Estimation of Cronbach's alpha for sparse datasets

Mike Lopez

mike.lopez@manukau.ac.nz

Abstract

Cronbach's alpha is widely used to evaluate the internal consistency of a psychometric instrument. Its popularity is largely based on a straightforward interpretation in terms of correlations, its ease of calculation and the guidance it gives to building a single dimensional scale. The standard calculation of alpha, however, requires a complete dataset and can give misleading results with sparse datasets.

An alternative method of calculating an equivalent to Cronbach's alpha is proposed that retains the essence of alpha and can be readily calculated for sparse datasets. A theoretical basis is given and the method is evaluated and validated against generated datasets.

Keywords: numerical methods, Cronbach's alpha, sparse datasets, psychometrics.

1 Problem identification and motivation

This paper is organised according to the Design Science Research Process (DSRP) (Peffers, Tiuuanen, Gengler, Rossi, Hui, & Virtanen, 2006)

1.1 Background

In 1951, Lee Cronbach proposed "alpha" as a measure of the internal consistency of a test (Cronbach, 1951). Alpha may be seen as a generalisation of the earlier KR20 formula, extending it from dichotomous to continuous variables. Louis Guttman developed the same measure under the name "lambda-2" in 1945.

Alpha is defined as:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \left(\frac{\sum_{i=1}^{k} \text{var}(S_i)}{\text{var}\left(\sum_{i=1}^{k} S_i\right)}\right)\right)$$

Where k is the number of items in the instrument and S_i represents the score for item i.

Bland and Altman (1997) characterise alpha thus:

> "Cronbach's alpha has a direct interpretation. The items in our test are only some of the many possible items which could be used to make the total score. If we were to choose two random samples of k of these possible items, we would have two different scores each made up of k items. The expected correlation between these scores is alpha". (p 572)

Alpha can thus be seen as combining two ideas: the average correlation of items in a dataset and a correction that adjusts this to reflect the correlation expected between samples drawn from the population of all the possible conceptual items that relate to what we are measuring.

This correction is analogous to the Spearman-Brown correction we apply when we step up the correlation in split-half reliability.

These two ideas form the strength of alpha in giving guidance when scale building. In general, the more items there are, the higher the reliability. On the other hand, items with a poor item-total correlation reduce the reliability. Many researchers follow the guideline of deleting items with low item-total correlations if and only if the deletion would result in an improvement in alpha.

Many researchers also follow the guideline that alpha should be at least 0.7. This is based on the intuitive interpretation that:

$$0.7 \approx 1/\sqrt{2} = Cos(45^\circ)$$

This intuition is based on two ideas:

- The concept of angle may be generalised to multiple dimensions by defining the cosine of the angle as the standardised dot product of two vectors:

$$\cos(\theta) = \frac{a \bullet b}{|a||b|}$$

- An angle below 45° suggests that the measure is "more on target than off" or equivalently that the "signal" component in the instrument is stronger than the "noise".

1.2 Problem identification

Deliberately sparse datasets offer considerable potential for increasing the efficiency of data collection.

This quality assured paper appeared at the *20th Annual Conference of the National Advisory Committee on Computing Qualifications* (NACCQ 2007), Nelson, New Zealand. Samuel Mann and Noel Bridgeman (Eds).

For example, we may create parallel forms of an assessment instrument by using a number of "booklets" each of which has a number of items drawn from a common bank. A typical goal of this approach is to achieve the higher accuracy that is possible with a larger number of items without overburdening each participant. Often, each participant combines some common items with some matrix-sampled items from the item bank. Further details of this approach are given in Mislevy, Beaton, Kaplan, & Sheehan (1992).

We may also use adaptive sampling techniques to target items to cases, based on the data collected to date. This enables the use of a smaller number of items to achieve the level of information and accuracy required. In a report on a large scale (N>250,000) study, Kingsbury and Hauser (2004) comment:

> "The results from the impact analysis identify the percentage of students who aren't being measured with the precision required for instructional decision making. In no condition did the adaptive test have more than 1 percent of the students imprecisely measured. In no condition did either fixed-form test have less than 6 percent of the students imprecisely measured. " (p 11)

This is a compelling argument for the use of adaptive testing. The need for selection at the item level inevitably leads to an approach based on item response modelling (IRM). Lord & Novick (1968) give a good overview of item response modelling theory.

Nevertheless, classic test theory still has a role to play in the analysis and interpretation of results at the test, rather than item, level. The sparse datasets that result, however, can cause problems for many of the classic test measures.

The issue with alpha may be illustrated by the artificial sample data shown in table 1:

Table 1

	Item1	Item2	Item3	Item4	Total
Case 1	1	2	3	4	10
Case 2	2	3	4	5	14
Case 3	3	4	5	6	18
Case 4	4	5	6	7	22

These data are designed to give a perfect correlation. As expected, the perfect correlations give an alpha of 1.0 with the standard definitional formula. However, if we delete (say) the value for Case 1, item 4, we will get an alpha coefficient of 1.16, which is clearly wrong.

Many statistical packages (eg SPSS) simply delete cases with incomplete data. Clearly this is inappropriate with deliberately sparse data; it might delete the entire dataset. Other techniques, such as substituting the mean or interpolating through a regression technique also give misleading results.

One option is to use alternate measures to alpha. For example, Andrich (1982) has defined a person separation index which is a logically similar measure from an item response modelling perspective. However, alternate measures may fail to utilise the intuitive understanding that researchers have built through experience with alpha on complete datasets.

1.3 Motivation

This work was motivated by development of the OTIS software package (Lopez, 2007).

Although the software is based mainly on item response modelling, one goal of the package was to build a bridge between a traditional test approach and IRM. IRM is new to many practitioners, so it was important to utilise, wherever possible, familiar measures so as to help build confidence in the IRM approach.

IRM approaches handle sparse data simply and naturally because they shift the focus from test evaluation to item evaluation. The challenge was to provide an equivalent measure to Cronbach's alpha that would work for sparse datasets.

1.4 Relevance

We come across sparse datasets in many assessment, measurement and research situations. Simply deleting cases because they do not fit the measure is at best inappropriate. If a measure does not fit the data, we should change the measure, not the data.

Conversely, suitably modified measures and techniques can become an enabling technique, leading researchers to consider the deliberate use of sparse datasets.

2 Objectives of a solution

From the above discussion, the central objective is to define an alternative to Cronbach's alpha that is:

- Computationally tractable for sparse datasets
- Capable of an equivalent interpretation to Cronbach's alpha.

For the measure to be acceptable to the research community, it is also important that the measure can be simply and readily calculated within a spreadsheet, or within a researcher's software package of choice.

3 Design and development

This section sets out the theoretical basis for the algorithm and the algorithm itself.

3.1 Theoretical basis

Following the Spearman-Brown prediction formula, an alternate conceptualisation of alpha is:

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$

Where k is the number of items and $\bar{r}$ is the average item-item correlation. In interpreting this formula, it is worth noting that the numerator is what the total of each column of the correlation matrix would be if all cells were $\bar{r}$ and that the denominator is what the column total would be with one cell (the correlation of an item with itself) being 1 and the rest $\bar{r}$.

This alternate definition is tractable for sparse datasets and forms the basis of the algorithm. It is also convenient. In practice, we usually calculate a correlation matrix as part of our instrument evaluation and all the data we need is contained in the correlation matrix. We can thus regard alpha as a function of the correlation matrix and it can be calculated simply by a formula within a spreadsheet or statistical package.

In programming terms, it can simply be a method of a correlation matrix object.

3.2 Algorithm

We can express the algorithm thus:

1) Calculate the correlation matrix.

2) Calculate the mean of the cells off the main diagonal

3) Evaluate alpha with the formula given above.

Applying this algorithm to the sparse version of table 1 gives the correct answer of 1.0 for the example shown.

4 Demonstration

No scale is ever truly one dimensional. Every item brings its own uniqueness as well as (hopefully) telling us something about the overall characteristic we are trying to assess.

To simulate this, each item in the dataset was given a component which was unique as well as one that tapped into the overall characteristic. No secondary factors were used; each item's unique contribution component was orthogonal.

To demonstrate the algorithm, a sparse dataset was constructed using Microsoft Excel with these characteristics:

- There were 30 cases
- Four items were used
- The main factor accounted for 2/3 of the variability; the independent second factor for each item, 1/3 of the variability.
- The independent factors were constructed using Excel's RAND function.
- All factors used a rectangular distribution.
- A sparse dataset was created with 10 cases having responses to items 1 and 2, 10 with responses to 3 and 4, and 10 with responses to all four items.
- Cronbach's alpha was calculated for the full dataset using both the standard and proposed algorithms.

5 Evaluation

To meet the set objectives, the algorithm must be correct, easy to calculate and capable of an interpretation that is consistent with the standard alpha. These are considered in turn.

5.1 Correctness

A series of 20 runs was carried out, with Excel allocating new values for the random variables on each run.

Table 2 shows the results from 20 runs of the spreadsheet. Given that by design, 2/3 of the variability comes from the main factor, the true average correlation is √ (2/3) (0.8165) and alpha (from Spearman-Brown) is 0.9468.

Table 2

Dataset	Full	Full	Sparse	Sparse
Method	Standard	Proposed	Standard	Proposed
1	0.94	0.94	1.14	0.95
2	0.95	0.95	1.08	0.92
3	0.95	0.95	1.10	0.96
4	0.95	0.95	1.09	0.95
5	0.93	0.94	1.14	0.93
6	0.93	0.94	1.14	0.93
7	0.93	0.93	1.15	0.93
8	0.94	0.95	1.08	0.95
9	0.95	0.95	1.12	0.95
10	0.96	0.96	1.10	0.96
11	0.95	0.95	1.09	0.95
12	0.91	0.92	1.11	0.91
13	0.94	0.95	1.09	0.95
14	0.93	0.93	1.12	0.93
15	0.95	0.95	1.07	0.96
16	0.93	0.93	1.10	0.95
17	0.95	0.95	1.10	0.95
18	0.94	0.94	1.12	0.95
19	0.94	0.94	1.10	0.94
20	0.94	0.95	1.08	0.94

The use of random values guarantees the independence of the factors. We would not expect to recover the parameters exactly, due to sampling error. However, the variability in the estimates can also give us a rough guide as to the accuracy of the estimates.

Table 3

Method	Data Set	Mean Estimate	Diff.	Std. Dev.	P value (sig.)
Standard	Full	0.9405	.0063	0.55	0.5828
Proposed	Full	0.9435	.0033	0.33	0.7376
Standard	Sparse	1.1060	.1592	6.84	<0.0001
Proposed	Sparse	0.9430	.0038	0.28	0.7833

Table 3 sets out, for each of the methods and conditions shown:

- The mean estimate of alpha.
- The difference from the true value.
- The difference expressed as multiples of the standard deviations observed.
- The two-tailed probability of getting results this extreme under a normal distribution for each of the measures treated separately.

The estimates given by the standard algorithm on sparse data are clearly ($p<0.0001$) outside the range expected by sampling variability.

The estimates given by the standard algorithm on the full dataset and by the proposed algorithm on both full and sparse datasets are all consistent with the true figure.

5.2 Ease of calculation

The calculation is straightforward, requiring only a correlation matrix and basic arithmetic. It can therefore be implemented easily in a spreadsheet; a spreadsheet approach was chosen for this paper specifically to demonstrate this.

Specifically, in Excel, cells in the correlation matrix are defined using the CORREL function, the average correlation ($\bar{r}$) is defined with the AVERAGE function and alpha with the formula given in 3.1 above.

A sample Excel spreadsheet is available from the author.

5.3 Interpretation

When used with a complete dataset, alpha as defined herein has an identical interpretation to alpha defined by the traditional method.

When used with a sparse dataset, the alpha defined herein can be interpreted as the value that would have been obtained if the dataset were complete.

It can therefore be used for the same scale-building purposes as the conventional calculation, but with the additional capability that a sparse dataset may be used to identify items with a poor scale fit, if the researcher so chooses.

6 Communication

Two examples of the use of sparse datasets are given below.

6.1 Scale building

The ability to use sparse datasets for scale building has a major practical implication. In practice, we often start with many more items than we intend to use in the final scale, culling them (often using alpha) as we progress through the trialling and scale building process. To administer the full set of items may be an unacceptable burden to the trial participants.

With a sparse dataset approach, we can divide the proposed items into a number of sets and allocate sets to participants in such a way that:

- No participant responds to more than two sets.
- Each set is responded to by at least two groups of participants.
- Each set is connected to every other set through some participants.

For example, 100 potential items could be split into five sets of 20 and each participant given two of these sets for a total of 40 items. One possible scheme for this is given in table 4 below:

Table 4

	Set A	Set B	Set C	Set D	Set E
Unit 1	Yes	Yes			
Unit 2		Yes	Yes		
Unit 3			Yes	Yes	
Unit 4				Yes	Yes
Unit 5	Yes				Yes

In this design, participants should be randomly allocated to units for classical test theory measures such as alpha. The random allocation is not required for item response modelling approaches.

This Balanced Incomplete Block (BIB) is a popular approach: details are given in the NAEP 1998 Technical Report (Allen, Donoghue, & Shoeps, 1998). The authors comment:

> "Both the BIB and PBIB designs provide for booklets of interlocking blocks of items, so that no student receives too many items, but all receive groups of items that are also presented to other students."

6.2 Adaptive assessment

Parallel test forms, using a booklet approach have been used for some time, but with the increasing use of computer administered assessment, it becomes possible to select the items administered dynamically, based on previous responses.

This enables the use of a smaller number of items to achieve the level of information and accuracy required, thus reducing the burden on the student.

A typical approach would be to form a rough estimate of ability from part one of a test, and then use this to target items of appropriate difficulty in a second part. This typically requires an item modelling approach which, in turn, requires uni-dimensionality of scale items.

Alpha forms a natural measure of uni-dimensionality.

7 Conclusions

For rigour, further work is needed on the distribution, before confidence intervals for the estimate can be given.

However, the proposed algorithm is correct from a theoretical perspective and it is clear that the estimates produced are at least "in the ballpark". This is clearly not true for the standard algorithm when used with sparse datasets.

It is hoped that this proposed algorithm will enable researchers to consider the deliberate use of sparse datasets, without the necessity of abandoning the familiar measures of alpha.

8 References

Allen, N., Donoghue, J., & Shoeps, T. (1998). The NAEP 1998 Technical Report. *Education Statistics Quarterly , 3* (4).

Andrich, D. (1982). An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. *Education Research and Perspectives , 9* (1), 95-104.

Ariel, A., Veldkamp, B., & van der Linden, W. (2002). *Constructing Rotating Item Pools for Constrained Adaptive Testing.* AE Enschede, The Netherlands: Faculty of Educational Science and Technology, University of Twente.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ , 314* (7080), 572-.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. , 16*, 297-334.

Kingsbury G.G., & Hauser, C. (2004). *Computerized Adaptive Testing and No Child Left Behind.* San Diego, CA: American Educational Research Association.

Lopez, M., (2007). *Otis software*. Auckland, New Zealand: Department of Computing and Information Technology, Manukau Institute of Technology.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement , 29*, 133-161.

Peffers, K., Tiuuanen, T., Gengler, C. E., Rossi, M., Hui, W., & Virtanen, V. B. (2006). The Design Science Research Process: A model for producing and presenting Information Systems research. *DESRIST 2006.*