Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 47 (2003) 32–46
Journal of Mathematical Psychology
http://www.elsevier.com/locate/jmp

Avoiding the dangers of averaging across subjects when using multidimensional scaling

Michael D. Lee[a,*] and Kenneth J. Pope[b]

[a] Department of Psychology, University of Adelaide, SA 5005, Australia
[b] School of Informatics and Engineering, Flinders University of South Australia, Australia

Received 28 April 2000; revised 28 March 2002

Abstract

Ashby, Maddox and Lee (Psychological Science, 5 (3) 144) argue that it can be inappropriate to fit multidimensional scaling (MDS) models to similarity or dissimilarity data that have been averaged across subjects. They demonstrate that the averaging process tends to make dissimilarity data more amenable to metric representations, and conduct a simulation study showing that noisy data generated using one distance metric, when averaged, may be better fit using a different distance metric. This paper argues that a Bayesian measure of MDS models has the potential to address these difficulties, because it takes into account data-fit, the number of dimensions used by an MDS representation, and the precision of the data. A method of analysis based on the Bayesian measure is demonstrated through two simulation studies with accompanying theoretical analysis. In the first study, it is shown that the Bayesian analysis rejects those MDS models showing better fit to averaged data using the incorrect distance metric, while accepting those that use the correct metric. In the second study, different groups of simulated 'subjects' are assumed to use different underlying configurations. In this case, the Bayesian analysis rejects MDS representations where a significant proportion of subjects use different configurations, or when their dissimilarity judgments contain significant amounts of noise. It is concluded that the Bayesian analysis provides a simple and principled means for systematically accepting and rejecting MDS models derived from averaged data.

1. Introduction

Multidimensional scaling (MDS) techniques (Shepard, 1962; Kruskal, 1964; see Cox & Cox, 1994 for a recent overview) generate representations of stimulus sets based on the similarities or dissimilarities between each pair of stimuli. Within MDS models, each stimulus is associated with a point in a coordinate space of some dimensionality, so that the distance between two points corresponds to the dissimilarity of the associated stimuli. MDS representations have their origins in, and some considerable status as, plausible models of human conceptual structure, particularly in relation to low-level, continuous sensory stimulus domains (Shepard, 1957, 1987, 1994). For this reason, MDS representations have been used as a tool for analyzing the psychological similarity structure of a variety of stimulus domains (e.g., Glushko, 1975; Jones, Roberts, & Holman, 1978; Heaps & Handel, 1999). In addition, they are used as the representational basis of a number of successful cognitive models, including the Generalized Context Model (Nosofsky, 1984, 1986), and ALCOVE (Kruschke, 1992).

A long-established and pervasive practice in both of these types of applications (e.g. Ekman, 1954; Gati & Tversky, 1982; Gregson, 1976; Johnson & Tversky, 1984; Kruschke, 1993) is to use 'pooled' similarity or dissimilarity matrices, obtained by averaging individual measures across a number of subjects. The rationale for averaging is the familiar one of reducing the effects of errors in whatever measurement process is used to collect the dissimilarity values. In a recent critique, however, Ashby, Maddox, and Lee (1994) highlight some undesirable consequences of this averaging process when fitting MDS models. They present theoretical arguments and a simulation study arguing that 'averaging across subjects changes the underlying psychological structure of the data' (p. 147). In particular, they

*Corresponding author. Fax: +61-8-8303-3770.
E-mail address: michael.lee@psychology.adelaide.edu.au (M.D. Lee).

0022-2496/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-2496(02)00019-6

show that artificial data generated using a particular metric structure, when averaged, may be fit better by an MDS model that uses a different distance metric. They conclude that 'it would be extremely dangerous to fit an MDS model to data that have been averaged across a large group of subjects and then to claim that the resulting MDS solution is a valid representation of the underlying psychological space' (p. 147).

The first goal of this paper is to demonstrate that the application of a Bayesian analysis for selecting MDS models, described by Lee (2001), addresses the problems highlighted by Ashby et al. (1994). Unlike the approach adopted by Ashby et al. (1994), which relies solely on data-fit in comparing competing MDS models, the Bayesian analysis is sensitive to data-fit, model complexity, and the inherent precision of the dissimilarity data. This means that it is possible to compare a candidate MDS representational model with a 'null' or 'zero-dimensional' MDS model, explaining 0% of the variance of the data using one parameter. When coupled with the sensitivity to the precision of dissimilarity data, simulations show that making these comparisons causes MDS models with the incorrect metric structure to be rejected on a systematic basis.

The second goal of this paper is to demonstrate the applicability of the Bayesian analysis to a more general situation, where different groups of subjects use fundamentally different underlying representations to generate dissimilarity data. By varying the proportion of subjects who favor one representation over another, as well as manipulating the level of noise in the measurement process for collecting data, simulations show that the Bayesian analysis rejects MDS models that do not capture the structure of the dominant original representation. Taken together, these two simulation studies suggest that the application of the Bayesian analysis has the potential to avoid the dangers inherent in averaging data across subjects when fitting MDS representational models.

2. Summary of Ashby, Maddox, and Lee (1994)

In their analysis, Ashby et al. (1994) concentrate on the so-called 'metric' variety of MDS, which involves the application of some form of optimization method to minimize an error measure of the form

$$E \propto \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2, \tag{1}$$

where d_{ij} is the given dissimilarity between the ith and jth stimuli, and $\hat{d}_{ij}$ is the distance between representative points $\mathbf{p}_i = (p_{i1}, \ldots, p_{im})$ and $\mathbf{p}_j = (p_{j1}, \ldots, p_{jm})$ in an m-dimensional space. Following common practice in metric MDS (see, for example Cox & Cox, 1994), this distance is measured according to one of the family of Minkowskian r-metrics, given by

$$\hat{d}_{ij} = \left[\sum_{k=1}^{m} |p_{ik} - p_{jk}|^r \right]^{\frac{1}{r}} + c, \tag{2}$$

where c is an additive constant.

On the theoretical front, Ashby et al. (1994) use the framework developed by Furnas (1989) to demonstrate that averaging across subjects will tend to eliminate any violations of the triangle inequality that might be evident in single-subject dissimilarity data using $r<1$. In particular, they simulate dissimilarity data for 50 subjects using a known spatial configuration, which consists of a complete 3×3 grid with unit spacing, rotated by 30° anticlockwise about a point near the center of the grid (Ashby, pers. comm., July, 1999).[1] Individual differences between subjects are simulated by adding zero-mean Gaussian noise with a specified standard deviation to each coordinate independently for each subject. Fig. 1 shows the underlying grid configuration, and overlays examples of noisy individual subject configurations derived using standard deviations of 0.1, 0.2, 0.3 and 0.4. The distances between these noise perturbed points, according to the Minkowskian r-metric with $r = 0.5$, are used as dissimilarity measures for each subject, and an averaged dissimilarity matrix is also calculated across subjects. Ashby et al. (1994) note that the individual data, because it was generated using a 'semi-metric' distance measure, contains many violations of the triangle inequality, but that the averaged data contains no such violations. Accordingly, they observe that 'although these data were generated from a model that is incompatible with all versions of MDS, there exists some MDS model that fits the averaged data perfectly' (Ashby et al., 1994, p. 147).[2]

In their simulation study Ashby et al. (1994) use the same configuration, and generate dissimilarity matrices for 50 subjects, and an averaged matrix, in the same way, but do this for each of three distance metrics, corresponding to the choices $r = 0.5$, 1 and 2 from the Minkowskian family. In the broad context of cognitive modeling, these values are of particular interest. The $r = 1$ (City-Block) and $r = 2$ (Euclidean) cases have long been associated with, respectively, so-called 'separable' and 'integral' stimulus domains (Garner, 1974; Shepard, 1991). Meanwhile, the adoption of metrics with $r<1$ has been given a psychological justification (Gati & Tversky, 1982; see also Shepard, 1987, 1991) in terms of modeling

[1] The exact location of the center of rotation, although difficult to determine from Ashby et al. (1994, Fig. 1), does not matter, since it does not affect the required inter-point distances under any of the metrics being considered.

[2] As this statement implies, Ashby et al. (1994) do not refer to the $r = 0.5$ 'semi-metric' as an MDS model. A difference in terminology between our study and theirs is that we do refer to the $r = 0.5$ case as an MDS model.

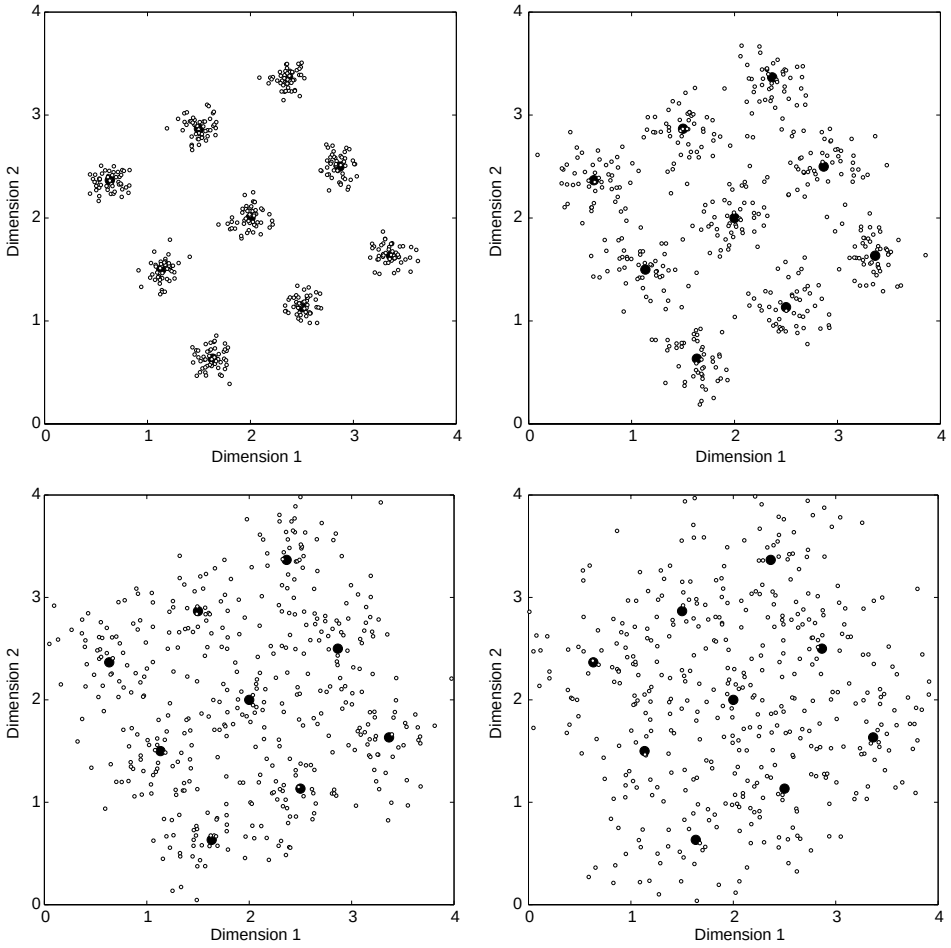

Fig. 1. The configuration used by Ashby et al. (1994). The large circles indicate the underlying 3×3 grid, while the small circles indicate the stimulus position for 50 simulated subjects using noise standard deviations of 0.1 (top left), 0.2 (top right), 0.3 (bottom left) and 0.4 (bottom right).

stimuli with component dimensions that 'compete' for attention. Ashby et al. (1994, Table 1) examine the patterns of data-fit when MDS models with the same range of metric structures, $r = 0.5$, 1 and 2, are fitted to both the single-subject and averaged data for all three generating metrics. As expected, for the single-subject data, perfect data-fit is achieved when the same metric is used to generate the data and recover the MDS representation. The important finding, however, is that for the averaged data, the best data-fit is not necessarily achieved using the recovery metric that corresponds to the generating metric. In particular, for the averaged data generated using $r = 0.5$, the MDS representation using $r = 1$ fit the data significantly better that the representation using $r = 0.5$.

Taken together, the theoretical arguments and these simulation results suggest that it can be inappropriate to average data across subjects when fitting MDS models. The averaging process may have effects that extend beyond the reduction of measurement error, and serve to alter fundamental metric properties of the data. Indeed, the simulation study seems to realize Ashby et al.'s (1994) fears that 'the worst possible effect of averaging would be to alter the underlying psychological structure of the data in such a way that an invalid model appears valid' (p. 144).

Ashby et al. (1994), through their focus on critiquing existing practice in MDS modeling, test the validity of their models using a measure of data-fit. A more sophisticated approach to model selection, recently popularized in psychological modeling (e.g., Myung & Pitt, 1997; Myung, Forster, & Browne, 2000b; Myung, Balasubramanian, & Pitt, 2000a; Pitt, Myung, & Zhang, 2002), argues that both data-fit and model complexity should be taken into account when comparing models. Once this trade-off between accuracy and simplicity is established in the process of model selection, the notion of data precision also becomes important. If it is proposed that a model should be made more complicated to provide some improvement in data-fit, it is necessary to have some understanding of the inherent precision of the constraining data itself. Precise data will tend to warrant the improved data-fit offered by additional complexity, whereas noisy data will tend not to warrant any such change. Since the Bayesian approach to MDS model selection developed by Lee

(2001) is founded on this trade-off between data-fit and model complexity, and explicitly incorporates a measure of data precision, it is worth applying to the problem highlighted by Ashby et al. (1994). Our approach to doing this is to compare MDS representations with a 'null' zero-dimensional MDS model, which has poor data-fit but low complexity, to gauge whether data warrant any MDS model at all.

3. A Bayesian approach to MDS model selection

The method for MDS model selection developed by Lee (2001) is based on the well-known Bayesian Information Criterion (BIC: Schwarz, 1978; see also Kass & Raftery, 1995; Myung & Pitt, 1997). The BIC takes the general form

$$\mathrm{BIC} = -2\log p(\mathrm{ML}) + P\log N, \tag{3}$$

where $p(\mathrm{ML})$ is the likelihood of the data given the model when the model parameters are estimated using maximum likelihood, P is the number of parameters in the model, and N is the sample size. Qualitatively, it can be seen that this measure increases whenever either model complexity, as measured by the number of model parameters, increases or when the data-fit of the model worsens. Accordingly, the candidate model with the minimal BIC value is to be preferred.

The application of the BIC to MDS uses a probabilistic formulation of the data-fit, based on the assumption that the probability of a set of target dissimilarities, given a particular MDS representation of dimensionality m, has a Gaussian distribution with common variance σ_d^2,

$$p(\mathbf{D}\mid m,\hat{\mathbf{D}}) \propto \exp\left(-\frac{1}{2\sigma_d^2}\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2\right), \tag{4}$$

where $\mathbf{D}=[d_{ij}]$ and $\hat{\mathbf{D}}=[\hat{d}_{ij}]$.

In the context of MDS models, parametric complexity is related to the number of dimensions used by a representation. In an m-dimensional space, representing n points uses mn parameters. However, the representation is translation invariant, reducing the number of free parameters by m. The additive constant c constitutes another free parameter, giving a total of $(m(n-1)+1)$. Since the representation is generated from an $n\times n$ symmetric dissimilarity matrix, these parameters are constrained by a total of $n(n-1)/2$ data values. Accordingly, the MDS formulation of the BIC measure takes the form

$$\mathrm{BIC} = \frac{1}{s_d^2}\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2 + (m(n-1)+1)\log\left(\frac{n(n-1)}{2}\right), \tag{5}$$

where s_d is a sample estimate of the data precision population parameter σ_d.

Ashby et al. (1994) consider the experimental methodology in which there are individual dissimilarity matrices $\mathbf{D}^k=[d_{ij}^k]$ describing the data collected from each of $k=1,2,\ldots,K$ subjects, and it is the averaged dissimilarity matrix $\mathbf{D}=\frac{1}{K}[\sum_k d_{ij}^k]=[d_{ij}]$ that is used to generate an MDS representation. In this case, as argued by Lee (2001), one approach to determining s_d is to calculate the average of the sample standard deviations for each of the pooled cells in the final averaged matrix, as follows:

$$s_d = \frac{1}{n(n-1)/2}\sum_{i<j}\sqrt{\frac{\sum_k (d_{ij}^k-d_{ij})^2}{K-1}}. \tag{6}$$

This estimate of data precision is entirely determined by the raw data, and may be calculated before fitting MDS representations with different dimensionalities or metric structures to the averaged data. The evaluation of BIC measures for each of these candidate representations is then straightforward, requiring the substitution of s_d into Eq. (5), and using the known parametric complexities and residual errors. The representation with the minimal BIC value may then be taken as constituting an appropriate compromise between the need to accommodate the original data, and the requirement to minimize the parametric complexity of the MDS model.

As a special case of this comparison of BIC values, we consider the 'null' MDS model to gauge whether an MDS model is warranted at all. The 'null' model may be thought of as a degenerate 'zero-dimensional' model that explains 0% of the variance in the data using only one parameter. To see this, note that the variance accounted for by an MDS model is given by

$$v = 1-\frac{\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2}{\sum_{i<j}(d_{ij}-\bar{d})^2}, \tag{7}$$

where $\bar{d}$ is the arithmetic mean of the dissimilarity measures. This means that setting the additive constant in the distance measure to the mean of the dissimilarity data, but not specifying any coordinate locations for representative points, effectively creates a one-parameter MDS model explaining 0% of the variance. If this model has a BIC value that is less than those for substantive MDS representations, we can conclude that it is not appropriate to model the data using MDS. This may occur when MDS representations need a large number of dimensions to achieve relatively poor data-fit, signaling the presence of fundamentally non-metric data, or when the inherent precision of the data itself is too poor to sustain a meaningful representational model.

4. Simulation study 1: same underlying configuration

The application of the BIC to MDS model selection was examined in a simulation study based on that reported by Ashby et al. (1994).

4.1. Method

The same nine-point two-dimensional spatial configuration was used to generate dissimilarity matrices for 50 subjects and an averaged matrix, for the three generating metric choices $r = 0.5, 1$ and 2. The Ashby et al. (1994) study considered only two levels of data precision, corresponding to two levels of standard deviation for the Gaussian noise added to coordinate locations, given by $\sigma_p = 0.50$ and 0.67. In the present study, separate simulations were undertaken for σ_p values ranging from 0 to 1, increasing in steps of 0.02. Once dissimilarity matrices were calculated for each subject, the sample estimate s_d of the population parameter σ_d was calculated from the 'raw data' according to Eq. (6).

As with the Ashby et al. (1994) study, the averaged matrix for each level of data precision was recovered using the same three distance metrics $r = 0.5$, 1 and 2. Rather than simply assuming a two-dimensional MDS representation, however, each metric structure was applied to MDS representations with 1, 2 and 3 dimensions. The metric MDS algorithm employed for model fitting was based on the standard Levenberg–Marquardt approach to non-linear least-squares optimization (More, 1977), and used 200 attempts at fitting each model, starting at a different random configuration in an attempt to avoid local minima. When the correct metric structure was used and σ_p was near zero, the recovered MDS models explained almost all of the variance in the data, suggesting that the algorithm was adequate. On this basis, BIC values for each of the best-fitting representations of each dimensionality, together with that for the 'null' zero-dimensional alternative, were calculated using Eq. (5), and the preferred model with lowest value was determined.

4.2. Results

Figs. 2–4 summarize the results of these simulations when the generating metrics were set to $r = 0.5$, $r = 1$, and $r = 2$, respectively. The x-axis shows the data precision, as quantified by σ_p, while the y-axis shows the data-fit for the recovery metric choices $r = 0.5$, 1 and 2, as quantified by the percentage of variance accounted for (PVAF) by the best-fitting MDS representation. Each marker, showing the data-fit of a particular recovery metric at a particular precision level, is colored according to the BIC comparisons. Black markers indicate that a two-dimensional MDS model was preferred, while white markers indicate that another dimensionality, or the 'null' zero-dimensional alternative was preferred. In this way, the summary figures allow the data-fit approach to model selection used by Ashby et al. (1994) to be compared directly with those arising from the Bayesian analysis.

In each of Figs. 2, 3 and 4, it can be seen that, for small values of σ_p, the best-fitting configuration is consistently found when the recovery metric matches the generating metric. These configurations also display near perfect data-fit, and are regarded by the BIC measure as the 'appropriate' spatial representation of the data. It is interesting to examine the point at which Euclidean representations recovered from Euclidean data are rejected, with reference to the different noise standard deviations shown in Fig. 1. As Fig. 4 shows, once σ_p exceeds about 0.28, derived representations explaining almost all of the variance in the data are rejected by the BIC. This means, in terms of the natural visual interpretation of 'straight-line' distance in Fig. 1, that when $\sigma_p = 0.1$ and 0.2 the averaged data would be regarded as sufficiently precise to sustain an MDS representation, but not when $\sigma_p = 0.3$ and 0.4, with the 'cut-off' point being somewhere near the case $\sigma_p = 0.3$.

The important result, however, in terms of the difficulties noted by Ashby et al. (1994), concerns the patterns of $r = 0.5$ and 1 recovery from the generating metric $r = 0.5$. Fig. 2 shows that, when σ_p exceeds 0.24, the configurations recovered using $r = 1$ fit the data better than those using the correct metric $r = 0.5$. Effectively, this result replicates those on which the conclusions of Ashby et al. (1994) are based. What Fig. 2 also shows, however, is that the BIC begins rejecting derived configurations,[3] whatever the distance metric employed, at about the level of precision where the change in ordering of data-fit occurs. Only when $\sigma_p = 0.26$ does the BIC incorrectly accept an $r = 1$ representation with better data-fit than the $r = 0.5$ representation. In other words, the evaluation of the MDS models using the BIC tends to provide a basis for rejecting those representations that, because of the effects of the averaging process, show better fit when assuming the incorrect metric structure.

4.3. Theoretical discussion

There are two significant features shared by Figs. 2–4. First, the data-fit of representations using the correct metric deteriorates as σ_p increases. Secondly, the data-fit of representations using the recovery metric $r = 2$ improves as σ_p increases. This means that, as the level of noise added to the underlying representations is increased, the metric that best fits the averaged distance

[3] Although it is not depicted in Fig. 2, this rejection is in favor of the 'null' model.

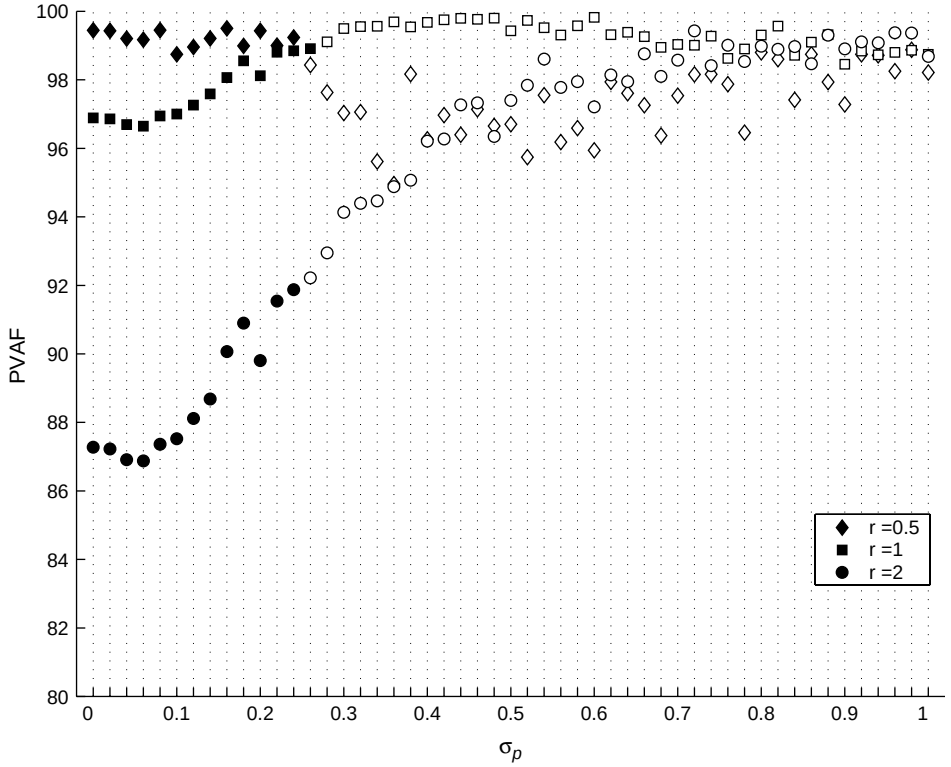

Fig. 2. PVAF as a function of σ_p for the best-fitting two-dimensional MDS representations, using each of the recovery metrics $r = 0.5, 1$ and 2. The generating metric for the dissimilarity data uses $r = 0.5$. Representations accepted by the BIC are indicated by black markers, while rejected representations are indicated by white markers.

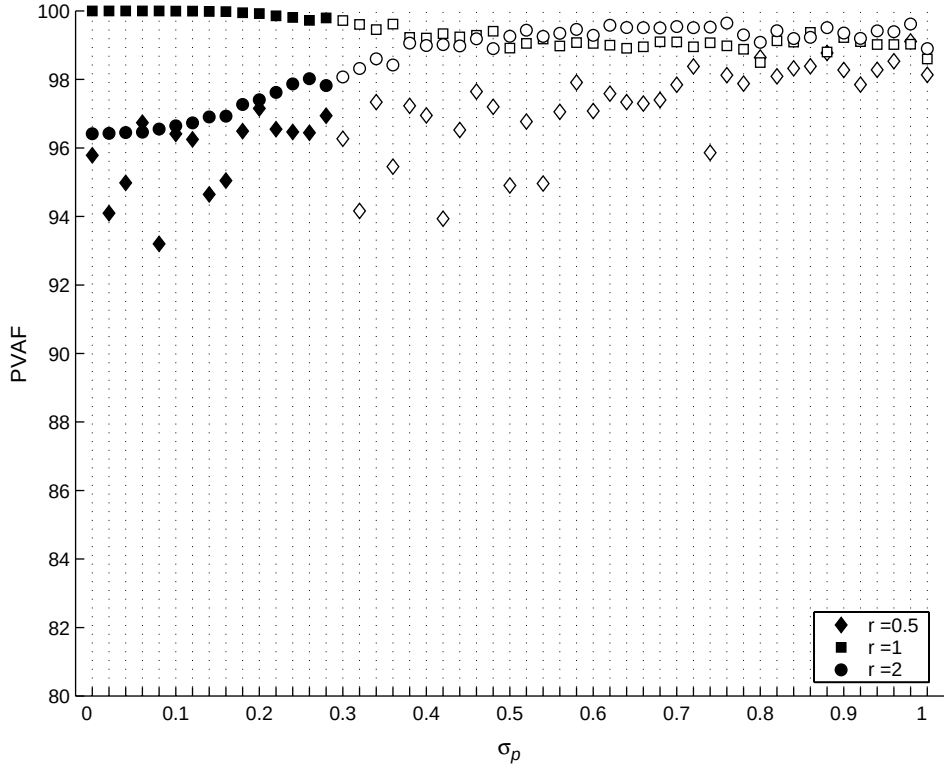

Fig. 3. PVAF as a function of σ_p for the best-fitting two-dimensional MDS representations, using each of the recovery metrics $r = 0.5, 1$ and 2. The generating metric for the dissimilarity data uses $r = 1$. Representations accepted by the BIC are indicated by black markers, while rejected representations are indicated by white markers.

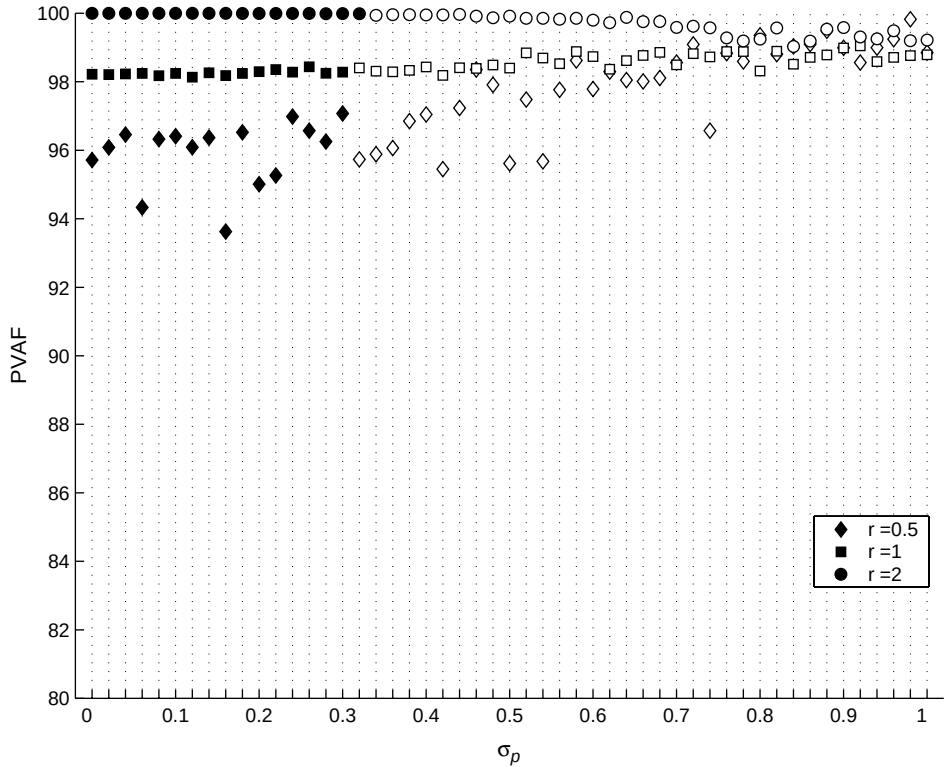

Fig. 4. PVAF as a function of σ_p for the best-fitting two-dimensional MDS representations, using each of the recovery metrics $r = 0.5, 1$ and 2. The generating metric for the dissimilarity data uses $r = 2$. Representations accepted by the BIC are indicated by black markers, while rejected representations are indicated by white markers.

matrix tends towards a larger value of r. In particular, Fig. 2 shows the best-fitting metric to be $r = 0.5$ at no and low noise, before moving through the $r = 1$ and then 2 metrics as σ_p increases. Similarly, Fig. 3 shows the best-fitting metric change from $r = 1$ to 2 as σ_p increases.

The reason for this change in the best recovery metrics can be explained in terms of the different effects adding noise has on distances under different metrics. It is simplest to think of the problem of recovering the unit square. Across different metrics, this four-point configuration is characterized by the relationship between the horizontal (or vertical) distance between neighboring points, and the diagonal distance across the square. For the $r = 2$ metric, these distances are 1 for the horizontal and $\sqrt{2}$ for the diagonal, for $r = 1$ they are 1 and 2, and for $r = 0.5$ they are 1 and 4. In general, the relationship between horizontal and diagonal distances uniquely identifies the underlying Minkowskian metric.

When Gaussian noise is added independently to the points on a number of unit squares, and the individual distances averaged, the final horizontal and diagonal distances change depending on the metric being used, and the level of noise added. Fig. 5 provides a graphical means of understanding these changes, showing the unit square for each of the $r = 0.5, 1$ and 2 metrics with both small and large amounts of added noise. The unit square is shown by the points, each of which is surrounded by a broken line giving the circular contour of equal likelihood once noise is added. The contours of equal distance for the original horizontal and diagonal distances are also shown by the solid lines.

Consider first the left-hand column of Fig. 5, where a small amount of noise has been added. In the $r = 0.5$ case, the horizontal distance will increase on average, because most of the distances between random points on the noise contours are longer than the original distance indicated by the metric contour. For the diagonal point, however, the majority of the noise contour falls inside the metric contour, and so the average diagonal distance will decrease. Using the same geometric argument, it can be seen that horizontal distances will increase and diagonal ones will remain the same for $r = 1$, and both horizontal and diagonal distances will increase for $r = 2$.

For the larger level of noise shown in the right-hand column of Fig. 5, a different pattern of change emerges. This is because the noise contours now begin to overlap, increasing the distances between points on the overlapping segments that had previously been decreasing, and so increasing the average distance between all points on the noise contours. This means that, even in cases such as the diagonal distance under $r = 0.5$ where the metric acts to decrease average distance, the addition of

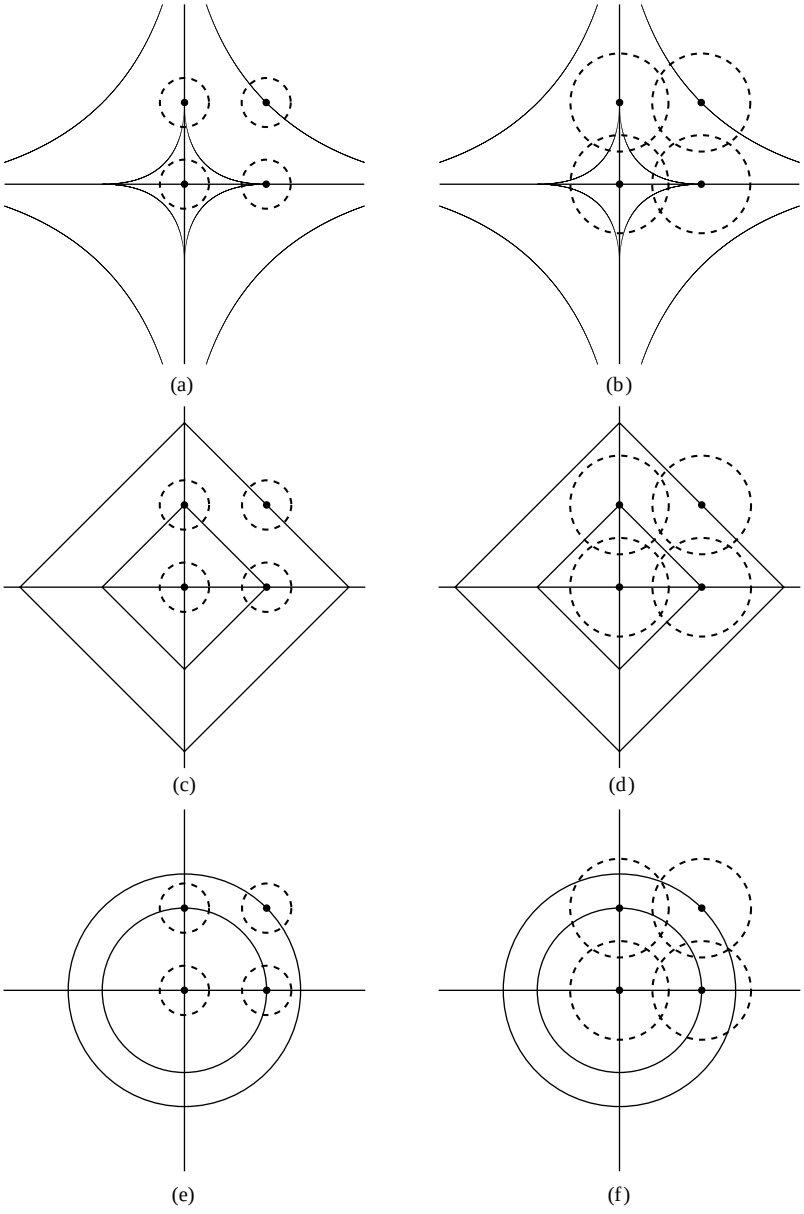

Fig. 5. Geometrical interpretation of the effect adding noise to the unit square has on horizontal, vertical and diagonal distances, shown for six cases: (a) $r = 0.5$ with small noise, (b) $r = 0.5$ with large noise, (c) $r = 1$ with small noise, (d) $r = 1$ with large noise, (e) $r = 2$ with small noise, (f) $r = 2$ with large noise. Contours of equal noise are shown by the broken circles, while contours of equal distance are shown by the solid lines. See text for discussion.

large enough noise will eventually lead to an increase. The other important effect is that the overlap between noise contours will always occur first for the horizontal distances, because they are closer for all metrics with $r < \infty$. This means that the addition of noise will eventually lead to horizontal distances increasing more rapidly than diagonal distances, regardless of the metric.

Fig. 6 confirms these geometric insights by showing the results of a simulation measuring the increase for both horizontal and diagonal distances under each of the $r = 0.5, 1$ and 2 metrics. This simulation is based on averaging across 1,000,000 noise perturbed unit squares under each metric at each level of noise.[4] It shows that all of the horizontal distances are 1 when there is no noise, but that there is a rapid increase as noise is added for the $r = 0.5$ metric, a less rapid increase for $r = 1$, and an even gentler increase for $r = 2$. For the diagonal distances, there is an increase after an initial decline for $r = 0.5$, an increase after initial stability for $r = 1$, and a more immediate increase

[4] In the appendix, we provide an analytic derivation of the increase for the $r = 1$ case.

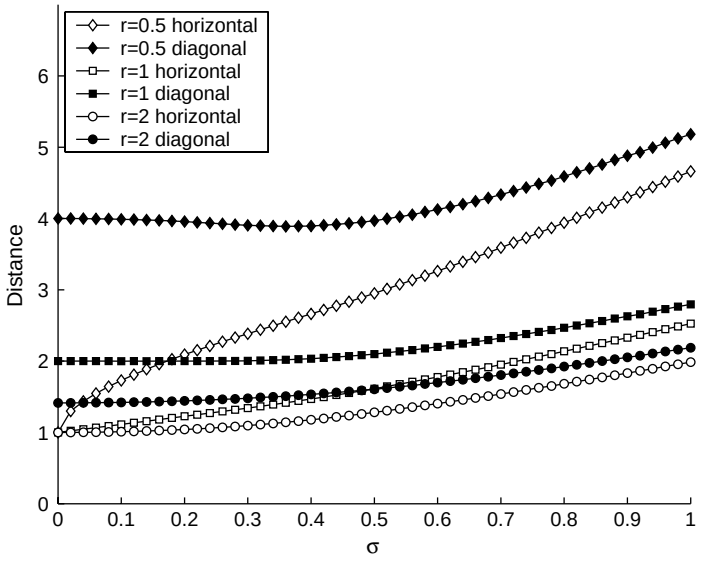

Fig. 6. The increase for both horizontal and diagonal distances under each of the $r = 0.5, 1$ and 2 metrics.

for $r = 2$. Most importantly Fig. 6 also shows that, for all three metrics, the horizontal distances increase more rapidly than the diagonal distances as more noise is added.

It is the difference in the way the horizontal and diagonal distances increase under noise that explains the change in the best recovery metric. This is because the relationship between the two distances, which defines the metric of the unit square, has been changed. Suppose we tried to recover the unit square configuration using the $r = 0.5$ metric when noise with $\sigma_p = 0.2$ has been added. As Fig. 6 indicates, the new horizontal length between neighboring points would have increased from 1 to about 2, while the diagonal distance has not yet increased much beyond its original value of 4. These distances make it impossible to recover the original unit square configuration using the $r = 0.5$ metric. Indeed, the pattern of distances between all of the points does not correspond exactly to any two-dimensional representation under this metric, and so MDS will return a representation that has a significant level of error. Using the $r = 1$ metric, however, the unit square would be recovered without error, because the diagonal distance is twice the horizontal. In this way, adding noise to a representation using the $r = 0.5$ metric has led to accurate recovery being achieved in the $r = 1$ metric.

In general, it is possible to find the best recovery metric for any given averaged horizontal and diagonal distance values. If the horizontal distance is d_h and the diagonal distance is d_d, the best metric r^* is the one that satisfies the relationship

$$d_d = (d_h^{r^*} + d_h^{r^*})^{\frac{1}{r^*}},$$

which gives

$$r^* = \frac{1}{\log_2 d_d - \log_2 d_h}. \tag{8}$$

Fig. 7 shows the results of applying Eq. (8) to the distances in Fig. 6. This means that Fig. 7 shows the best recovery metric for the unit square, as a function of the noise level σ_p, for each of the generating metrics $r = 0.5$, 1 and 2. It can be seen that the best recovery metric increases as more noise is added, which is entirely consistent with the results of Ashby et al. (1994), and the replications and extensions shown in Figs. 2–4. It is interesting to note that the best-fitting metric continues to increase for all generating metrics, heading towards the dominance or supremum metric ($r = \infty$) at the limit of arbitrarily large levels of noise. This makes sense: Horizontal and vertical distances increase more than diagonal ones, and eventually become so much larger that the dominance metric is appropriate.

For the problem raised by Ashby et al. (1994), the important point is that the simulations and theoretical analysis show the averaging of distances formed from noisy approximations to the same representation results in a distance matrix that does not accurately reflect the metric of the true representation. As the level of noise increases, Minkowskian r-metrics with larger values of r than used originally will provide the best fit. This means that model selection based on PVAF is guaranteed to select the incorrect model. In contrast, the BIC is sensitive to the presence of noise, and so at least has the potential to avoid selecting incorrect models. What the results in Figs. 2–4 suggest is that the rejection occurs at or near the level of noise where incorrect metrics start to achieve the best fit, although a full theoretical account of the generality of this finding awaits future research.

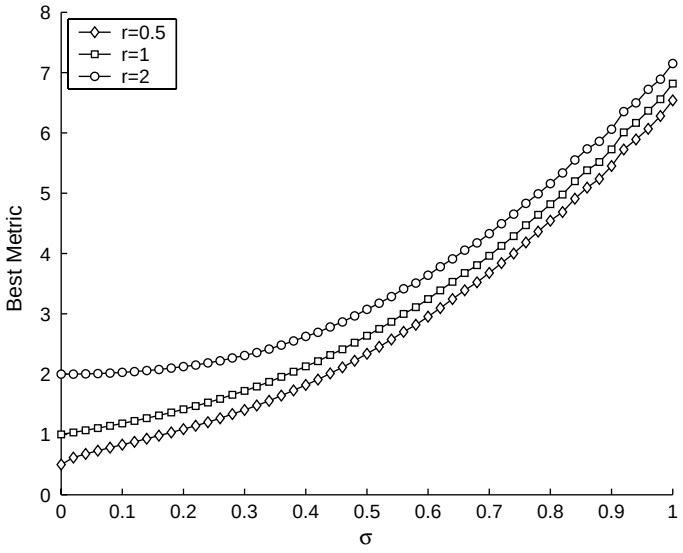

Fig. 7. The best fitting metric r^* for each of the generating metrics $r = 0.5, 1$ and 2 as σ_p increases from 0 to 1.

5. Simulation study 2: different underlying configurations

Given the success of the BIC in accepting and rejecting representations when averaging noisy data derived from the same underlying configuration, the obvious extension is to examine its capabilities when there are different underlying configurations. In many respects, this is the scenario of the greatest relevance to the methodological issue of averaging similarity or dissimilarity data across subjects. If, for example, different groups of subjects use significantly different underlying spatial mental representations when making judgments, the use of averaging is not justified. This is because there are now effectively two sources of data variation: that caused by noise in the measurement process, as in the previous simulation study, and that caused by fundamental differences in the underlying representations. While averaging may legitimately be employed to reduce the effects of the decision noise (i.e., if the above BIC model selection procedure is used), it should not be allowed to distort the underlying representations that MDS seeks to recover.

5.1. Method

To provide a concrete test of averaging multiple representations, two spatial configurations resembling the letters 'A' and 'I' were constructed, each containing 14 stimuli. The configurations are shown in Fig. 8. Single-subject dissimilarity data was generated from these configurations by measuring the distances between points using both the $r = 1$ and $r = 2$ metrics. A total of 100 subjects were considered, with the proportion of subjects, β, using the 'A' configuration varying from 0% up to 100% in 10% intervals. Along similar lines to the previous simulation study, zero-mean Gaussian noise with a standard deviation ranging from $\sigma_p = 0$ to 0.50 in steps of 0.02 was added independently to the coordinate locations.

For each level of decision noise σ_p and proportion of subjects β using the 'A' configuration, the same MDS algorithm developed for the first simulation study was used to find best-fitting representations in spaces of 1, 2 and 3 dimensions.[5] For each averaged dissimilarity matrix, precision estimates s_d were calculated as before, and these were used to find BIC measures for each derived representation. The acceptance or rejection of the best-fitting two-dimensional configuration was then determined on the basis of which BIC measure, including that for the zero-dimensional 'null' alternative, was minimal.

5.2. Results

Figs. 9 and 10 summarize the results of this process for the $r = 1$ and $r = 2$ cases, respectively. Each of these figures displays a grid of the best-fitting two-dimensional configurations as σ_p increases from left to right, and as the proportion of subjects β using the 'A' configuration increases from top to bottom. Those representations that are accepted by the BIC analysis are indicated by the presence of a bold border around the derived representation. Visual interpretation of the effectiveness of the BIC in accepting or rejecting the MDS configurations should be made in the context of an understanding of the distance-preserving

[5] Since the subject proportion β has effectively replaced the recovery metric as a dependent variable, the MDS representations assumed the same distance metric used to generate the data.

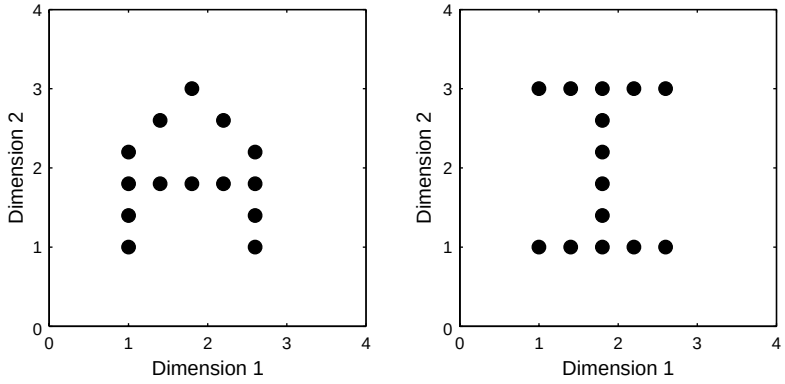

Fig. 8. The 'A' and 'I' representational configurations.

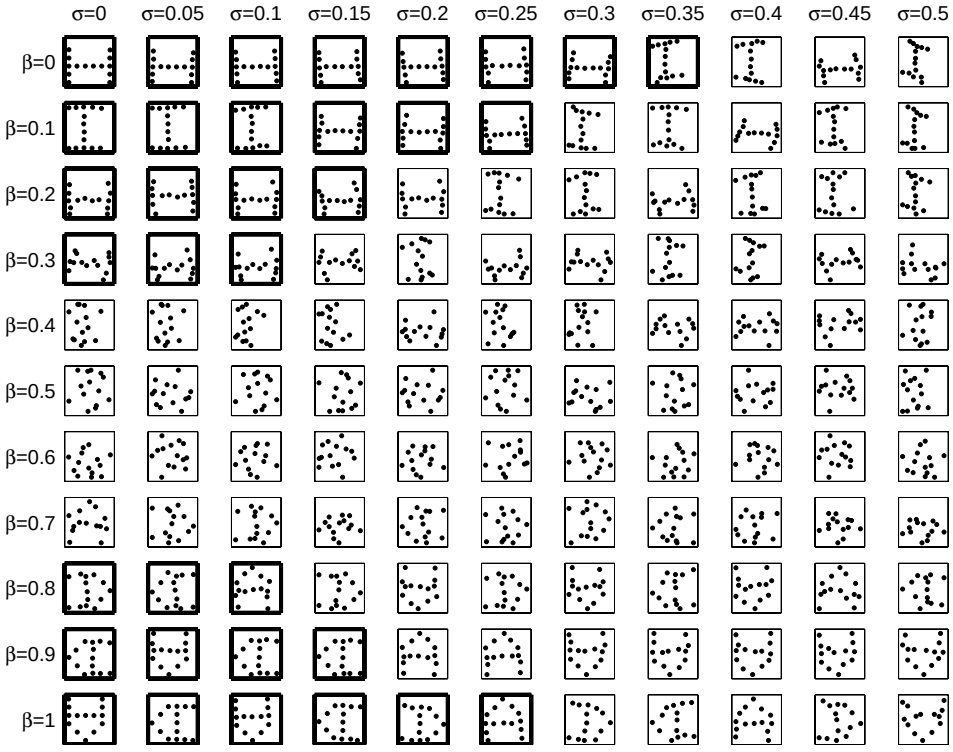

Fig. 9. Recovered City-Block configurations as the proportion of subjects with different underlying representations changes from $\beta = 0, 0.1, \ldots, 1$, and the Gaussian noise takes different standard deviations $\sigma_p = 0, 0.05, \ldots, 0.50$. The acceptance of a recovered configuration by the BIC is indicated by a solid border.

transformations afforded by the $r = 1$ and 2 metrics. Given the symmetries of the original configurations shown in Fig. 8, this means that any rotation is permissible in the $r = 2$ case, while any rotation by a multiple of 90° is permissible in the $r = 1$ case.

With these guidelines in mind, Figs. 9 and 10 suggest that the BIC correctly accepts representations when most subjects are using one of the configurations, and there is little decision noise. In this situation, MDS clearly recovers representations that capture the structure of the dominant original configuration. As subgroups of subjects start to use both configurations, however, the BIC only accepts representations generated from data that have been subjected to relatively less decision noise. Eventually, when significant numbers of subjects are using both configurations, the derived representations become uninterpretable, and are rejected by the BIC even when there is no decision noise. These results demonstrate the usefulness of the BIC in avoiding the dangers inherent in averaging data across subjects, since they indicate that the BIC analysis provides a systematic mechanism for rejecting representations derived from data where the underlying configurations differ significantly across subjects.

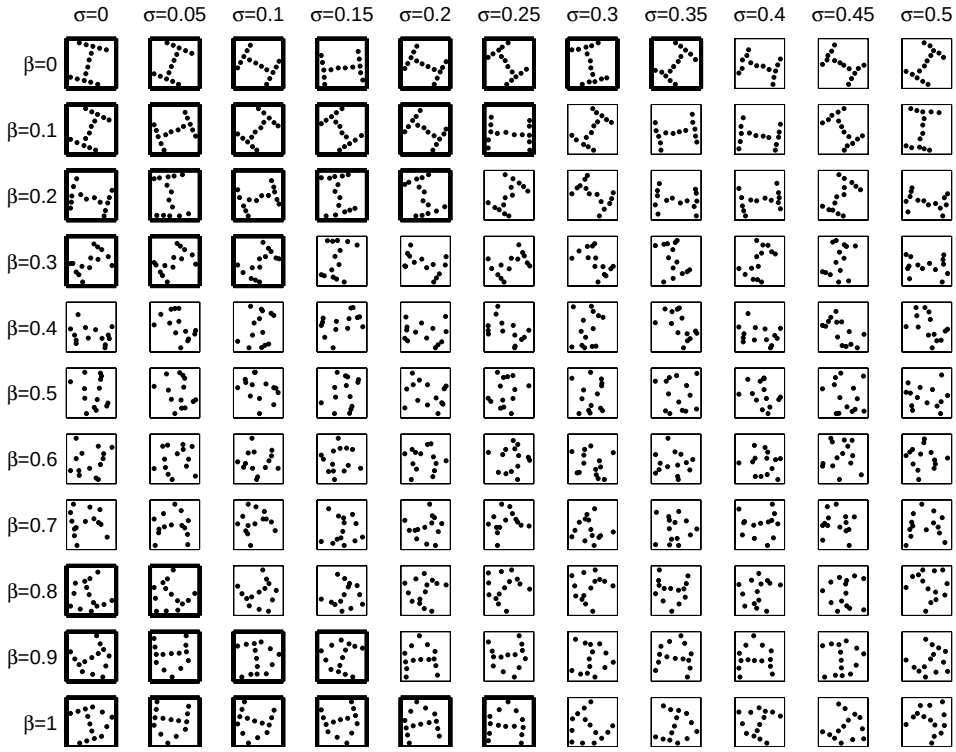

Fig. 10. Recovered Euclidean configurations as the proportion of subjects with different underlying representations changes from $\beta = 0, 0.1, \ldots, 1$, and the Gaussian noise takes different standard deviations $\sigma_p = 0, 0.05, \ldots, 0.50$. The acceptance of a recovered configuration by the BIC is indicated by a solid border.

5.3. Theoretical discussion

The results of this second simulation study rely on the ability of the BIC to accept MDS models based on averaged data when there is only one underlying spatial representation, but reject models based on averaged data across different configurations. An intuitive analysis of the BIC considering the properties of its data-fit term

$$\frac{1}{s_d^2} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2,$$

and its parametric complexity term

$$(m(n-1)+1) \log\left(\frac{n(n-1)}{2}\right),$$

shows how it is able to distinguish these two cases.

When there is only one underlying representation that is subject to decision noise, the averaging process should allow MDS to recover a configuration with a relatively low error measure. This is because a dissimilarity matrix that has been generated by averaging across matrices generated from the same configuration, but each corrupted by zero-mean additive Gaussian noise, will approximate the noise-free dissimilarity matrix, providing that the variance of the noise is not too large relative to the number of matrices being averaged. Of course, if the noise is too large, the metric problems raised by Ashby et al. (1994) will arise. For low levels of noise, however, it is possible to fit an accurate MDS model to the averaged data, which leads to a small value for the data-fit term. When there is no single underlying representation, however, the data-fit term will be relatively greater, because the averaged data will not be consistent with any of the generating configurations, and is unlikely to be consistent with any other MDS representation of reasonable dimensionality.

The parametric complexity term, in contrast, is not affected by whether or not there is a single underlying configuration. It simply serves to measure whether the achieved level of data-fit provides sufficient evidence to warrant the number of dimensions used by the representation. BIC values, therefore, will be lower when there is a single underlying representation than where there are multiple representations, even at the same level of data precision. This means that, when a comparison is made with the BIC value of the 'null' model, it is more likely that MDS representations based on subjects using the same underlying configuration will be accepted.

The important point, therefore, is that the BIC does not simply accept and reject representations based on the measured level of data precision. Through its combination of residual error and data precision, the BIC is sensitive to different sources of imprecision in averaged data. In particular, it naturally distinguishes

between decision noise acting on a single representation, and variance in data that arises from the existence of two or more fundamentally different representations.

Perhaps the least impressive aspect of the performance of the BIC in the second simulation study is that, when there is a large degree of decision noise, but the subjects remain relatively homogeneous in terms of their underlying representations, representations are rejected that seem to preserve the structure of the original configurations. This pattern of rejection highlights what is sometimes termed the 'conservative' nature of the BIC measure (Raftery, 1999). Basically, the BIC is conservative in the sense that it tends to provide relatively less evidence for additional parameters than the representations themselves suggest. In its practical application to selecting MDS models, as is evident in Figs. 9 and 10, this causes the BIC to have a tendency to reject what could reasonably be regarded as meaningful representations.

There are strong grounds, however, for arguing that the conservatism of the BIC biases it towards making the type of error that constitutes the lesser of two evils. As Grünwald (2000, p. 148) concludes: 'If you overfit, you think you know more than you really know. If you underfit, you do not know much, but you know that you do not know much. In this sense, underfitting is relatively harmless, while overfitting is dangerous'. Figs. 9 and 10 suggest that the application of the BIC is likely to accept most useful MDS representations, and will reject all of those that have been significantly altered by differences in underlying configurations or noise in the measurement process. The rejection of some representations at the margins of interpretability might be regarded as a reasonable sacrifice for this capability.

6. General discussion

Ashby et al. (1994) raise a fundamental issue for MDS models, by questioning whether it is appropriate to use averaged data. Taken together, the two simulation studies presented suggest that the BIC provides part of the answer to this question. The first study showed the effectiveness of the BIC analysis in the case where subjects use the same underlying representation, but display individual differences because of noisy dissimilarity judgments. Under these conditions, the BIC analysis is able to reject MDS models that showed a better fit to averaged data when using the incorrect distance metric, while accepting those models that maintained the correct metric. The second study showed the effectiveness of the BIC analysis in the case where different subjects have fundamentally different underlying configurations, and their dissimilarity judgments remain noisy. Here, the BIC analysis is capable of accepting most of the substantial models, where most subjects use the same configuration and give reasonably precise judgments, while rejecting uninterpretable MDS models based on data that has been transformed by the averaging process.

It would be wrong to conclude that the BIC analysis presented here casts doubt on the existence of the problems raised by Ashby et al. (1994). It is true that averaging dissimilarity data across subjects can change the structure of the data in ways that make it inappropriate to fit MDS models. What the BIC analysis does provide, however, is a partial means of addressing and avoiding the pitfalls of dealing with averaged data in the context of MDS. The BIC analysis constitutes a simple, principled, and effective means of deciding when averaged data is sufficiently precise to warrant the observed levels of data fit achieved by MDS representations. Averaged data derived from single-subject data that contains significant individual differences—through the existence of different underlying configurations, or through noise in the process of generating dissimilarity judgments, or through a combination of both—is identified as inappropriate for MDS modeling. In this way, the use of the BIC prevents averaged data that is not amenable to meaningful representation being misused as the basis for an MDS model.

One course of action, when the BIC rejects the representation of averaged data, is to fit a separate MDS model to each of the single-subject data matrices. In this way, models of the spatial representations assumed to have generated the data may be recovered, rather than all representations being rejected on the basis of the BIC analysis. Unfortunately, however, this single-subject approach guarantees that the resultant MDS models will have considered all of the decision noise in the data, and does not use the potential benefits of averaging. Ideally, what is required is a method that is able to identify those subgroups of subjects using the same underlying spatial representation, average the data within these subgroups, and fit separate MDS models for each. The development of such a method constitutes a worthwhile topic for future research.

Acknowledgments

The authors wish to thank Greg Ashby, Simon Dennis, Todd Maddox, In Jae Myung, Richard Schweickert and Chris Woodruff for helpful comments on previous versions of this article.

Appendix A. The effect of noise on average distance

This appendix derives the change in distance from the origin (using $r = 1$) of an initial point (x, y) that is

perturbed in both directions by additive white Gaussian noise of zero mean and variance σ_p^2. The initial distance from the origin is $d_0 = |x| + |y|$, and the final distance is $d = |x + n_x| + |y + n_y|$. Hence we seek to derive the average change in distance, $E[d - d_0] = E[|x + n_x|] + E[|y + n_y|] - |x| + |y|$. It is useful to consider the quantity in the x dimension only, and exploit the symmetry in the expression.

$$\begin{aligned} E[|x+n|] &= \int_{-\infty}^{\infty} |x+n| p(n)\, dn \\ &= -\int_{-\infty}^{-x} \frac{(x+n)\exp(\frac{-n^2}{2\sigma_p^2})}{\sqrt{2\pi\sigma_p^2}}\, dn \\ &\quad + \int_{-x}^{\infty} \frac{(x+n)\exp(\frac{-n^2}{2\sigma_p^2})}{\sqrt{2\pi\sigma_p^2}}\, dn. \end{aligned}$$

The first half of each term is an incomplete integral of a Gaussian, which is the error function:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)\, dt.$$

The different limits on the integral can be handled by noting that $\operatorname{erf}(\infty) = 1$ and $\operatorname{erf}(-\infty) = -1$. Making the substitution $t = \frac{n}{\sqrt{2\sigma_p^2}}$ and dividing by two yields a more useful form

$$\frac{1}{2}\operatorname{erf}(z) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \int_0^{\sqrt{2}\sigma_p z} \exp\left(\frac{-n^2}{2\sigma_p^2}\right) dn.$$

The second half of each term can be solved using the relation:

$$\int z \exp\left(\frac{-z^2}{2\sigma_p^2}\right) dz = -\sigma_p^2 \exp\left(\frac{-z^2}{2\sigma_p^2}\right) + c.$$

Therefore,

$$\begin{aligned} E[|x+n|] &= -\frac{x}{2}\left(1 + \operatorname{erf}\left(\frac{-x}{\sqrt{2\sigma_p^2}}\right)\right) \\ &\quad + \sqrt{\frac{\sigma_p^2}{2\pi}}\left[\exp\left(\frac{-n^2}{2\sigma_p^2}\right)\right]_{-\infty}^{-x} + \frac{x}{2}\left(1 - \operatorname{erf}\left(\frac{-x}{\sqrt{2\sigma_p^2}}\right)\right) \\ &\quad - \sqrt{\frac{\sigma_p^2}{2\pi}}\left[\exp\left(\frac{-n^2}{2\sigma_p^2}\right)\right]_{-x}^{\infty} \\ &= x \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma_p}\right) + \sqrt{\frac{\sigma_p^2}{2\pi}} \exp\left(\frac{-x^2}{2\sigma_p^2}\right). \end{aligned}$$

Thus

$$\begin{aligned} E[d - d_0] &= x \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2\sigma^2}{\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \\ &\quad + y \operatorname{erf}\left(\frac{y}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2\sigma^2}{\pi}} \exp\left(\frac{-y^2}{2\sigma^2}\right) \\ &\quad - |x| - |y|. \end{aligned}$$

There are two special cases of note, when the initial point is at a vertex of the equidistant contour (i.e., one of x and y will be zero) and when the initial point is in the middle of the straight section of the equidistant contour (i.e., $|x| = |y|$). As an example, consider $x = 1$ and $y = 0$. Therefore,

$$\begin{aligned} E[d - d_0]|_{x=1,y=0} &= \operatorname{erf}\left(\frac{1}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2\sigma^2}{\pi}} \exp\left(\frac{-1}{2\sigma^2}\right) \\ &\quad + \sqrt{\frac{2\sigma^2}{\pi}} - 1. \end{aligned}$$

In contrast, for $x = y = 0.5$.

$$E[d - d_0]|_{x=y=0.5} = \operatorname{erf}\left(\frac{1}{\sqrt{8\sigma^2}}\right) + 2\sqrt{\frac{2\sigma^2}{\pi}} \exp\left(\frac{-1}{8\sigma^2}\right) - 1.$$

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.

Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman & Hall.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, *38*, 467–474.

Furnas, G. W. (1989). Metric family portraits. *Journal of Classification*, *6*, 7–52.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 325–340.

Glushko, R. J. (1975). Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical cluster analysis. *Perception & Psychophysics*, *17*(2), 158–162.

Gregson, R. A. M. (1976). A comparative evaluation of seven similarity models. *British Journal of Mathematical and Statistical Psychology*, *29*, 139–156.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*(1), 133–152.

Heaps, C., & Handel, S. (1999). Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 299–320.

Johnson, E. J., & Tversky, A. (1984). Representations of perceptions of risks. *Journal of Experimental Psychology: General*, *113*(1), 55–70.

Jones, F. N., Roberts, K., & Holman, E. W. (1978). Similarity judgments and recognition memory for common spices. *Perception & Psychophysics*, *24*(1), 2–6.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

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

More, J. J. (1977). The Levenberg–Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Lecture notes in mathematics*, vol. 630 (pp. 105–116). Berlin: Springer.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000a). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.

Myung, I. J., Forster, M., & Browne, M. W. (2000b). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Raftery, A. E. (1999). Bayes factors and BIC: Comment on Weakliem. *Sociological Methods and Research*, *27*, 411–427.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *I. Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. R. Pomerantz, & G. L. Lockhead (Eds.), *The perception of structure*: *Essays in honor of Wendell R Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28.