Using cross-validation to determine dimensionality in multidimensional scaling

Russell Richie (drrichie@sas.upenn.edu)
Department of Psychology, Philadelphia, PA, USA

Steven Verheyen (verheyen@essb.eur.nl)
Department of Psychology, Rotterdam, Belgium

Keywords: similarity; multidimensional scaling; cross-validation; model selection; dimensionality

Introduction

Multidimensional scaling (MDS) is a popular technique for embedding items in a low-dimensional spatial representation from a matrix of the dissimilarities among items (Shepard, 1962). MDS has been used simply as a visualization aid or dimensionality reduction technique in statistics and machine learning applications, but in cognitive science, MDS has also been interpreted as a cognitive model of similarity perception or similarity judgment, and is often part of a larger framework for modeling complex behaviors like categorization (Nosofsky, 1992) or generalization (Shepard, 2004). However, a persistent challenge in application of MDS is selecting the latent dimensionality of the inferred spatial representation; the dimensionality is a hyperparameter that the modeler must specify when fitting MDS.

Perhaps the most well-known procedure for selecting dimensionality is constructing a scree plot of residual stress (the difference between empirical dissimilarities and dissimilarities implied by the model) as a function of dimensionality, and then looking for an elbow: the dimensionality where stress has decreased dramatically but then plateaus. This elbow is taken to indicate that extending the space with additional dimensions does not substantially improve the fit of the model to the input similarities. Unfortunately, this procedure is highly subjective. Often such elbows do not exist, and instead the scree plots show a smooth decrease in stress as MDS increasingly overfits to noise at higher dimensionalities. In response, various more principled statistical techniques for model selection have been proposed that account for the trade-off between model complexity (dimensionality) and model fit (stress), including likelihood ratio tests (Ramsay, 1977), BIC (Lee, 2001), and Bayes factors (Gronau and Lee, in press). While such techniques are valuable, they can be prohibitively computationally complex for novice MDS users, and rely on a number of assumptions that are not necessarily met (e.g., Storms, 1995).

An alternative technique that may avoid such problems is cross-validation. Under this approach, MDS of a given dimensionality would be fit to some subset of available dissimilarity data, the model's predicted distances for held-out dissimilarity data would be evaluated, and the dimensionality which maximizes performance on the held-out data would be selected. Despite the simplicity and generality of cross-validation as a model selection procedure, cross-validation has seen relatively little application to MDS or related methods (Steyvers, 2006; Roads & Mozer, 2019; Gronau & Lee, in press), with no systematic exploration of its capabilities, as there has been for likelihood ratio tests, BIC, and Bayes factors (Ramsay, 1977; Lee, 2001; Gronau & Lee, in press). In the present work, we therefore examine the usefulness of cross-validation over cells of a dissimilarity matrix in simulations and applications to empirical data.

Simulations

We conducted a standard model recovery exercise, whereby we simulated spaces of known dimensionality, from which we collected and aggregated noisy dissimilarity data, and applied cross-validation to attempt to recover the true dimensionality. Our simulations were conducted as follows:

1. Sample *n* items uniformly from the unit hypercube of dimensionality between 1 and 7
2. For each simulated subject, add noise ~ *N(0, sd)* to the item coordinates, and compute the inter-item Euclidean distances
3. Average over subject distance (dissimilarity) matrices
4. Derive a weight for each cell of the average distance matrix equal to the inter-subject precision of that cell
5. Generate 10 random 80-20 train-test splits of the averaged matrix such that each row of each training matrix is missing no more than 75% of its cells
6. For each train-test split:
    1. For each dimensionality from 1 to 7:
        1. Fit ratio MDS to the training dissimilarities and the cell weights given by (4), using the *smacof* library in R (de Leeuw & Mair, 2017)
        2. Use the fitted MDS to obtain distances for the cells of the test split
        3. Compute Pearson correlation between the MDS distances and held-out dissimilarities
7. Select the dimensionality with the highest median correlation across all train-test splits

Figure 1 shows distributions of best-fitting dimensionalities (y-axis) over 50 simulations of a particular true dimensionality (x-axis), number of subjects (hue), noise level (columns), and number of items (rows). Figure 1 shows the true dimensionality is recoverable across a range of conditions.

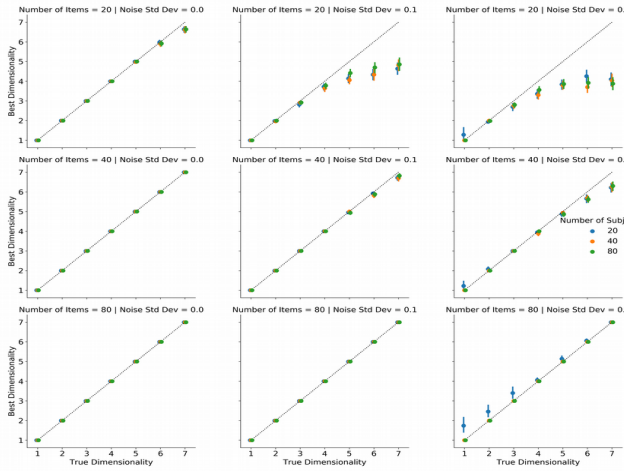

Figure 1. Error bars are 95% confidence intervals.

Empirical application

We applied steps 3-7 above to an empirical dataset of similarity judgments from Hout, Goldinger, and Ferguson (2013), who had 92 subjects use the Spatial Arrangement Method to judge similarity among a set of 27 artificial 'bugs' which varied on 3 dimensions (darkness, pincer shape, number of legs). Figure 2 shows distributions of Pearson correlations between MDS distances and held-out dissimilarities under 100 train-test splits for each fitted dimensionality from 1 to 7. Cross-validation correctly selects a 3-dimensional spatial representation for these data.

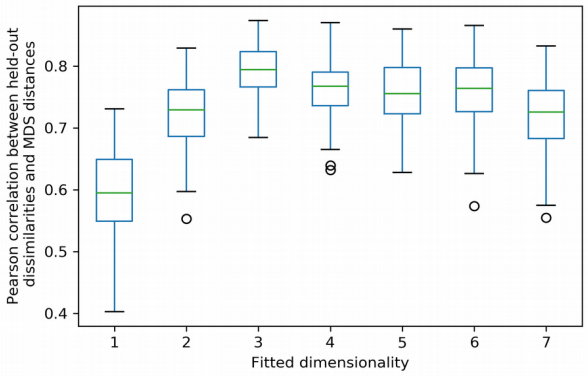

Figure 2

Discussion

We have demonstrated the utility of cross-validation for determining the dimensionality of multidimensional scaling models, given subject-averaged similarity data and assumptions (or knowledge) that dissimilarity data are on a ratio scale and were generated from a Euclidean distance metric. We cross-validated across individual cells of a dissimilarity matrix, whereas previous applications of cross-validation to MDS cross-validated over subjects (Steyvers, 2006). We believe our approach has certain advantages, e.g., it can be applied to single subject data, and might eventually be applicable in individual differences scaling, a direction we are now pursuing. This latter extension may be especially important, because averaging dissimilarity matrices might not always be warranted (Ashby, Maddox, & Lee, 1994). We are also currently exploring simulations and empirical applications when certain current constraints are relaxed, e.g., when similarity data are on a likert scale.

Acknowledgments

This work was funded by National Science Foundation grant SES-1626825, and the Alfred P. Sloan Foundation.

References

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5*(3), 144-151.

Gronau, Q.F., & Lee, M.D. (in press). Bayesian inference for multidimensional scaling representations with psychologically-interpretable metrics. *Computational Brain & Behavior*.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General, 142*(1), 256–281. https://doi.org/10.1037/a0028860

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45* (1), 149–166.

de Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software, 31*(3), 1-30.

Nosofsky (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43* , 22 –53.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42*(2), 241-266. https://doi.org/10.1007/bf02294052

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*. doi: 10.3758/s13428-019-01285-3

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*, 125–140.

Shepard, R. N. (2004). How a cognitive psychologist came to seek universal laws. *Psychonomic Bulletin & Review, 11*, 1-23. https://doi.org/10.3758/bf03206455

Steyvers, M. (2002). Multidimensional Scaling. In: *Encyclopedia of Cognitive Science*. Nature Publishing Group, London, UK.

Storms, G. (1995). On the robustness of maximum-likelihood scaling for violations of the error model. *Psychometrika, 60* (2), 247-258. https://doi.org/10.1007/BF02301415