EDUCATIONAL DATA SYSTEMS, INC

OBJECTIVITY AND MULTIDIMENSIONALITY

AN ALTERNATING LEAST SQUARES ALGORITHM FOR IMPOSING RASCH-LIKE STANDARDS OF OBJECTIVITY ON HIGHLY MULTIDIMENSIONAL DATASETS

Mark H. Moulton, Ph.D.
markhmoulton@gmail.com

May 14, 2013

ABSTRACT

To an increasing degree, psychometric applications (e.g., predicting music preferences) are characterized by highly multidimensional, incomplete datasets. While the data mining and machine learning fields offer effective algorithms for such data, few specify Rasch-like conditions of objectivity. On the other hand, while Rasch models specify conditions of objectivity—made necessary by the imperative of fairness in educational testing—they do not decisively extend those conditions to multidimensional spaces. This paper asks the following questions: What must a multidimensional psychometric model do in order to be classified as “objective” in Rasch’s sense? What algorithm can meet these requirements? The paper describes a form of “alternating least squares” matrix decomposition (NOUS) that meets these requirements to a large degree. It shows that when certain well-defined empirical criteria are met, such as fit to the model, ability to predict “pseudo-missing” cells, and structural invariance, NOUS person and item parameters and their associated predictions can be assumed to be invariant and sample-free with all the benefits this implies. The paper also describes those conditions under which the model can be expected to fail. Demonstrations of NOUS mathematical properties are performed using an open-source implementation of NOUS called Damon.

BLANK

1.0 BACKGROUND

1.1 SPECIFIC OBJECTIVITY

A student takes a test with items 1 - 20. Another student takes a test on the same subject with items 10 - 30. Though 10 items are overlapping, the tests are different. The problem of psychometrics -- perhaps most clearly identified by Danish mathematician Georg Rasch in 1960 -- is to answer the question, "How can we compare two students who take different tests as if they had taken the *same test*?" The answer Rasch proposed is to impose a model and scoring procedure that generates scores for which it can be demonstrated that each score does not depend on the sample of items a student takes. Any sample of items should yield the same score. This sample-independence property he called "specific objectivity", and it is the defining characteristic of the Rasch model. A test that meets the conditions of this model can be called "fair". The Rasch model is typically formulated as:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}, \qquad \text{Eq. 1}$$

where *Xni* is a dichotomous (0,1) observation, *ßn* is the ability of person *n* and *δi* is the difficulty of item *i*. Note that the probability of success is a function solely of a person (row) parameter and an item (column) parameter, with no interaction terms. Rephrased as a deterministic model in matrix notation and recasting item difficulty as item easiness, the following is conceptually equivalent:

$$\mathbf{E}_{ni} = \mathbf{R}_n \cdot \mathbf{C}_i \qquad \text{Eq. 2}$$

where each element $\mathbf{E}_{ni}$ of estimates matrix **E** is the dot product of vector *n* of row (person) matrix **R** and vector *i* of column (item) matrix **C** under the constraints that $0.0 < \mathbf{E}_{ni}, \mathbf{R}_n, \mathbf{C}_i < 1.0$ and vectors *n* and *i* contain only one element, i.e., the dimensionality is 1.

Over the next 30 years or so, the Rasch model gained acceptance in the fields of education and licensure as the strictest of a family of Item Response Theory models in its adherence to the specific objectivity requirement. The model has been used widely in those fields for several reasons:

1. Fairness. It is a strong requirement in high stakes testing that all examinees be comparable solely in terms of the construct of interest and that scores differ for no other reason than ability on this construct. The Rasch model provides defensible statistical criteria for making the claim that fairness has been achieved.

2. Unidimensionality. The Rasch model requires unidimensionality (all items testing for the same kind of ability) in order to achieve its special objectivity property. As it happens, the datasets generated in the field of educational are generally, though not necessarily, reducible to a single ability dimension. This is in part due to deliberate test design, in part due to the widespread use of multiple-choice questions which, taken one by one, tend to be so indeterminate in the dimension they imply that they yield a single (somewhat messy) composite dimension when aggregated.

3. Missing Data Designs. Missing data is a characteristic of most real-world datasets but it is especially important in the field of education. This is because it is often desired to compare students from different grades or ability levels. Because students at different ability levels need to be given tests appropriate to their level (though containing items that "link" to other levels), the aggregate dataset of all students will contain large blocks of non-randomly missing cells. Because Rasch student scores do not depend on which items a student takes, these missing blocks do not compromise the ability to compare students from different ability levels.

The chief limitation of the Rasch model is that it derives its properties by restricting itself to one dimension in a given analysis. Fields such as spectral analysis, latent semantic analysis, image recognition, genetics, consumer preference prediction, market forecasting, artificial intelligence, and so on, with datasets that are often highly and intrinsically multidimensional, are beyond the reach of unidimensional models.

1.2 SINGULAR VALUE DECOMPOSITION

Discovered independently several times since the 1800's, Singular Value Decomposition (SVD) has been widely used since practical methods for implementing it were developed in the 1950's and 1960's. SVD decomposes a data matrix, potentially of a high number of dimensions, by modeling it with the *mxn* matrix **M** defined as the product of three matrices:

$$\mathbf{M} = \mathbf{U\Sigma V^*} \qquad \text{Eq. 3}$$

where **U** is an *mxm* real or complex unitary matrix, **Σ** is an *mxn* rectangular diagonal matrix with nonnegative real numbers on the diagonal, and **V*** (the conjugate transpose of **V**) is an *nxn* real or complex unitary matrix (Wikipedia, *Singular value decomposition*, May 14, 2013). Matrix **Σ**, the numbers on the diagonal of which are called "singular values", controls the rank or dimensionality of the SVD representation of **M**. The number of singular values can be no larger than *m* (the number or rows/columns in **U** and the size of **M** along its smallest axis). When all *m* values are greater than zero, the resulting matrix **M** exactly reproduces the data matrix the SVD is applied to. The dimensionality is reduced by setting some of the singular values to zero. An SVD where **Σ** consists of three non-zero singular values produces a matrix **M** which is a 3-dimensional representation of the data.

Once the rank (dimensionality *d*) of Equation 3 has been decided, Equation 3 can be restated as the outer product of an *mxd* row matrix **R** and a *dxn* column matrix **C**. Changing **M** to **E** to make notation consistent the value of each element $\mathbf{E}_{ni}$ is:

$$\mathbf{E}_{ni} = \mathbf{R}_n \bullet \mathbf{C}_i \qquad \text{Eq. 4}$$

which is the same formally as Equation 2, the matrix restatement of the Rasch model, except that $-\infty < \mathbf{E}_{ni}, \mathbf{R}_n, \mathbf{C}_i < \infty$ and vectors *n* and *i* contain *d* elements. Thus, the Rasch model can be thought of as a special case of SVD in which the vectors are required to be positive (expressible as probabilities) and the dimensionality is 1.

Despite their formal similarities, the algorithms for implementing the Rasch model and SVD are quite different, with important practical implications:

1. Missing Data. SVD, applied to the data matrix as a whole using the techniques of linear algebra, requires complete data. Rasch algorithms such as JMLE (Joint Maximum Likelihood Estimation) iteratively solve for each row and column individually and can ignore missing cells.

2. Probabilities. SVD assumes interval data and computes row and column parameters accordingly. Rasch assumes ordinal data and computes probabilities for its row and column parameters and for each "step" between adjacent categories.

3. Convergence Goal. SVD minimizes the Frobenius norm (the sum of squared residuals) between observations matrix **X** and estimates matrix **E**. Rasch finds parameters that maximize the likelihood of **X**.

In addition, the goals of the two methodologies are quite different. The goal of Rasch is to create a generalizable measurement structure and to filter out all data that might potentially or actually get in the way. It does this by forcing analysts to collect and edit data in such a way that it can be described *sufficiently* by locations along a single dimension. The goal of SVD is to describe a dataset concisely as the product of two matrices. While singular values are often examined with an eye toward reducing the rank (dimensionality) used to describe the data, the criteria for choosing a dimensionality are not always clear.

The big advantage that SVD has over Rasch is that it can model highly multidimensional datasets that are intractable with 1-dimensional models.

1.3 ALTERNATING LEAST SQUARES

The Alternating Least Squares (ALS) family of algorithms offers an alternative to SVD for calculating **R** and **C**. It is simple, efficient, converges, is robust to variation in data types, handles constraints gracefully (e.g, that **R** and **C** be nonnegative), generalizes to higher order tensors (analogous to "facets" in the Rasch literature (Linacre, 1994), and most important, is able to handle missing data. ALS has arisen in multiple fields, often independently, since at least the 1970's. Important early contributors in the field of psychometrics in the 1960's and '70's are Tucker, Carroll and Chang, Kruskal, Harshman, and Kroonenberg and de Leeuw. Appelloff and Davidson introduced similar ideas in chemometrics (1981). Lee and Seung were seminal in developing nonnegative ALS (1999). Kolda and Bader provide a very useful survey of the work done generalizing ALS to higher order tensor ("many-facet") decompositions (2008). The last two decades have seen a flowering of ALS-related algorithms, and they have played a leading role in such data-mining competitions as the Netflix Prize (2009) and the Yahoo! KDD-CUP (2011). Hu, Koren, and Volinsky (2008) describe how Alternating Least Squares can be used to develop recommender systems (e.g., movies) and how it is well suited for massively parallel distributed processing.

Parallel to the development of multidimensional ALS decomposition algorithms, but quite distinct from it, the field of Multidimensional Item Response Theory (MIRT) has evolved a family of probability models for handling multidimensional test data. Mark Wilson and Mark Reckase are seminal figures in this field, though their models are quite distinct. Reckase has published a useful history and explanation of MIRT (2009). Wilson, Adams, and Wang (1997) published an important paper describing the multidimensional random coefficients multinomial logit model.

The form of ALS discussed in this paper, called NOUS, was independently invented by Howard Silsdorf in 2001 in collaboration with the author, who subsequently elaborated it. It is implemented in the open-source Python package *Damon*. In many ways it is a rediscovery of previous forms of ALS. It differs primarily in the philosophical emphasis that it places on finding the optimal rank, or dimensionality, of **R** and **C**, and in the procedures used to do so.

1.4 SPECIFICATIONS OF AN "OBJECT-ORIENTED" MULTIDIMENSIONAL MODEL

It is beyond the scope of this paper to attempt a comparison of the various models and algorithms referred to here. Instead, we focus on the properties one would *hope* to find in an object-oriented multidimensional model. The term "object-oriented" is used in preference to "objective" to indicate that objectivity is a goal of analysis, an analytical *orientation*, not a property of algorithms *per se*. It also makes the important epistemological point that the goal of such a model is to abstract entities (persons, items) from the datasets in which they appear, to view them as "objects" separable from their context.

An object-oriented multidimensional model should have the following properties:

1. Fit multidimensional data. It should produce cell estimates that fit observed values within statistical tolerances, even when the items are not all measuring along the same dimension or construct and are to varying degrees uncorrelated or negatively correlated, i.e., the dataset is multidimensional.

2. Objective Dimensionality. It should be possible to determine unambiguously the number of dimensions of the common space within which the items and persons are located. In other words, persons and items should not be the only objects under consideration. The space itself should be interpretable as an object with invariant attributes, and the most important of these attributes is its dimensionality.

3. "True" estimates. Each cell estimate should be as close to the "true" value for that cell as possible given the data available. It should approximate the value that would be obtained by independently repeating the observation for that cell an infinite number of times and averaging the observations.

4. Predict missing values. It should accurately predict values of missing cells, regardless of whether they are randomly missing (the examinee did not respond) or missing by design (as in test equating). It should also handled sparse datasets, where more than 90% of the data are missing.

5. Person invariance. It should locate persons in space such that their positions remain the same regardless of the sample of items they answer, assuming that each sample of items erects the same space.

6. Item invariance. It should locate items in the same space as the persons, and the positions of the items should remain the same regardless of the sample of persons that responds to the items.

7. Misfit when invariance (objectivity) is not achieved. When a person's response to an item varies due to variation in dimensions that are not part of the space erected by the remaining items on the test, that response should be significantly different from the value predicted for it. In other words, the model's predictions and the observed values should disagree. Put another way, model estimates should not *overfit* the data.

8. Minimal influence by observations. The model should make it possible to minimize the influence of any given observation on its corresponding cell estimate.

9. Maximal use of information. The model estimate for each cell should use all relevant information in the dataset and be correspondingly precise.

10. Portability. Person and item parameters should be portable across datasets of the same kind, so that results of one analysis can be applied to another.

11. Sufficiency. Item parameters should on their own be sufficient to summarize the data corresponding to each item without recourse to person parameters, and vice versa.

12. Standard errors. Each cell estimate—how a person is expected to score on an item—should be accompanied by its own variance statistic and standard error. The variance statistic should represent the expected residual between the estimate and the observation. The standard error statistic should represent the expected residual between the estimate and the "true" value.

13. Small, unrepresentative samples. The model should not depend for its properties on the statistical benefits of large or representative samples.

This paper offers evidence that NOUS, suitably applied, leads to results that enjoy these properties to a large degree.

2.0 THE NOUS ALGORITHM

2.1 CALCULATION OF **R**, **C**

Consider an *nxi* matrix **X** of data values. For ease of visualization, think of the rows as persons, the columns as items, each data value as a crude measure of how a person performed on an item. Let us assume for now that the cell values are interval measures, that the data matrix may be multidimensional, and that items may be more or less correlated, or negatively correlated, with each other.

The user specifies a range of dimensionalities $d = \{D_1, D_2, ..., D_k\}$. For the first dimensionality (order does not matter) NOUS initializes an *nxd* matrix **R** to contain person coordinates and a *dxi* matrix **C** to contain item coordinates. (The row and column vectors $\mathbf{R}_n$ and $\mathbf{C}_i$ are called coordinates, as they are spatial coordinates mapping a space of *d* dimensions.) **C** is populated with a set of random numbers called a "seed". **C** is used in conjunction with the data in the first row to compute a set of coordinates for that row, the first row in **R**, using ordinary least squares. The process is repeated for the second row, and so on, until **R** is populated with values. Then **R** is used in conjunction with the data in the first column to compute a set of improved coordinates for the first column in **C** using ordinary least squares. The process is repeated for each column in **C**. **C** is used to recompute **R** and **R** used to recompute **C**, improving them iteratively until a stopping condition is met. The alternating calculation of least squares solutions is why the algorithm is called "alternating least squares."

R and **C** are called tensors, or facets. There is no limit to the number of facets that can be used to model the data. A 3-facet example might be raters evaluating how persons perform on items—each data value is modeled as the product of a person vector, an item vector, and a rater vector. However, this paper limits itself to the 2-facet case.

Thus, given the system of equation **Uv** = **x**, where **U** is an array of row or column coordinates, **x** is the data corresponding to a specified row or column, and **v** is a vector of coordinates for that row or column, a least squares solution to the system denoted by **v**[*solution*] will also be a solution to the associated normal system,

$$\mathbf{U}^{\mathbf{T}}\mathbf{U}\mathbf{v} = \mathbf{U}^{\mathbf{T}}\mathbf{x} \qquad \text{Eq. 5}$$

If **U** has linearly independent rows and the system is over-conditioned (there are more observations than unknowns (dimensions) in **v**, then a unique least squares solution is given by,

$$\mathbf{v}[solution] = (\mathbf{U}^{\mathbf{T}}\mathbf{U})^{-1}\ \mathbf{U}^{\mathbf{T}}\mathbf{x} \qquad \text{Eq. 6}$$

so that the least squares solution for each row and column is

$$\mathbf{v}^{\mathbf{R}}[solution] = (\mathbf{U}^{\mathbf{RT}}\mathbf{U}^{\mathbf{R}})^{-1}\ \mathbf{U}^{\mathbf{RT}}\mathbf{x}^{\mathbf{R}} \qquad \text{Eq. 7}$$

$$\mathbf{v}^{\mathbf{C}}[solution] = (\mathbf{U}^{\mathbf{CT}}\mathbf{U}^{\mathbf{C}})^{-1}\ \mathbf{U}^{\mathbf{CT}}\mathbf{x}^{\mathbf{C}} \qquad \text{Eq. 8}$$

If any cells are missing in **x** for a given row or column, the corresponding vectors in **U** are ignored when calculating **v**.

So far, this describes a fairly conventional interpretation of ALS. Where NOUS differs is that for a given dimensionality *d* it computes three statistics to assess the "objectivity" of the solution. First, it computes estimates **E**[*pseudo-missing*] for a random selection of cells that have been made "pseudo-missing" prior to the analysis. The product-moment correlation of these estimates and their corresponding original data values **X**[*pseudo-missing*] is defined to be the Accuracy statistic:

$$\mathbf{E}[\textit{pseudo-missing}] = \mathbf{R}[\textit{pseudo-missing}] \bullet \mathbf{C}[\textit{pseudo-missing}] \qquad \text{Eq. 9}$$

$$\text{Accuracy} \equiv \textit{correl}(\mathbf{X}[\textit{pseudo-missing}], \mathbf{E}[\textit{pseudo-missing}]) \qquad \text{Eq. 10}$$

NOUS also calculates a "stability" statistic for dimensionality *d*. This is obtained by applying ALS to separate quadrants of **X**, as follows

1. Column coordinates. Calculate column coordinates using only *half* the rows ("the first half').

2. Group 1 row coordinates. Divide the columns into two groups, Group 1 and Group2. Then using only data from the "second half' of rows (not used to compute column coordinates), calculate a set of row coordinates, called the Group1 row coordinates.

3. Group 2 row coordinates. Repeat Step 2 with the Group 2 columns, resulting in Group 2 row coordinates.

4. Correlate Group 1 and Group 2 coordinates. If the correlation between the Group 1 and Group 2 coordinates is 1.0, Rasch's person invariance requirement (Property 3 above) has been exactly met for at least the Group 1 and Group 2 samples. The person coordinates are the same regardless of which sample of items is used to calculate them.

Thus,

$$\text{Stability} \equiv \textit{correl}(\mathbf{R}[\textit{Group 1}], \mathbf{R}[\textit{Group 2}]) \qquad \text{Eq. 11}$$

NOUS defines a third statistic, "objectivity", to be the geometric mean of Accuracy and Stability:

$$\text{Objectivity} \equiv (\text{Accuracy} * \text{Stability})^{(1/2)} \qquad \text{Eq. 12}$$

ALS is applied at each dimensionality *d* in the specified range and objectivity statistics are calculated. The dimensionality with the highest objectivity *D*[*objective*] is selected as "final" and defined to be the "objective dimensionality" of the dataset. ALS is applied again at *D*[*objective*], with the pseudo-missing cells restored, to compute **R** and **C**.

In addition to being used to assess each dimensionality, Objectivity is used to assess "seed" random starter coordinates. When the data is free of noise, it can be shown that all seeds will lead to the same solution **E** (though a different **R** and **C**). However, as noise is introduced the choice of starter coordinates makes a difference. NOUS computes objectivity statistics for each of a specified sample of seeds and selects the seed that leads to the most objective solution.

The above procedure suffices as an overview of the NOUS algorithm, but it is far from complete. For instance, at each iteration the row coordinates matrix **R** is (optionally) converted into its orthonormal equivalent using **QR** decomposition. This has a number of important benefits. It improves the numerical stability of the algorithm by avoiding ill-conditioned matrices. It improves the interpretability and usability of the column coordinates by making it possible to use column coordinates to estimate the variance of the column estimates, as well as to estimate the correlation between any two columns (equivalent to the cosine of the angle between those columns) without having actually to compute those estimates. To the degree that **R** and **C** are objective (Objectivity = 1.0), these statistics will enjoy comparability for every subsample of persons, regardless of data errors or missing cells, and the cosine/correlation statistic between two columns will be the same for each subsample of rows.

In addition, NOUS (optionally) down-weights the influence of very large coordinates on downstream solutions. This avoids “influence traps” caused by a few coordinates inadvertently dominating downstream calculations due merely to the accident of choice of starter coordinates.

2.2 MISSING DATA

It is important to note that the above procedure is robust to missing data. This is due to the piece-meal nature of ALS -- for any given row or column, only those coordinates are used for which there exists observed data. All other cells are ignored. For example, if the first cell in a row with 10 observations is missing, the least squares solution for the row is computed using only observations 2 - 10 and only the corresponding column coordinates **C**[2 - 10]. The solution is valid to the degree that the system of equations solved by least squares is *over-determined*, i.e, there are more equations than unknowns. This is a well-known property of solving simultaneous equations and is the basis of Gaussian least squares. That means in order to calculate a least squares solution for a given row, the number of observations must exceed the dimensionality under consideration. Otherwise, an error is returned for that row.

This approach to missing data has several important implications:

1. It does not matter how *many* cells are missing in a given row or column, so long as there are at least as many observations as the specified dimensionality (the more the better, obviously).

2. If the model fits the data at the objective dimensionality, it does not matter *which* cells are missing in a given row or column. The remaining observations combined with the corresponding coordinates will lead to (approximately) the same least squares solution.

3. There is no need to impute values for missing cells in order to perform an analysis. They are ignored anyway.

4. NOUS can be applied to sparse data matrices, with high percentages of missing cells.

5. If the model fits the data at the optimal dimensionality, NOUS estimates for missing cells are not merely statistically plausible values. They are definite predictions that can be expected to approximate the “true” value.

2.3 ESTIMATES, VARIANCE, FIT, STANDARD ERROR, SEPARATION, RELIABILITY

Having calculated row coordinates **R** and column coordinates **C** at the objective dimensionality *D*[*objective*], it is straightforward to compute statistics corresponding to each individual cell of the complete data matrix. The primary statistic, the cell estimate for each cell, is given by:

$$\mathbf{E}_{ni} = \mathbf{R}_n \bullet \mathbf{C}_i \qquad \text{Eq. 13}$$

These estimates exist for every cell, including those that are missing. They can be interpreted as the most likely value of the cell given the rest of the data in the array. They can also be interpreted as an estimate of the “true” value of the cell, which is defined as the value that would be obtained if an infinite number of independent observations for that cell were averaged. Yet again **E** can be interpreted as the orthogonal projection of **X** into the *D*[*objective*] subspace. The accuracy of the estimate is governed by the total number of observations in the dataset and the objectivity of the system.

Each cell with an observation has a residual:

$$\mathbf{Res}_{ni} = \mathbf{X}_{ni} - \mathbf{E}_{ni} \qquad \text{Eq. 14}$$

NOUS defines the variance of each cell in terms of the absolute residual that is expected for it. This statistic is called the Expected Absolute Residual (EAR). It is calculated by applying NOUS to the matrix of absolute residuals, specifying a dimensionality of 1 ($d = 1$). Absolute residuals are preferred to squared residuals in this context because they are easier to analyze with NOUS.

$$\mathbf{EAR}_{ni} = \mathrm{NOUS}(\mathbf{Res}_{ni}, d = 1) \qquad \text{Eq. 15}$$

Because NOUS can estimate missing cells, an EAR statistic can be calculated for every cell in the data matrix, including missing. It is interpreted as the expected absolute difference between the estimate and the raw observation given the other residuals observed for that cell's row and column. The EAR statistic is used to calculate fit statistics. The misfit for a given cell is given by the ratio of the observed to expected absolute residual:

$$\mathbf{Misfit}_{ni} = \mathbf{Res}_{ni} / \mathbf{EAR}_{ni} \qquad \text{Eq. 16}$$

Cell misfit statistics can be aggregated across the rows and columns in various ways to generate row and column-level fit statistics.

While the EAR statistic is good for estimating the "noisiness" of a given cell, it is not appropriate for significance tests. It estimates the residual between the *observation* and the estimate, not between the *"true" value* and the estimate. For a large dataset, it is quite possible that an estimate is highly accurate and reproducible even as its observed data are very noisy. The statistic that captures the true accuracy of the estimate is its standard error. In NOUS, for a 2-facet system, the standard error is:

$$\mathrm{SE}_{ni} = \frac{2 \cdot \mathrm{E}AR_{ni}}{\left(\left(\frac{N}{D}-1\right)^{\frac{1}{2}}\left(\frac{I}{D}-1\right)^{\frac{1}{2}}\right)^{\frac{1}{2}}} \qquad \text{Eq. 17}$$

where N is the number of observations in the nth row, I is the number of observations in the ith column, and D is the objective dimensionality. A perusal of the formula shows that it is similar in form to the ordinary statistical standard error: SE = SD/sqrt(N). The difference is that it takes into account the number of unanchored facets (2 in this case), the number of observations in the row *and* column, and the dimensionality (equivalent to degrees of freedom). As the number of observations in either the row or column approaches the number of dimensions, the standard error grows to infinity. The formula easily generalizes to more than two facets.

Having cell standard errors makes it possible to perform significance tests with estimates. In addition it makes it possible to compute Separation and Reliability statistics for each row and column.

$$\mathrm{Separation}_i = (\mathrm{SD}_i^2 - \mathrm{RMSE}_i^2)^{1/2} / \mathrm{RMSE}_i \qquad \text{Eq. 18}$$

$$\mathrm{Reliability}_i = \mathrm{Separation}_i^2 / (1 + \mathrm{Separation}_i^2) \qquad \text{Eq. 19}$$

where i refers to the ith column or, alternatively, to the nth row and RMSE refers to the Root Mean Squared Error, an aggregation of the cell standard errors for that row or column. Reliability is the NOUS equivalent of the Cronbach-alpha statistic and, in conjunction with other item statistics, is a useful indicator of item quality.

2.4 DATA REQUIREMENTS

NOUS has definite limits on the types of data it can analyze successfully, i.e., for which it accurately predicts the values of missing cells and derives stable coordinate structures.

1. Structured data. The noisier the data and the higher the measurement error associated with each cell, the less successful NOUS will be. At the extreme of perfectly random data, NOUS will be no more accurate predicting the values of missing cells than simply taking the average of the whole array.

2. Common metric. NOUS requires that all cells be in the same metric whatever that may be. This often requires "pre-standardizing" the data, generally to an interval metric. Binning data values to create a dichotomous array, then converting the 0's and 1's into "pseudo-logits", has also proven to be a useful strategy in some cases, though it invites problems caused by violation of the item independence requirement.

3. Interval data is ideal. Ordinary least squares requires interval-type data, and to that extent NOUS does as well. However, it is often applied successfully to ratio, ordinal, and dichotomous data. When NOUS has trouble with a type of data, it will often find a higher dimensional solution that leads to satisfactory solutions.

4. Gauss-Markov requirements. The Gauss-Markov theorem states that ordinary least squares, the basis of ALS, yields the best linear unbiased estimator (BLUE) when the errors have an expectation of zero, are uncorrelated, and have equal variances (i.e., are homoscedastic). (Wikipedia, *Gauss-Markov Theorem*, April 8, 2012). Ordinal and dichotomous data violate the homoscedasticity requirement, though this has less of an effect on NOUS than one might expect due to the mutability of **R** and **C**. NOUS includes an Iteratively Reweighted Least Squares (IRLS) option to address heteroscedasticity, though this has proven unnecessary.

5. Common space. An exact analog of the Rasch unidimensionality requirement, NOUS requires that either row entities or column entities (generally column entities, items) be sensitive to the same set of dimensions and insensitive to all other dimensions. This is the canonical requirement of NOUS. What this means mathematically is that each item have *non-zero* values on each dimension in the common space and *zero* values on all other dimensions. What this means in practice is that each item should contain the same content dimensions as every other item, albeit in different proportions, the usual example being word problems which require a mix of reading and math ability. This is called "within-item multidimensionality". "Between-item multidimensionality", where items are sensitive to content dimensions not shared by the other items, violate the common space requirement. Random error (noise) is a relatively benign violation of the common space requirement, depending on the extent of noise. The purpose of analysis of fit in the context of NOUS is to identify items that do participate in the common space.

6. Linear structure. All data that can be conceptualized as the dot product of two vectors whose components (dimensions) are linearly independent is in theory analyzable by NOUS. Many, many functions can be formulated in this way, but not all. An important counter-example is distances. The formula for the distance between two points does not fall cleanly into a linear structure, and as a consequence NOUS cannot replicate a distance matrix perfectly (unless special conditions are applied to **R** and **C** in each iteration). Thus, NOUS can be conceived as a test of linear structure.

2.5 SCALABILITY

An important property of alternating least squares is that its piece-meal approach to matrix decomposition opens the door to massively parallel distributed data processing.

1. Analysis by slice. Instead of having to load the entire data array into memory in order to process it, ALS requires only one row or column's worth of data at a time, plus the coordinates that are associated with it. If even that slice of data is too large, it can easily be sampled. This makes it possible to analyze unlimited data with a fixed amount of memory.

2. Ignores missing. Because missing cells are simply ignored, they do not enter into the calculation at all and do not need to be stored. As datasets become very large, they tend to have larger proportions of missing cells, which translates into computational savings for ALS.

3. Parallelism. ALS does not require that its row and column entities be analyzed in any particular order. Therefore, given thread-safe input/output access to the coordinate arrays, it is in theory possible (given enough processors) to compute least squares solutions for all entities at the same time.

4. New data. Once coordinates have been calculated for all entities, they can be stored in a bank and reused when new data is added to the system. That means there is no need to recalibrate the system when new data is introduced. Only the coordinates for the new entity are calculated.

While the Damon package does not as yet exploit the parallelism property, there are other projects that do. An example is the ALS algorithm implemented on the Mahout/Hadoop machine learning platform described by Sean Owen (2012) which is based on a recommender system described by Hu, Koren, and Volinsky (2008). Its scalability suggests that ALS has an important role to play in the analysis of big data.

3.0 PROPERTIES OF NOUS STATISTICS

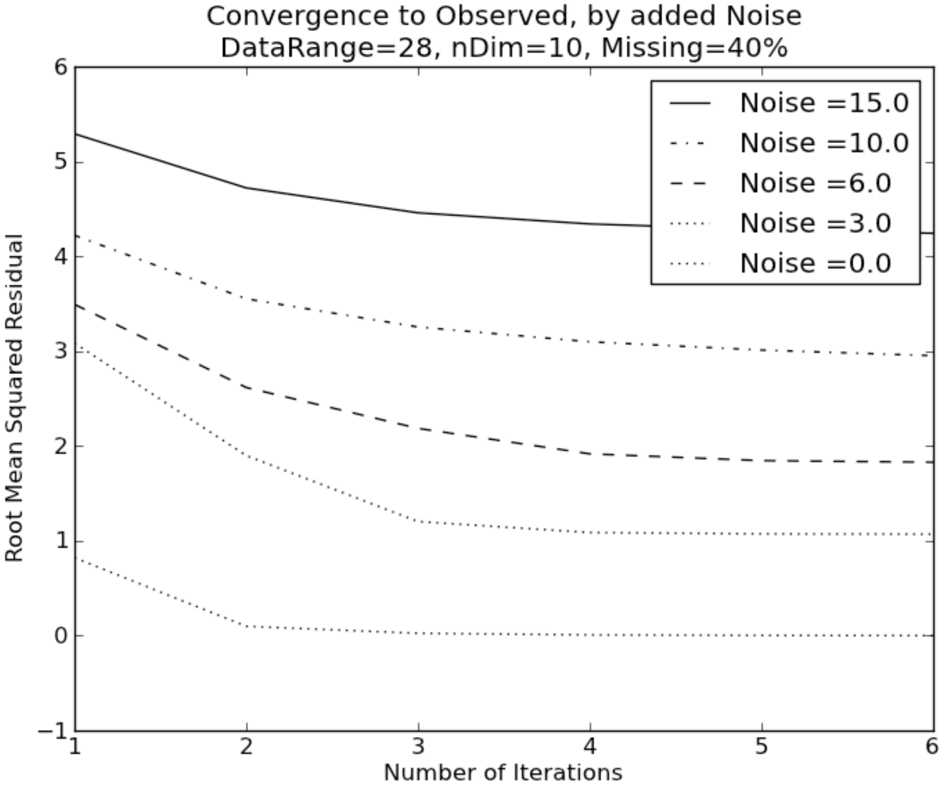
While it is beyond the scope of this paper to provide formal mathematical proofs of claims regarding NOUS, it can offer empirical demonstrations. These demonstrations are intended to show how NOUS succeeds in meeting the specifications laid out for an object-oriented multidimensional model in section 1.4. We begin with the claim that NOUS converges on a solution.

3.1 CONVERGENCE

In ALS, the quantity that is minimized is the Euclidean distance *S* between the model's estimates and the corresponding observed data (where present). *S* is their root mean squared residual. The intuition is that during the iteration between **R** and **C**, each new set of coordinates will yield a closer fit to the observations than the iteration before, inasmuch as it is a least squares solution (the smallest possible distance *S* given the data and the previous **R** or **C**).

Figure 3.1 shows that convergence to a plateau happens fairly rapidly, even at 10 dimensions with high "noise" (random numbers added to the data) and missing data. The data matrix was 100 x 80.

Fig. 3.1

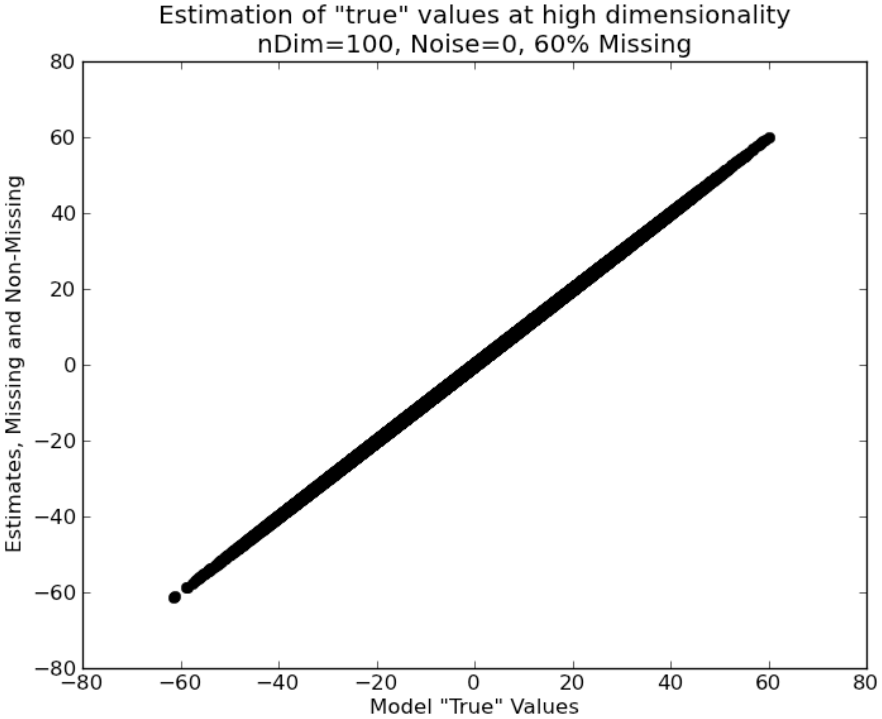


3.2 MULTIDIMENSIONALITY

There is no theoretical limit on the number of dimensions that can be analyzed for a given dataset, only practical limitations posed by the size of the dataset needed to accommodate a large number of dimensions and the computing load involved with high-dimensional solutions.

Figure 3.2 compares estimates calculated from a 1000 x 1000 data array built to be 100-dimensional with model "true" values. 60% of the cells are missing. No noise was added. The chart shows that in the absence of noise NOUS exactly predicts the values of the "true" model for missing *and* non-missing cells. Number of dimensions is not an obstacle to building the correct model for a given dataset.

Fig 3.2



3.3 OBJECTIVE DIMENSIONALITY

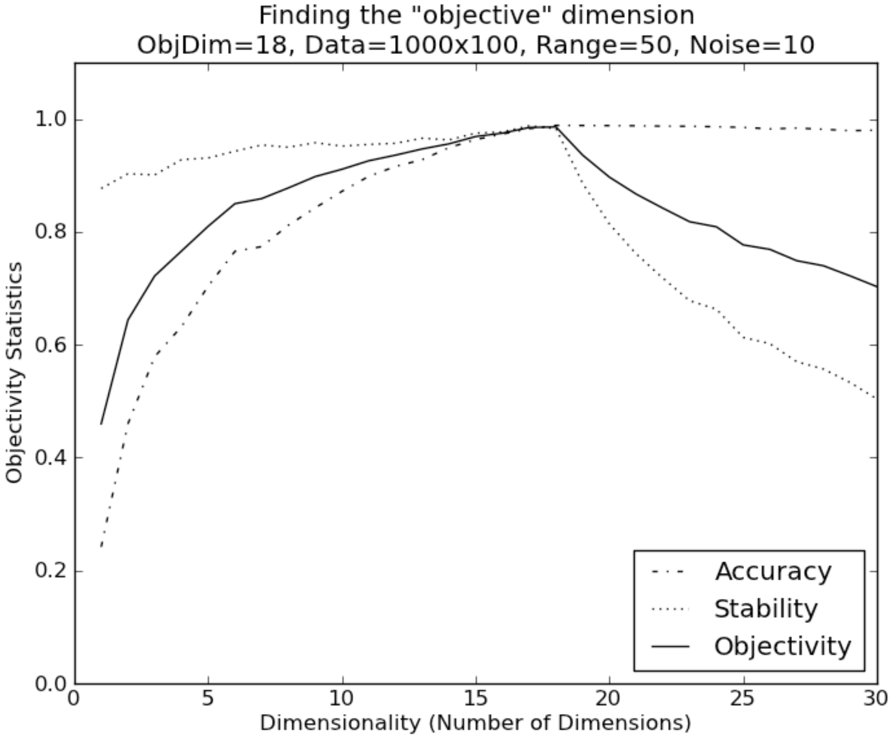
It is claimed that the dimensionality that yields the highest objectivity statistic (combining Accuracy and Stability) is, in general, the objective dimensionality of a dataset, i.e., the dimensionality that best predicts the “true” values. This is not a trivial claim, and requires proof on two fronts: a) proof that the dimensionality that best predicts pseudo-missing cells (Accuracy) is also the dimensionality that best predicts the “true” values; and b) proof that the dimensionality that produces the most stable coordinate structure (Stability) is also the dimensionality that best predicts the “true” values, and therefore that Accuracy and Stability agree in the objective dimension they imply.

Though the proofs are complicated, the intuition is simple enough. Ability to predict missing observations, aside from having intrinsic value, certainly seems like a reasonable proxy for predicting “true” values. And the same can be said for ability to set up a stable coordinate structure. However, there is more to it than that. A little reflection suggests that the Accuracy and Stability statistics are likely to behave somewhat differently.

We expect the model to do poorly predicting missing cells at lower than the objective dimension, but perhaps not all that badly above. Accuracy should grow quickly and plateau at the objective dimension. For stability, we expect the reverse. We expect the coordinate structure to be fairly stable at less than the objective dimension but to degrade above the objective dimension as the extra dimensions are used to model noise. Therefore, objectivity, which is a combination of stability and accuracy, should be a mountain-shaped curve with a noticeable peak at the objective dimensionality.

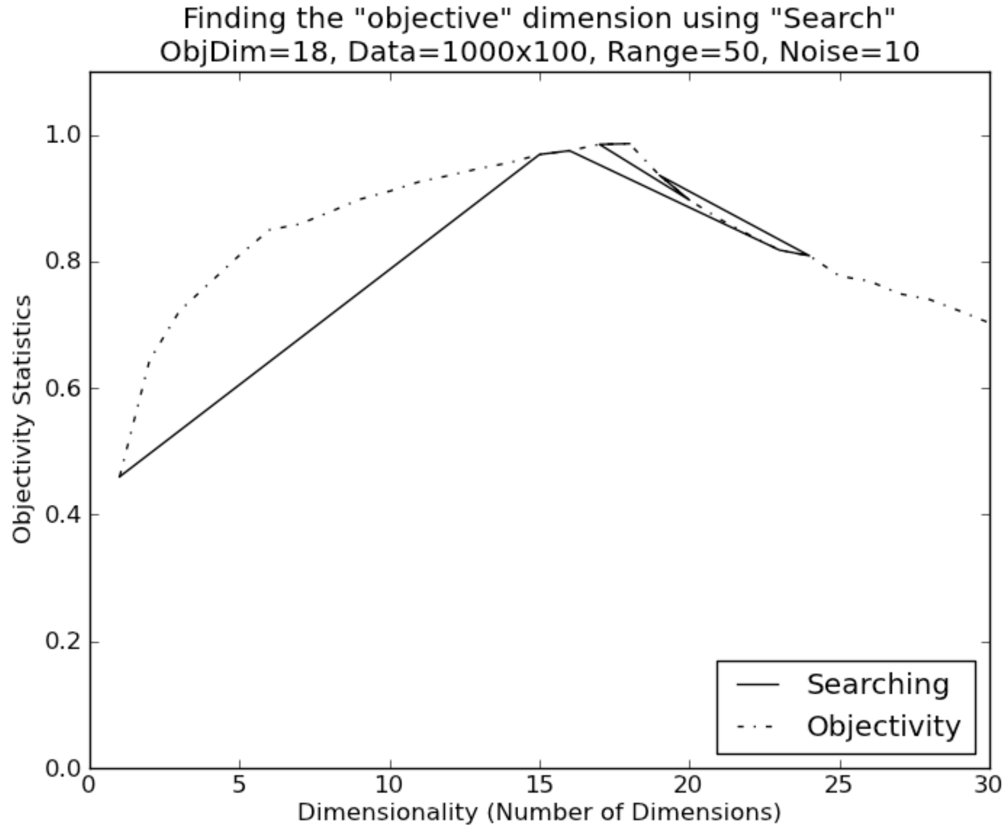
And that is what we find. Figure 3.3a shows the result of analyzing a dataset that was constructed to be 18-dimensional, to which noise was added. The NOUS objectivity curve peaks at 18 dimensions.

Fig. 3.3 a



In this case, to determine the objective dimensionality, NOUS calculated objectivity for every dimensionality from 1 to 30 -- computationally expensive. However, the fact that the objectivity curve has a well-defined peak suggests that it can be found more efficiently. Figure 3.3 b employs a binary search routine to home in on the objective dimensionality. The range is defined (30) and a baseline is calculated for dimension 1. A dimensionality is chosen in the middle of the range (15), and objectivity is calculated both for dimensionality 15 and dimensionality 16. If their slope is positive, a new midpoint is chosen between 15 and 30; if negative a midpoint is chosen between 1 and 15. A new pair of objectivity statistics is calculated and their slope calculated. In this way, it is possible to “scale the objectivity mountain” in an efficient manner.

Fig. 3.3 b



Instead of doing 30 complete NOUS runs (including the extra runs needed to estimate objectivity), we have located the objective dimensionality in 9. However, it is important to note that the objectivity curve is itself subject to error and there is no guarantee that all slopes to the left of the objective dimension will be positive or that all slopes to the right will be negative, in which case it is possible for the "Search" method to become stuck in local minima.

In addition, there is a point at which, as noise is added, the peak of the objectivity curve begins to shift to the left. At the extreme, where the noise is so great that it overwhelms the structure, the peak shifts all the way to 1. The dataset is essentially a matrix of random numbers.

3.4 "TRUE" ESTIMATES, PREDICTIONS OF MISSING CELLS

Conceptually, the "true" value of a given cell is the average of independent observations of that cell as *N* goes to infinity. For purposes of simulation in NOUS, the "true" or model value of a cell is the dot product of a randomly generated row vector and column vector. The full model array **T** is the (outer) product of randomly generated **R** and **C** matrices, whose rank determines the objective dimensionality *D*[objective] of the dataset. Data matrix **X** is simulated by adding random numbers whose average is zero (noise) to **T** and making some cells missing.

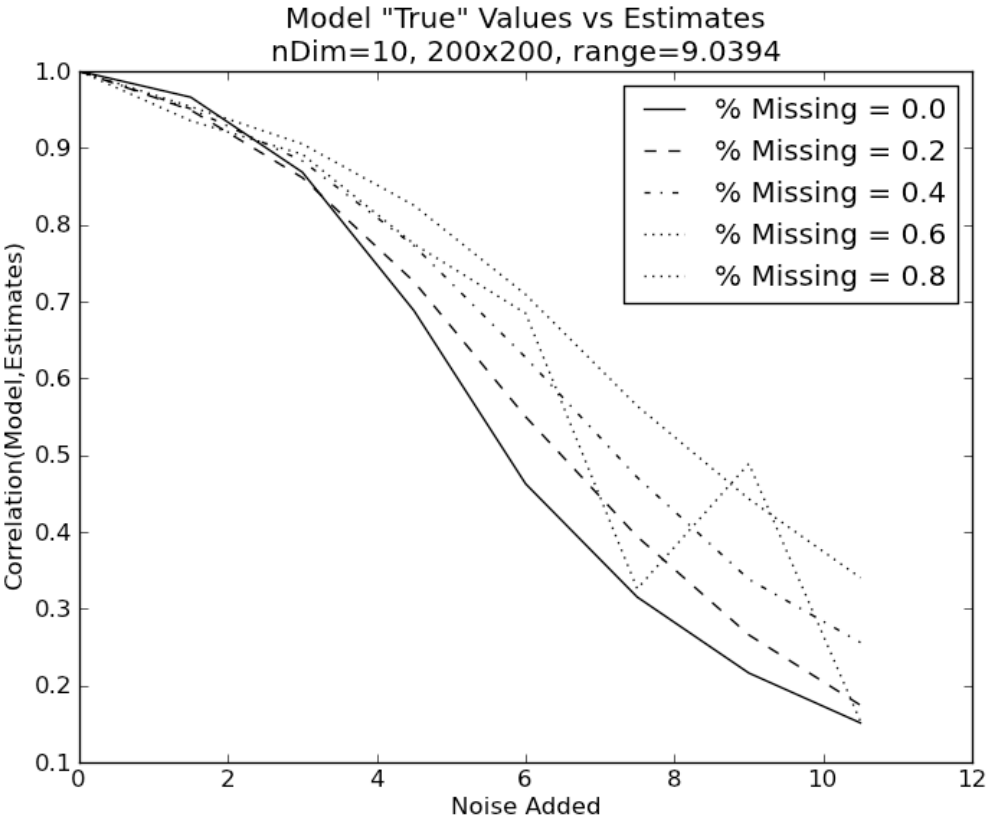
The intuitive basis of the claim that estimates matrix **E** approximates **T**, the "true" values, is that data array **X** can be assumed to be orthogonal to **T**. That is because adding random numbers whose average is zero to **T** creates **X** in such a way that: a) **X** ≠ **T** , and b) **X** is not biased in one direction or another away from **T**, and is thus perpendicular to **T**. When NOUS computes estimates **E** from **X**, it can be shown that

E is the orthogonal projection of **X** into the *D*[*objective*] subspace, which is the same subspace as **T**. Because **X** is orthogonal to **T**, and **E** is the orthogonal projection of **X** back into the **T** subspace, **E** will approximate **T**. In other words, computing **E** essentially reverses the process of adding noise to **T**. This, more or less, is why NOUS estimates approximate the true values. You can also see from this argument why it is so important that **E** be computed using the same dimensionality as **T**. Otherwise, it projects **X** into the wrong space.

As more noise is added to **T** in creating **X**, the farther apart the two matrices become. As **X** is increasingly buffeted by individual large errors, it is more likely to be accidentally thrown off its perpendicular to **T**, which in turn will disturb **E**, the projection of **X** back into the **T** subspace. This, among other reasons, is why the addition of noise can be expected to progressively degrade the accuracy of **E**.

Figure 3.4 shows the degradation that occurs in the correlation between the NOUS estimates and the "true" values as noise is added. It also shows the effect of missing data on the rate of degradation. The correlations are computed using the whole array of estimates, including those for missing cells.

Fig. 3.4



Broadly speaking, Figure 3.4 demonstrates that NOUS meets two essential specifications of an object-oriented multidimensional model:

1. "True" estimates. To the degree that the data fit the model, i.e., the noise is zero, the NOUS *estimate* for each cell matches the "true" model value.

2. Predicts missing data. To the degree that the data fit the model, the NOUS *prediction* for each missing cell matches the "true" value. This is evident in the graph by the fact that when noise = 0.0, the correlation between the model values and the estimates is 1.0, even when 80% of the data are missing.

We also see that as noise is added to **T** and the model violated, the NOUS estimates degrade as expected. The noise values on the *x*-axis represent the range of random values *above and below* a given model value. Thus, we can expect that when the noise is 2.0, i.e., when the observed value is within plus or minus 2 of the true value, the correlation we can expect is in the neighborhood of 0.90. Since, the range of data values is 9.0, the variation around each data value when noise = 2.0 represents almost 50% of the whole data range. That is quite a lot of error. As the noise increases to 4.0 and 5.0, the amount of variation around each data value takes up the entire data range and the accuracy of the estimates degrades rapidly.

We also note several oddities. First, of course, is that one of the curves (this is actually the 80% missing curve, though it's not clear in the legend) jumps around quite a bit when the noise exceeds 6.0. The causes for this are not known, but it is apparent that the degree of noise has surpassed what a linear system of equations consisting of, on average, 40 observations and 10 unknowns (dimensions) is able to handle. There are definite boundary conditions in the handling of noise and missing data, and in this case it appears we may have exceeded them.

Of equal interest is that, contrary to expectation, the curve with 0% missing data actually degrades the fastest to the right of the 2.0 noise threshold. When noise is less than 2.0 the percentage of missing data increases the rate of degradation, as one would expect. But to the right of the threshold, the percentage of missing data actually *decreases* the rate of degradation. This phenomenon is not yet understood.

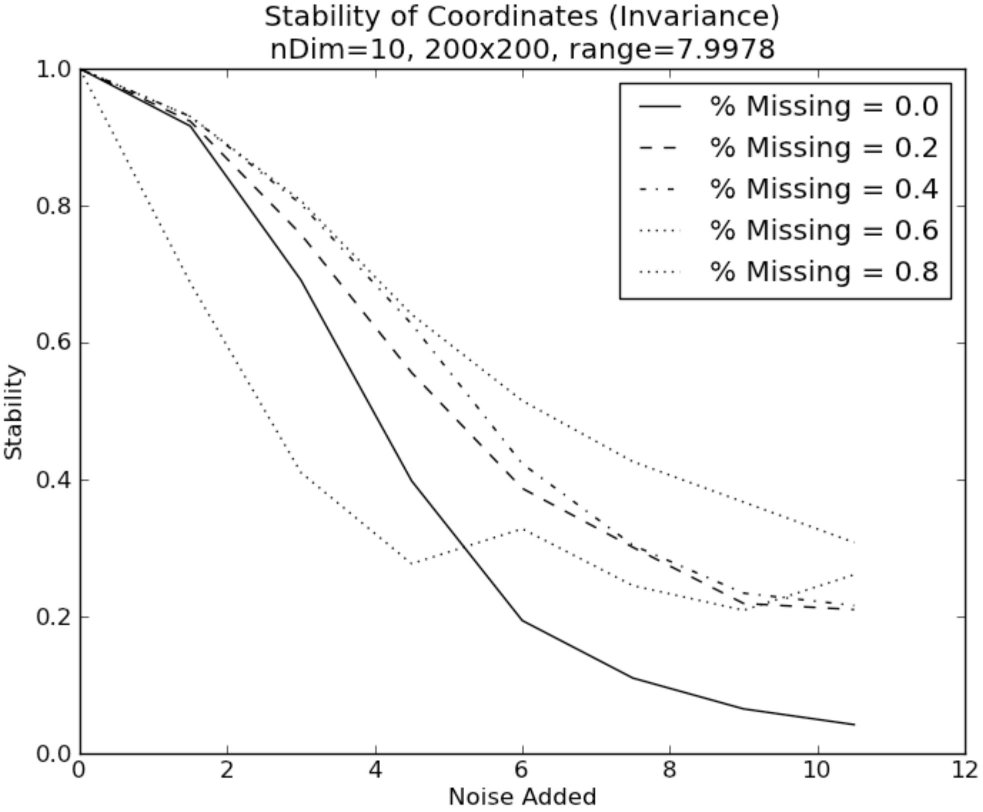
Rate of degradation is also affected by the size of the dataset and the number of dimensions. The larger the dataset and the smaller the number of dimensions, the slower the rate of degradation, though that is not shown here.

3.5 PERSON, ITEM INVARIANCE (STABILITY)

The Rasch model property of item and person invariance is operationalized in NOUS as "coordinate stability." It is the correlation between person coordinates **R** calculated using two different samples of items. In principle, stability would capture all variation in **R** associated with the use of all possible subsamples of **C**, but this is impractical and appears to be unnecessary.

Figure 3.5 tracks the stability/invariance of the coordinate system with different levels of noise and missing data.

Fig. 3.5



Like Figure 3.4, Figure 3.5 demonstrates that NOUS meets an essential specification of an object-oriented multidimensional model: To the degree that the data fit the model, i.e., the noise is zero, the person and item coordinates should be invariant. In this case, we see that Stability = 1.0 when noise = 0.0, which supports the claim. While the *y*-axis is technically only assessing row (person) stability, the demonstration is easily generalized to column (item) stability.

We see the same oddities as in Figure 3.4. We see that one of the curves (the 80% missing curve) degrades much faster than the others, reiterating the point that for this dataset, dimensionality, and amount of missing data, we appear to have exceeded some mathematical boundary condition. We also see, as before, that the curves with the least missing data in general degrade faster than the others. Again, this is a mystery.

3.6 MISFIT, INFLUENCE, INFORMATION, PORTABILITY, SUFFICIENCY

Misfit. Figures 3.4 and 3.5 suffice to demonstrate that, to the degree the data in general fit the model at the "objective" dimensionality, when a cell observation is "wrong" (i.e., does not match the "true" value for that cell), it will differ from the estimate accordingly. In other words, it will misfit. This follows from the property that NOUS estimates approximate model values under these conditions. This important property makes it possible to identify cells, persons, and items that do not fit in the common objective space of the model, and to remove them for purposes of calibrating the items. In other words, it provides a mechanism for forcing a common space, so that the dataset meets the conditions necessary to maximize objectivity.

Influence. Influence is closely related to misfit, its opposite in fact. It is the influence that the observation in a given cell has on its corresponding estimate. It shows up as the tendency of estimates to be artificially close to their observations, also known as overfit, which will also cause their standard errors to be artificially small. Two factors cause high influence: a) too many dimensions; b) too little data. For a dataset with the correct dimensionality, influence begins to appear when there are fewer than around 80 rows and columns or so. To correct the rosy (and non-objective) picture cased by influence, and in general to free estimates from the effects of their observations, NOUS contains a procedure for ignoring a given column *I* while calculating **R**, then anchoring **R** and calculating coordinates and estimates for *I*. Because **R** is not influenced by *I*, the influence of each observation in *I* on its corresponding estimate is minimized. This is augmented by a procedure for subtracting out the effect of individual observations on least squares solutions.

Information. The Rasch model can be used to generate multidimensional measures so long as it is possible to classify items by the construct they embody and do separate analyses for each construct. One drawback of this procedure is that it wastes information, inasmuch as constructs will often have significant correlations with each other, even if they contain orthogonally distinct elements. NOUS uses all the data in a dataset to compute **R** and **C** and enjoys the benefits of least squares solutions (and indirectly, maximum likelihood solutions), regardless of how its entities are oriented in space, and to that extent makes maximal use of information.

Portability. The item invariance property as measured by the stability statistic makes it possible to compute **C** from one section of the data and apply it to another section. This extends to outside datasets as well. Coordinates are calculated for Items I - K using data collected in one dataset, then applied as anchors to another dataset that also contains Items I - K, or some subset thereof, so long as the number of items is sufficient given the dimensionality to compute a valid solution. Similar portability exists for person coordinates.

Sufficiency. The coordinate vectors in **R** and **C**, taken together as a complete system, are "sufficient statistics" in the sense that no other statistic which can be calculated from the same sample provides any additional information as to the values of those coordinates. (Wikipedia, *Sufficient Statistic*, May 14, 2012). However, it is important to bear in mind that **R** and **C** are arbitrary for a given space -- the choice of origin point is arbitrary, as is the orientation of the axes. Consistency across spaces is maintained only by forcing all vectors to participate in the same coordinate system through item or person "anchoring".

In addition, as mentioned in section 2.1, NOUS (optionally) converts **R** into its orthonormal equivalent at each iteration. This has two immediate effects: a) it makes each column (dimension) of **R** orthogonal to every other column, like Cartesian coordinates; b) it normalizes each column in **R** to have unit length (their root sums of squares equal 1.0). Because each column has a length of 1, it can be shown algebraically that each coordinate in **C** equals the sum of squares of its corresponding contribution to the estimates in its column, and that the length of each vector in **C** bears a close relationship with the standard deviation of the estimates in its column. Given these vector lengths, the cosine of the angle between any two vectors in **C** can be used to estimate the correlation of the estimates in their columns. **C** becomes, in this case, "sufficient" in the additional, practical sense of being able to describe on its own, without reference to **R** or **E**, some of the essential statistical attributes of its column of data, somewhat analogous to the way a Rasch measure is able to stand in for a raw score. The full meaning and implications of this property have yet to be explored.

3.7 STANDARD ERRORS

Equation 17 gives the NOUS formula for standard error (SE). In this context, standard error is conceptualized as the expected discrepancy between the "true" model value for each cell and the corresponding estimate. It is distinguished from the expected absolute residual (EAR), which is conceptualized as the expected discrepancy between the *observed* value for each cell and the corresponding estimate. The latter is relatively easy to calculate because the observed values are accessible. The standard error is more elusive, as it concerns the relationship between the estimate and an unknown and hypothetical quantity -- the "true" value for each cell.

Figure 3.7 compares the root mean squared residual between the estimates and the "true" model values for each column with the corresponding root mean squared standard errors for those columns, where the formula for computing each cell standard error is given by Equation 17.

Fig. 3.7

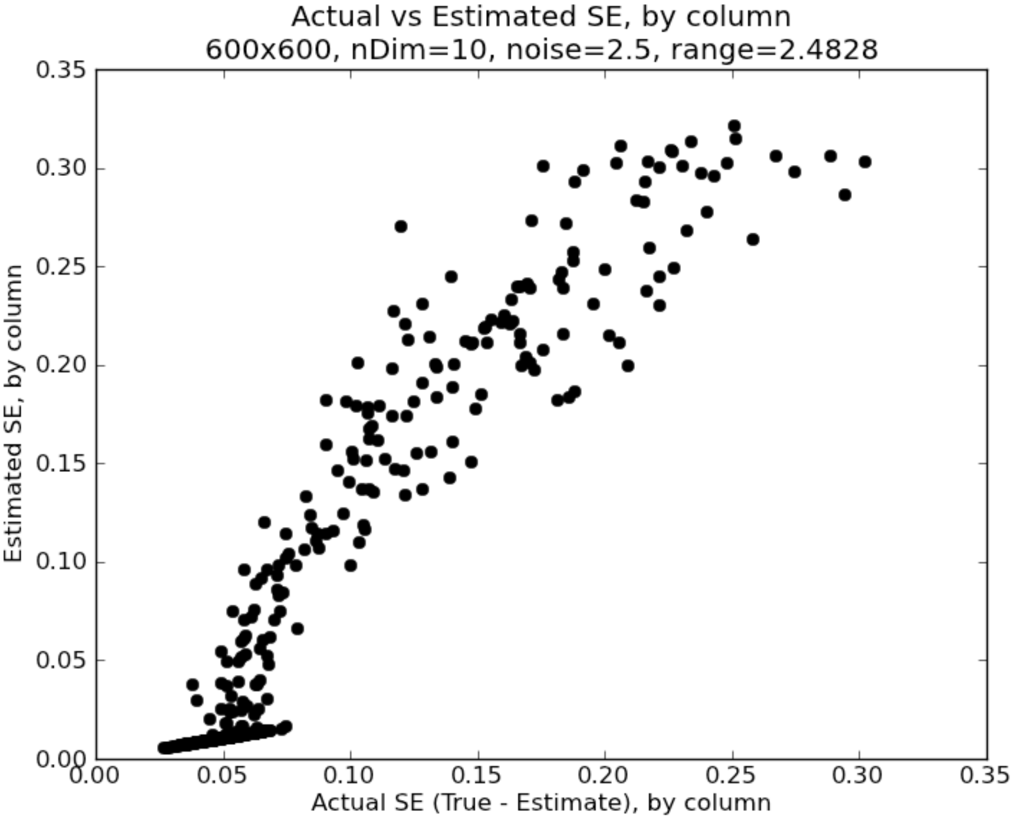


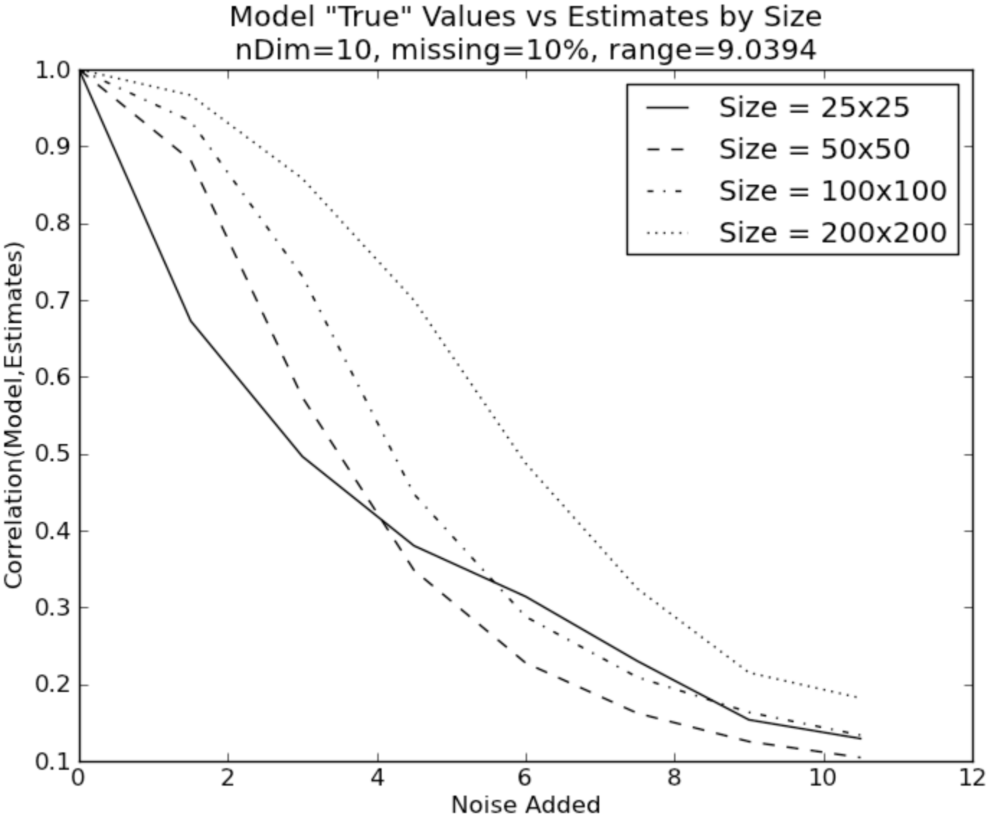
Figure 3.7 shows that Equation 17 works reasonably well. Though the relationship is somewhat non-linear close to zero, the estimated SE has a strong positive relationship with the "actual" standard error, and is in the correct range.

3.8 SIZE OF DATASET

Ability to scale *down* to small datasets is an important objective of multidimensional object-oriented models, not only because many datasets are small but also because a model cannot be considered object-oriented if it relies on representative person (or item) samples to be accurate or on large samples to be numerically stable. In NOUS, objects are persons, items, and other entities that interact to produce data. They are not populations or samples.

Figure 3.8 shows the rate of degradation in the correlation between "true" model values and NOUS estimates as noise is added for four differently sized datasets.

Fig. 3.8



We see, as expected, that the smaller the dataset, the more quickly it degrades with the addition of noise. What is interesting is that the estimates from smallest dataset, 25 rows by 25 columns, are not terribly different from the true values (0.60) when the noise added is plus or minus 2.0, or approximately 40% of the range of the data values. This is with a 10-dimensional dataset. While it is clear that datasets should ideally have hundreds of entities or more, this chart demonstrates that NOUS can be effective with small datasets, even when the dimensionality is relatively high.

3.9 DICHOTOMOUS DATA, NONNEGATIVE COORDINATES

In the world of probabilistic IRT models, the emphasis is on models that handle dichotomous or polytomous data, and the generalization to continuous interval data is awkward. In the world of ALS and NOUS, it is the other way around. To meet its homoscedasticity of errors requirement, linear least squares assumes interval data. This raises the important question, to what degree can NOUS, which depends on ALS, be generalized to dichotomous data?

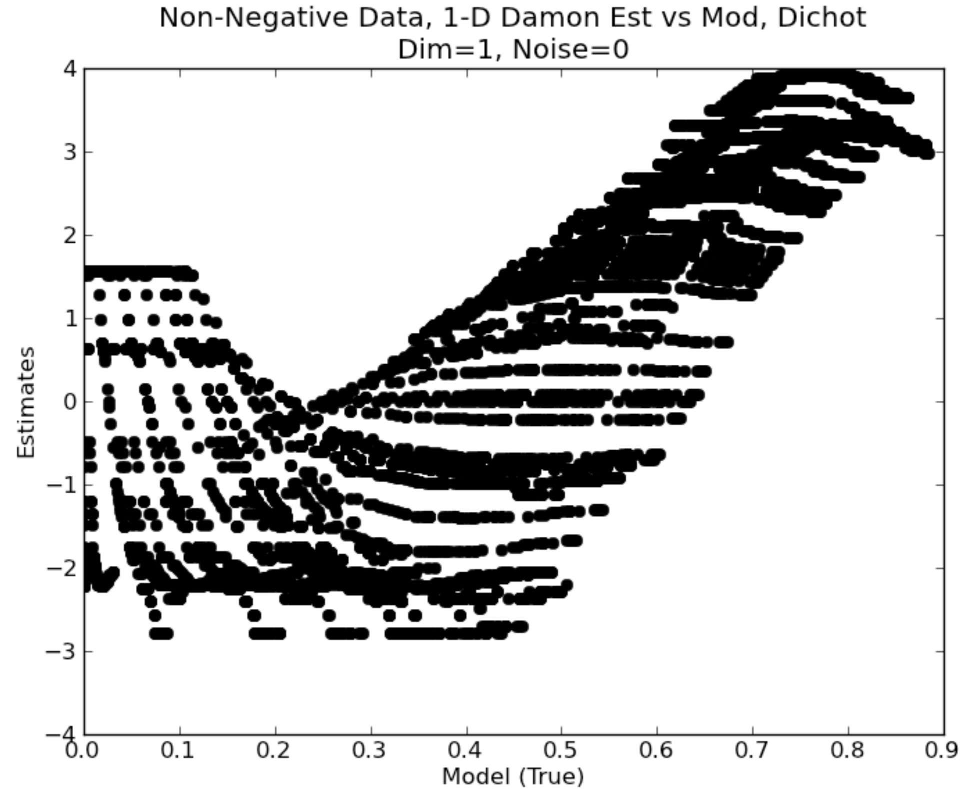
While this is a complicated question and outside the scope of this paper, the short answer is “to an acceptable degree”, but the results need to be interpreted somewhat differently than for interval data.

First, however, we need to be clear on what is meant by “nonnegative coordinates”. The Rasch model, as an example, requires positive coordinates. It assumes that person abilities and item difficulties can be expressed in terms of probabilities, which are positive (or “nonnegative” to use the linear algebra term), and this corresponds to datasets where the items are all positively correlated. Sometimes an item has a strong negative correlation with the other items. This suggests that it can be modeled only as a *negative* probability (a negative coordinate), which is not allowed in Rasch, so such items are either rejected or reverse-coded.

The key point, here, is that the sign of a coordinate value is to some extent a property of the item itself, not of the analysis technique applied to it. A simulated dataset built from all positive coordinates (nonnegative) behaves differently from a dataset built with a mix of positive and negative coordinates. Such coordinates I call “generating coordinates” as they are used to generate a dataset. In real life, the “generating coordinates” are not known -- they are the result of unseen natural forces that can only be inferred. The coordinates calculated by NOUS to estimate a data matrix (“estimating coordinates”) may be (in fact always are) quite different from the generating coordinates employed to create the data, with the sole exception of their rank or dimensionality, which must be discovered and is here termed the “objective dimensionality”. Ordinarily, the fact that the estimating coordinates are different from the generating coordinates is not a problem. One coordinate system is as good as another (for the most part). However, issues can arise when the proportion of negatives in the generating coordinates differs from their proportion in the estimating coordinates, which is what gave rise to the methodology known as nonnegative matrix factorization (NMF, see Lee & Seung, 1999).

For the most part, NOUS performs well with data created with nonnegative generating coordinates. The main exception is dichotomous data applied at a low dimensionality, such as one dimension. Figure 3.9a shows what happens when NOUS is run with a dimensionality of one on 1-dimensional, dichotomous, nonnegative data. No noise was added to the data aside from the (heteroscedastic) noise that naturally results from converting continuous values to dichotomous.

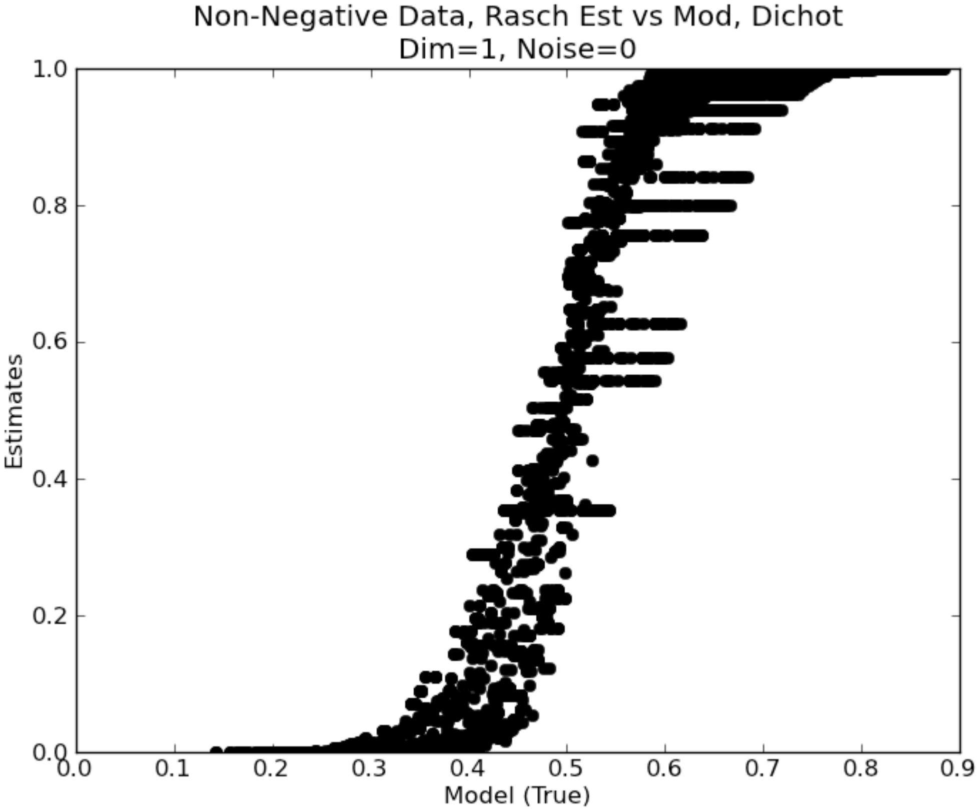
Figure 3.9a



NOUS does a poor job of capturing the "true" continuous values, disturbed particularly by a hook phenomenon at the lower end of the scale. This is due to the presence of negative coordinates in **R** and **C** calculated for a dataset that was built using only positive coordinates.

As Figure 3.9b shows, the same dataset run through Rasch is much better behaved.

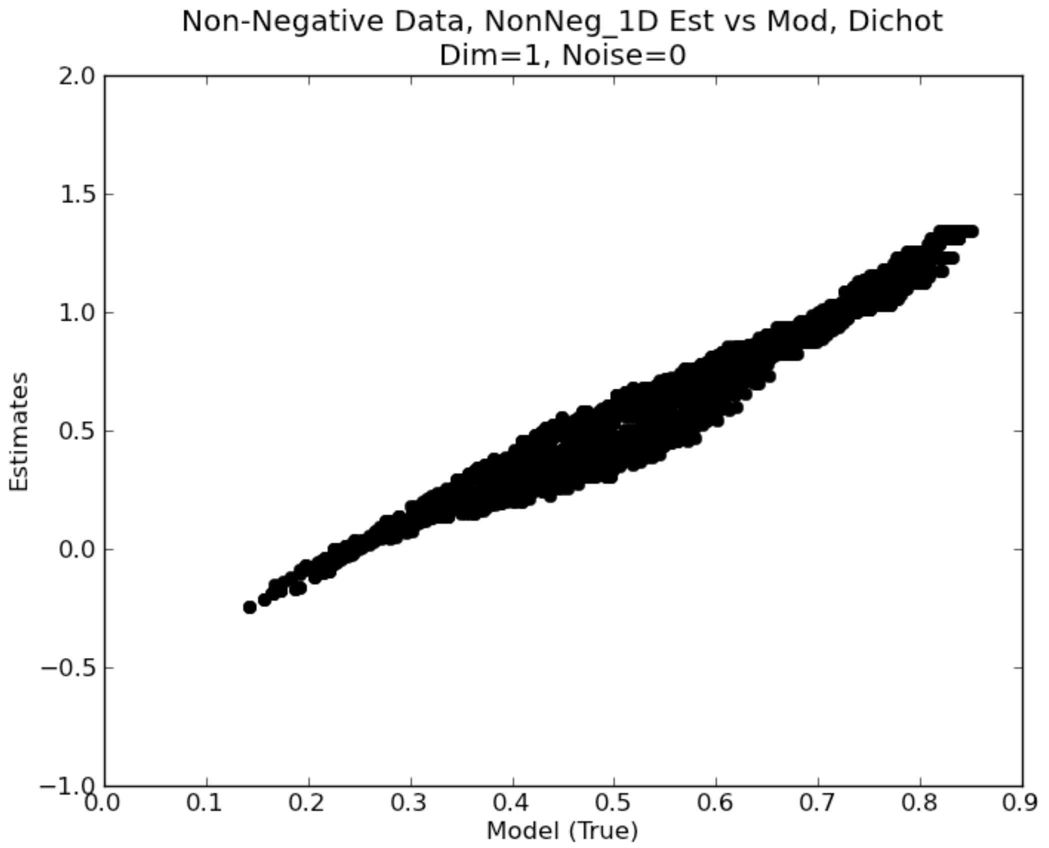
Fig. 3.9b



The Rasch estimates, being probabilities, exhibit a classic ogival relationship to the “true” model values. Thus, in a very real sense, the Rasch model is not just a special case of the model used by NOUS. It is a totally distinct model, applicable to datasets that are problematic for NOUS.

However, NOUS offers the option of applying special conditions to **R** and **C** in each alternating least squares iteration. Figure 3.9c shows what happens when **R** and **C** are constrained to be nonnegative.

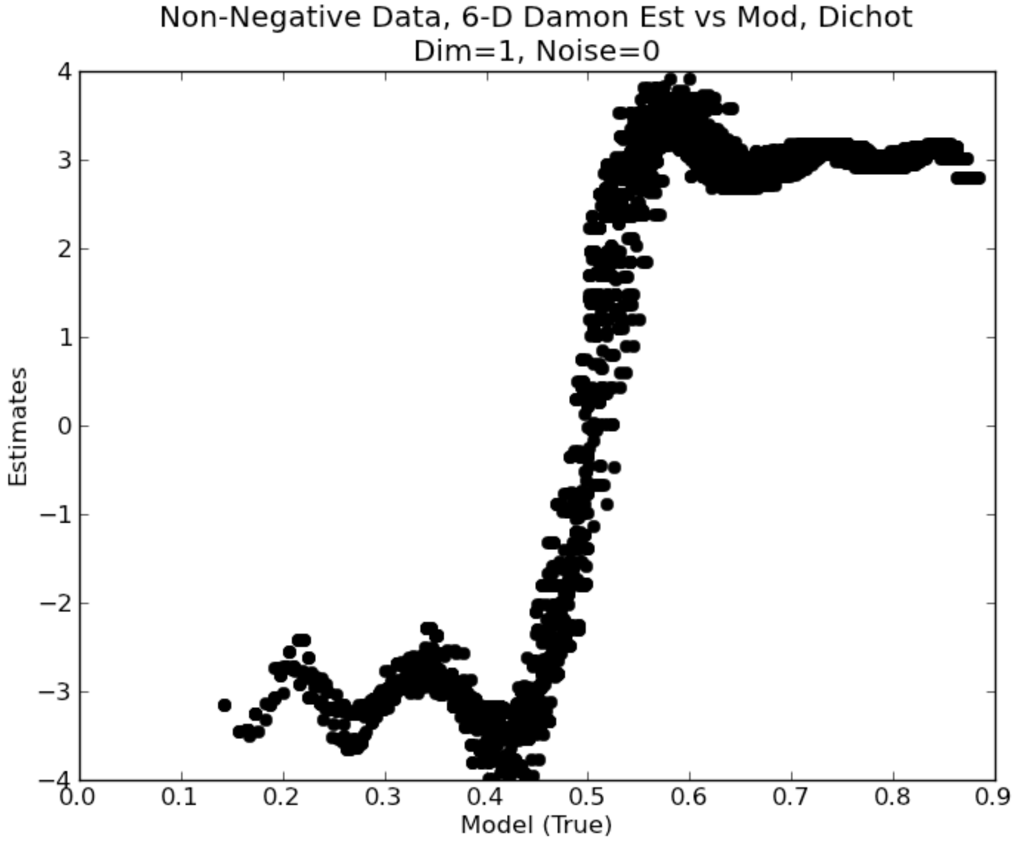
Figure 3.9d



Constraining NOUS to nonnegative coordinates makes all the difference. The resulting estimates have a strong linear relationship with the "true" model values, on par with Rasch.

Unfortunately, this approach requires insight by the analyst to invoke it at the appropriate time. Many applications require a more automated approach. In Figure 3.9a, NOUS was constrained to be 1-dimensional on theoretical grounds. Would NOUS find this dataset to be 1-dimensional if allowed to apply its objectivity criteria in the search for the optimal dimensionality? Figure 3.9e shows that the answer is no.

Fig. 3.9e



When left to its own devices, NOUS finds that a 6-dimensional solution maximizes the objectivity of the estimates. Yet the dataset was created using one dimension. What happened to the extra five dimensions? The answer is that NOUS has used them to model the tails with a complicated wave function. Another way to put it is that NOUS is trying to model the 0's and 1's as closely as possible, accepting them as a valid metric. Given enough dimensions, NOUS would convert the tails into (approximately) straight horizontal lines. It doesn't go past 6 dimensions, however, because this entails an unacceptable degradation in the objectivity statistic, driven primarily by its stability component.

Where does this leave us? As a metric representation of the the true values, the 6-dimensional solution is poor. As a set of dichotomous *predictions*, however, it is excellent. Because the y-axis is in a logit metric (the result of pre-standardization), every value above 0.00 represents a probability greater than 0.50, every value below a probability less than 0.50. Converting estimates to 0 and 1, we find that the 6-dimensional solution almost perfectly predicts the true *dichotomous* values.

Had the generating coordinates been an even mix of positive and negative, instead of all nonnegative, the story would be somewhat different. In the absence of noise, NOUS computes very accurate dichotomous predictions with few estimates around the 0.0 mark. Its main difference relative the nonnegative case, however, is that NOUS finds this solution at dimensionality 1 instead of dimensionality 6.

This allows us to state some general rules:

1. NOUS should be specified at the dimensionality that is found to produce the highest objectivity statistic (subject to analytic review), even (as in the case of nonnegative dichotomous data) when the dataset is known to be unidimensional.

2. When the data are interval, NOUS estimates for each column have linear metric properties; they are *measures*.

3. When the data are not interval but dichotomous or (to a lesser extent) polytomous, NOUS estimates for each column are not measures but *predictions*. They can be used to predict responses. To obtain measures in the dichotomous case, it is necessary to average or otherwise combine estimates across multiple columns to build a continuous construct with metric properties.

CONCLUSION

NOUS meets the specifications proposed for a true object-oriented multidimensional model. It handles highly multidimensional data, its estimates approximate "true" values, it is highly robust to missing data, and it predicts missing cells. It achieves these goals by combining the algorithmic properties of Alternating Least Squares with a set of objectivity criteria derived in part from the Rasch model. Among these criteria are that each analysis should be conducted at the "objective" dimensionality, an intrinsic property of the dataset. The "objective" dimensionality is discovered empirically by assessing the stability of the coordinate structure and the model's accuracy in predicting pseudo-missing cells for each of a set of possible dimensionalities, the objectivity curve tending to have a well-defined peak.

It is worth pausing to consider the overall goals of the model. Many machine-learning and data-mining methods deploy similar matrix factorization techniques, and better ones, but their goal seems to be to exploit patterns in the data to make useful predictions, not necessarily isolate new laws. The approach is somewhat analogous to that of Ptolemy and Copernicus who explained celestial phenomena by introducing new epicycles, new parameters, as needed. The goal of NOUS is more in line with that of Newton, to isolate laws, even at the expense of fit to the data. Inasmuch as laws are invariant relationships between two or more variables and NOUS selects item and person coordinates specifically to maximize their invariance across samples and their ability to predict unknowns, NOUS represents a step in the direction of using mathematical methods not merely to describe the real world but to discover its most stable underlying laws.

REFERENCES

Adams, R. J., Wilson, M., and Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, vol. 21, no. 1, 1-23.

Appellof, C. J., and Davidson, E. R. (1981). Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents, *Analytical Chemistry*, 53, pp. 2053-2056.

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition, *Psychometrika*, 35, pp. 283-319.

Carroll, J. D., Pruzansky, S., and Kruskal, J. B. (1980). CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters, *Psychometrika*, 45 (1980), pp. 3-24.

Eckart, G. and Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika*, 1, pp. 211-218.

Harshman, R. A. (1970). *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA working papers in phonetics, 16, pp. 1-84.

Hu, Y., Koren, Y., Volinsky, C. (2008). Collaborative Filtering for Implicit Feedback Datasets. IEEE International Conference on Data Mining (ICDM 2008), IEEE (2008). Yahoo! Research. Available at http://research.yahoo.com/pub/2433.

Kolda, Tamara G. and Bader, Brett W. (2009). Tensor Decompositions and Applications. *SIAM Review*, 51(3), 455-500. Permalink: http://dx.doi.org/10.1137/07070111X.

Kroonenberg, P. M., and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms, *Psychometrika*, 45 (1980), pp. 69-97.

Lee, Daniel D. and Seung, H. Sebastian (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature* **401** (6755): 788-791. doi:10.1031/44565. PMID 10548103.

Linacre, J. P. (1989/1994). *Many-Facet Rasch Measurement*, 2nd Edition. Institute for Objective Measurement. ISBN 0-941938-02-6. LC# 94-76939.

Moulton, M. (2004). One Use of a Non-Unidimensional Scaling (NOUS) Model: transferring information across dimensions and subscales. Educational Data Systems, Inc. Available at http://eddata.com/resources/publications/EDS_NOUS_Measurement.pdf.

Moulton, M. and Silsdorf, H. (2006). Multidimensional Equating: linking multidimensional test forms by constructing an objective n-space. Educational Data Systems, Inc. Presented at IOMW 2006, Berkeley, CA. Available at http://eddata.com/resources/publications/EDS_Moulton_IOMW_2006.pdf.

Netflix prize (2009). Information available at www.netflixprize.com.

Owen, S. (2012). Simple Matrix Factorization for Recommendation. Slide show presented at Data Science London. Available at HTTP://WWW.SLIDESHARE.NET/DATASCIENCELONDON/MATRIX-FACTORIZATION-MAHOUT-SPECIAL-SESSION.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with forward and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer. ISBN-10 0387899758.

Singular Value Decomposition (May 14, 2013). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Singular_value_decomposition.

Sufficient Statistic (May 14, 2013). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Sufficient_statistic.

Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. *Problems in Measuring Change*, C. W. Harris, ed., University of Wisconsin Press, 1963, pp. 122-137.

Yahoo! KDD Cup (2011). JMLR Workshop and Conference Proceedings, Volume 18: Proceedings of KDD Cup 2011. Available at http://jmlr.csail.mit.edu/proceedings/papers/v18/.