category implies lack of success in the category above it:

$$\log\left[P_{nij} \Big/ \sum_{k=0}^{j-1} kP_{nik}\right] = B_n - D_i - F_{ij}. \quad (10)$$

But which of these models matches the content of the rating scale? Consider the 7-level Functional Independence Scale (FIM®), used for rating medical patients during rehabilitation. Its 7 levels are shown in Table 1. Is this a scale in which the observed Level is reached from the full set of levels presented simultaneously in accordance with the Andrich model? Or is the Level reached by a process of increasing ability upwards from Total Assist in accord with the Steps-success model? Or is the Level reached by decreasing ability downwards from Complete independence in accord with the Steps-failure model? There is no clear-cut answer, and this ambiguity has proved pervasive in practice. Consequently, even for apparently strongly developmental scales, the Andrich rating scale has proved an effective model.

Table 1

Functional Independence Measure, FIM®

Rating Scale Level
7: Complete independence
6: Modified independence
5: Supervision
4: Minimal assist
3: Moderate assist
2: Maximal assist
1: Total assist

Fit Considerations

The statistical tradition is to select the model which best fits the data. An obvious approach is to compare the chi-square statistics of models with different numbers of estimated parameters. If the addition of a parameter reduces the chi-square by more than one unit, then the model is better fitting, but this tends to produce better-fitting models with many more parameters but only a small improvement in overall fit. This has motivated model comparison by means of the Akaike Information Criterion (AIC, Akaike, 1973) but, so far, this has not been reported in the Rasch literature as a notably successful approach to improving prediction.

Utility and Communication Considerations

Rasch analysis constructs linear measures. Producing a precise numerical summary of the current data set is important, but less so than producing a summary whose numbers have special properties. All the numbers we encounter in our day-to-day lives are linear, when considered only as numbers, but most numbers are not linear in their meanings. Rasch measures are numbers that are linear in meaning. This makes them useful in ways that other numbers are not. For instance, linear numbers are what most statistical methods expect.

The linearity of Rasch measures also makes them excellent for communicating findings. But this communication is undermined when there are too many numbers, or the numbers are too fragmented. A non-technical audience can understand one number per person and one number per item and one curve per rating scale, but when these multiply confusion soon reigns. For instance, for an instrument with a standard Likert 5-category rating scale, the audience expects to see one description of the relationship of the categories. It is disconcerting to be shown slightly different Partial Credit representations of the rating scale for each item on the instrument. If the rating scale

functioning for one item is noticeably different, then there is probably a mismatch between the item stem and the rating scale. For instance, the item stem may be asking a true-false question, but the rating scale is asking for degree of agreement. In this case, the item can be separated out for special attention, but, in communicating the instrument, it is probably better if such items are omitted entirely.

Predicting Future Responses

The previous considerations in model selection are well established in Rasch practice. They are always based on the analysis and reporting of currently existing data. But a new situation has arisen in which it is the prediction of future responses, yet to be collected, that is the paramount consideration. A fundamental assumption of much of statistics is that the model which best, or most usefully, fits the current data will also be the most successful at predicting future data. And this is true, up to a point. That point is when the model starts to over-fit the current data. If the model fits the current data too well, predicting too accurately its idiosyncratic aspects, then that model will fit future data less well because there will be different idiosyncratic aspects. But when has overfit occurred? In practice this can only be discovered by analyzing new data.

It is to be expected that each new dataset will produce slightly different estimates of the parameters, and so slightly different predictions about responses. These new estimates are generally expected to fluctuate around the previous estimates in a distribution described in some way by the probabilistic form of the model and the standard errors of measurement. Of course, there are always unmodeled sources of variance, so we are not surprised if the new estimates are somewhat more dispersed from their expectations than Rasch theory predicts. If the dispersion becomes too large, we start looking for item drift, differential item functioning, and other effects, but nearly always retrospectively. We rarely try to predict these in advance and to build them into our response predictions for future data.

The Netflix Prize

In October 2006, Netflix Inc. offered a one million dollar challenge prize to the winner of a competition to improve on Netflix's own method of predicting its customers' ratings of movies. Netflix rent out movies on DVD and they have over 6 million customers. Their customers are allowed to rate any movie in the Netflix database, whether they have rented it or not, using a rating scale of one to five stars. Netflix want to predict what rating a customer would give a movie that the customer has yet to rent or rate. This is so that Netflix can recommend to each customer movies that the customer is expected to rate highly and to avoid recommending movies to which the customer would give a low rating.

Superficially, the Netflix Prize appears trivial. In a standard Rasch analysis, the customers are the "persons" and the movies are the "items." The Rasch-Andrich measures of customers, movies and the rating scale structure supports the prediction of the rating of any movie by any customer. So, analyze the database of ratings with an Andrich model, and then recommend the most popular movies to everyone. The prediction would be 5 stars! This works successfully, but soon everyone has seen those most popular movies, such as "*Miss Congeniality.*" In order to keep their business active, Netflix need to keep recommending further movies to their customers. For these recommendations, the personal likes and dislikes of each customer need to be taken into account.

The Prize competition centers on the submission to Netflix of a set of predicted ratings for which Netflix provide the customer and the movie identification but not the ratings themselves. The basis for those predictions is found in a dataset of 100,480,507 "Training" ratings by 480,189 customers of 17,770 movies. Though the number of ratings is huge, it comprises only a little over 1% of the possible ratings. The data matrix is 99% missing. The 2,817,131 "Qualifying" ratings to be predicted form part of these missing data. These ratings have already been made by customers but are kept secret by Netflix. Participants in the Prize

competition submit datasets of their predictions of those secret ratings. About half of the secret Submission ratings form the "Quiz" subset. For this half, Netflix publish a summary statistic of the success of each participant's set of predictions. The other half of the secret ratings form the "Test" subset. For these Netflix provide no feedback at all. It is the accuracy of prediction of the Test subset that will decide the Prize winner. Participants in the competition may make multiple submissions.

The published summary statistic is the root-mean-square-error, RMSE, of the predicted values against the secret observed ratings in the Quiz subset. The smaller the RMSE, the better. On the 5-category rating scale, 1-5, the mean of the Quiz ratings is 3.67, and the standard deviation of the Quiz ratings around that mean is 1.1287. This would be the RMSE if the every prediction in the submitted dataset of predictions were to be the mean of the Quiz ratings.

The huge number of ratings in the Netflix dataset and the high proportion of missing data present no impediment to Rasch analysis in principle, but the operational aspects are taxing. A direct Andrich analysis of the data, followed by the generation of predictions, produces a Quiz RMSE of 0.9823. This looks good until it is compared with Netflix's own prediction RMSE of 0.9514. To win the Prize, the submitted prediction must have an RMSE of 0.8563 or less.

We might predict that each movie has its own partial credit scale. Perhaps there are some movies customers either love or hate, but other movies that evoke little extreme reaction. Applying a Partial Credit model to the movies allows each of the 17,770 movies to define its own 5-star rating scale structure. The result RMSE is 0.9867, worse than the Andrich-model prediction of 0.9823. This is surprising. Adding an extra 3 x 17770 = 53,310 parameters to the model has made the predictions worse!

Or we might predict that each of the 480,189 customers has a unique person-style rating scale. Perhaps there are some customers that either love or hate movies, but other customers who are more middle-of-the road. Applying a person-style partial credit model increases the number of estimated parameters from approximately 500,000 to 2,000,000. The resulting RMSE is 0.9907, worse yet! It appears that partial credit models, both for movies and for customers, overfit the Netflix dataset and make the prediction of future ratings worse. Modeling the rating scale structure to vary across customers or movies does not explain the variance in the Submission ratings. The data are multidimensional.

The Challenge of Empirical Dimensionality

Dirks (2008) suggests that there are 11 main genres of movie: Action, Adventure, Comedy, Crime/Gangster, Drama, Epics/Historical, Horror, Musicals, Science Fiction, War, Westerns. Informal research reported by a Netflix contestant, however, suggests that a strictly-genre based approach is not likely to be successful. "... genre, cast, director, etc are much more useful for predicting what someone will rent rather than what someone's rating for a particular movie will be" (bbame, 2007).

Accordingly, a major aspect of the Netflix Prize is devising and implementing methods which successfully extract empirical dimensions from the Training dataset, and then use these dimensions to predict the responses in the Qualifying dataset.

One approach to identifying the multidimensional structure within ordinal data is principal components analysis of residuals (Linacre, 1998), but this is ineffective here. The data are too sparse, and the data matrix is too large for the standard decomposition algorithms. The estimation errors overwhelm the sought-for components.

What Does Work in Predicting Future Ratings

One productive area of exploration is called "Collaborative Filtering." Researchers have exerted considerable effort in developing techniques which identify similar patterns within the data, and then using these patterns to predict the values of missing or future data points.

When attempting to predict the rating by a customer of a movie, an obvious starting point is to identify another customer who has already rated the target movie and whose ratings approximately match the ratings of the target customer on movies they both rated. The rating of the matching customer is then the predicted rating of the target customer. This can be extended to a set of K most closely-matching customers, the customer's K nearest neighbors, KNN, whose averaged rating on the target rating becomes the prediction. Or similarly the prediction can be based on finding the K nearest neighbors among the movies to the target movie, and using an average of the ratings by the target customer of those movies. In an experimental analysis, using a KNN approach yielded an RMSE of 0.9721, which is better than the Andrich-model RMSE of 0.9823. The sparseness of the data also weakens this approach because the number of close neighbors to many target customers or to many target movies is severely limited.

A more effective descriptive model is based on classical test theory. It is Singular Value Decomposition, SVD, which has been known for over a century. Here is an SVD model,

$$X_{ni} = x_n + x_i + \sum_{a=1}^{A} y_{na} y_{ia}. \qquad (11)$$

X_{ni} is the raw observation. It is initially decomposed into a customer component x_n and a movie component x_i. After these are estimated and fixed, the decomposition continues through A aspects (features, epochs) with the values fixed after each aspect is estimated prior to the next aspect being estimated. The aspect values are multiplicative. y_{na} is the contribution of person n in aspect a, and y_{ia} is the contribution of movie i in aspect a. Each aspect can be conceptualized as a dimension, and the substantive meanings of the first few aspects have been identified by participants in the Netflix Prize competition. According to the Netflix website, RMSEs of the order of 0.9132 have been obtained with the SVD approach. And there are indications that even lower values have been achieved by some participants.

The SVD approach is clearly productive, and it suggests a Rasch-SVD model of the form:

$$\log\left(P_{nij} / P_{ni(j-1)}\right) = B_n - D_i - F_j + \sum_{a=1}^{A} Y_{na} Y_{ia}, \qquad (12)$$

where Y_{na} and Y_{ia} are the contributions to the measures for aspect a by person n and movie i. An experiment with this model yielded an RMSE of 0.9579, better than the simple Andrich model of 0.9823 but still not as good as Netflix's own RMSE of 0.9514. For comparison, the leading competitor's RMSE in the competition as of June 2007 is 0.8808, well on toward the winning target of 0.8563.

What Might Work

A more elaborate approach to prediction is offered by Boltzmann machines, which some researchers are applying to the Netflix Prize data. Boltzmann machines (Ackley et al., 1985) are algorithms that make stochastic decisions based on data. They are composed of units, some of which are "visible," i.e., direct from the data, and others are "hidden," i.e., latent variables which must be inferred from the observed data. These machines were devised to model the learning process in neural networks.

A Boltzmann machine resembles a set of dynamic Rasch models. When unit u is given the opportunity to update its binary state of active $S_u = 1$ or of inactive $S_u = 0$, it first computes its total input, I_u, which is the sum of its own "bias" measure, U_u, and of the contribution, W_{uv}, coming from each of the other units, v, depending on that unit's binary state, S_v, so that

$$I_u = U_u + \sum_{v \neq u} S_v W_{uv}, \qquad (13)$$

Unit u then becomes active with a probability given by the logistic function:

$$\text{Prob}\left(S_u = 1\right) = \exp\left(I_u\right) / \left(1 + \exp\left(I_u\right)\right). \qquad 14$$

If the units are updated sequentially, the system will eventually stabilize into a Maxwell-Boltzmann distribution, hence the name "Boltzmann Machine."

Restricted Boltzmann Machines (Smolensky, 1986) consist of a layer of visible units and a layer of hidden units with no visible-visible or hidden-hidden connections. Consequently predictions from all the hidden units can be made in parallel. These restricted machines suggest the formulation of Boltzmann-Rasch Machines (BRM).

Let us start from the Rasch model for dichotomous data which models the item difficulty, D_i, of item *i,* where $i = 1, L$, when administered to a person *n* of ability B_n, where $n = 1, N$, producing an observed set of dichotomous responses $\{X_{ni}\}$. Then, the dichotomous Rasch model is

$$\log(P_{ni1} / P_{ni0}) = B_n - D_i , \qquad (15)$$

where P_{ni1} is the probability that $X_{ni} = 1$. The abilities and difficulties are estimated in the usual way.

Now let us add a hidden Boltzmann unit, termed a feature, to the Rasch model to construct a BRM. This feature models an off-dimension interaction between the persons and the items. W_i is the weight of the feature for item *i.* W_n is the weight of the feature for person *n*. S_{ni} is a binary 0-1 switch indicating whether the feature is active for person *n* and movie *i.*

$$\log(P_{ni1} / P_{ni0}) = B_n - D_i + S_{ni}(W_i + W_n) \quad 16$$

with

$$\text{Prob}(S_{ni} = 1) = \exp(I_{ni}) / (1 + \exp(I_{ni})) \quad 17$$

When $S_{ni} = 0$, the expected value of the rating, E_{uni}, is

$$E_{uni} = 1 / (1 + \exp(-(B_n - D_i)) \quad 18$$

But when $S_{ni} = 1$, the expected value of the rating, E_{wni}, is

$$P_{wni1} = 1 / (1 + \exp(-(B_n - D_i + W_i + W_n)) \quad 19$$

The setting of the binary switch, S_{ni}, for person *n* on movie *i* is given by the logistic model,

$$\log(S_{ni1} / S_{ni0}) = U + \sum_{j=1}^{L} W_j (X_{nj} - E_{unj}) + \sum_{k=1}^{N} W_k (X_{ki} - E_{uki}), \quad 20$$

where S_{ni1} is the probability that $S_{ni} = 1$, U is the "bias" measure of the feature, and the weight W_j only contributes to the switch setting for person *n* to the extent that X_{nj} departs from its baseline expectation E_{unj}, and similarly for W_k. Thus the combined stochastic BRM model of the expectation of E_{ni} of the predicted Boltzmann-Rasch rating becomes:

$$E_{ni} = S_{ni0} E_{uni} + S_{ni1} E_{wni} \quad 21$$

This model has attractive features because the $\{W_i\}$ and $\{W_n\}$ can be interpreted as logit distances in another dimension which interacts probabilistically with the main Rasch dimension. U quantifies the influence of the dimensional aspect on the ratings, and W_i and W_n quantify the dimensional aspect for movie *i* and customer *n.*

The Rasch-Boltzmann model can be extended to polytomous data by rewriting (16):

$$\log(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_j + S_{ni}(W_i + W_n), \quad 21$$

As before, when $S_{ni} = 0$, the expected value based on (21) becomes E_{uni}, and when $S_{ni} = 1$, the expected value becomes E_{wni}, and equations (19) and (20) hold. Initial experiments with this model indicate that it is difficult to estimate the parameter values. However, estimation of the parameters of other Boltzmann Machines has been successfully accomplished using Gibbs Sampling, so that is a promising approach to apply here.

Conclusion

It is clear that the prediction of future ratings requires effort beyond merely finding the best model that fits the current dataset and estimating its parameters. In fact, what conventional statistics might regard as the best-fitting descriptive model may be a relatively poor predictive model.

Rasch theory provides a good foundation for predictive models because its intention is to produce parameter estimates as independent as possible of the idiosyncrasies in the current dataset. Nevertheless, the necessity of including additional structural dimensions into the measurement framework suggests that extensions to the standard unidimensional Rasch models are required for effective prediction of future ratings.

References

Ackley, D., Hinton, G., and Sejnowski, T. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1),147-169.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaksi, (Eds.). 2nd International Symposium on Information Theory (pp. 267-281). Budapest, Hungary: Akademiai Kiado, .

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69-81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357-74.

Bbame (pseudonym). (2007, March 14). *Movie metadata*. Retrieved on January 19, 2008, from http://www.netflixprize.com

Beltrami, E. (1873). Sulle funzioni bilineari. *Giornale Math*. *Battaglini*, *11*, 98-106.

Dirks, T. (2008). Film genres. Retrieved January 19, 2008, from http://www.filmsite.org/genres.html

Glas, C. A. W., and Verhelst, N. D. (1991). Using the Rasch model for dichotomous data for analyzing polytomous responses. *Measurement and Research Dept Report 91-3*. Arnhem, The Netherlands: CITO.

Linacre, J. M. (1991). *Beyond Partial Credit, Rasch Measurement Transactions*, *5*(2), 155

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266-283.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Rasch, G. (1961). On general laws and meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, (4), 321-333. Berkeley, CA: University of California Press.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. van Duijn, and T. Snijders (Eds.), *Essays on item response theory*. Berlin/Heidelberg/New York: Springer.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland (Eds.). *Parallel distributed processing: Foundations* (Vol. 1, pp. 194-281). Cambridge, MA: MIT Press.

Verhelst, N. D., Glas, C. A. W., and De Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123-138) New York: Springer.

Wright B. D., and Masters G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.