Journal of Statistical Software

March 2008, Volume 25, Issue 2. http://www.jstatsoft.org/

MLDS: Maximum Likelihood Difference Scaling in R

Kenneth Knoblauch
Inserm, U846

Laurence T. Maloney
New York University

Abstract

The **MLDS** package in the R programming language can be used to estimate perceptual scales based on the results of psychophysical experiments using the method of difference scaling. In a difference scaling experiment, observers compare two supra-threshold differences (a,b) and (c,d) on each trial. The approach is based on a stochastic model of how the observer decides which perceptual difference (or interval) (a, b) or (c, d) is greater, and the parameters of the model are estimated using a maximum likelihood criterion. We also propose a method to test the model by evaluating the self-consistency of the estimated scale. The package includes an example in which an observer judges the differences in correlation between scatterplots. The example may be readily adapted to estimate perceptual scales for arbitrary physical continua.

Keywords: difference scaling, sensory magnitude, proximity, psychophysics, signal detection theory, GLM.

1. Introduction

Difference scaling is a psychophysical procedure used to estimate supra-threshold perceptual differences for stimuli distributed along a physical continuum. On each of a set of trials, an observer is presented with a quadruple, (I_a, I_b, I_c, I_d), drawn from an ordered set of stimuli, $\{I_1 < I_2 < \ldots < I_p\}$. For convenience, the two pairs (I_a, I_b) and (I_c, I_d) are often ordered so that $I_a < I_b$ and $I_c < I_d$ on the physical scale but they need not be. On each trial, the observer judges which pair shows the greater perceptual difference. The experimental data consist of an $n \times 5$ matrix with each row comprising the indices of each quadruple, $(a, b; c, d)$, from the larger set and the observer's response for each of n trials. The output of the scaling procedure are scale values $\{\psi_1, \psi_2, \ldots, \psi_p\}$ that best capture the subject's judgments of the perceptual difference between the stimuli in each pair (I_a, I_b) as we describe in detail below.

In seminal papers, Schneider and colleagues (Schneider 1980a,b; Schneider, Parker, and Stein 1974) applied this procedure to the perception of loudness and proposed a method for es-

timating the difference scale based on the proportion of times the fitted model reproduced the observer's judgments. This method does not explicitly model stochastic variability in the observer's responses. Boschman (2001) proposed a method based on numerical rating of perceptual differences. Subsequently, Maloney and Yang (2003) developed a maximum likelihood procedure, maximum likelihood difference scaling (MLDS), for estimating the parameters of the scale. Their method is based on direct perceptual comparison of pairs of stimuli.

MLDS has been successfully applied to characterize color differences (Maloney and Yang 2003), surface glossiness (Obein, Knoblauch, and Viénot 2004), image quality (Charrier, Maloney, Cherifi, and Knoblauch 2007), adaptive processes in face distortion (Rhodes, Maloney, Turner, and Ewing 2007) and neural encoding of sensory attributes (Yang, Szeverenyi, and Ts'o 2008). The mathematical basis for the method, including necessary and sufficient conditions for a difference scale representation in the absence of observer error, can be found in Krantz, Luce, Suppes, and Tversky (1971, Chapter 4, Definition 1, p. 147). We summarize the most relevant of these conditions below when we discuss diagnostic tests of the model.

In this article, we will describe the MLDS procedure using direct maximization of the likelihood as initially described by Maloney and Yang (2003) and an equivalent approach as a generalized linear model (GLM McCullagh and Nelder 1989). We present an R package (R Development Core Team 2008), **MLDS**, that implements both approaches as well as diagnostics of the estimated scale's validity. In the last section, we will demonstrate the package with an extended example in which we show how to use MLDS to evaluate perception of correlation in scatterplots (Cleveland and McGill 1984b). The package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=MLDS`.

In Figure 1 we show the kinds of stimuli used in the example: the $p = 11$ scatterplots each based on samples of 100 points drawn from bivariate Gaussian distributions that differ only in correlation r. The first ten values of r are equally spaced from 0.0 to 0.9 while the eleventh is 0.98. As explained in the next section, one goal of difference scaling is to develop a perceptual scale that predicts perceived differences between stimuli (here scatterplots) on the physical scale. An example of such a scale is shown as the twelfth graph in the figure. The example code is organized so that it can be readily adapted to other applications.

2. Maximum likelihood difference scaling

In this section, we develop the model of the observer's judgments in the psychophysical task of difference scaling. In each experimental session, the experimenter selects a set of p stimuli, $\{I_1, I_2, \ldots, I_p\}$ ordered along a physical continuum, such as the $p = 11$ correlations in Figure 1. On each trial the experimenter presents an observer with quadruples $(I_a, I_b; I_c, I_d)$, and asks him to judge which pair, I_a, I_b or I_c, I_d, exhibits the *larger perceptual difference*. Figure 2 contains an example of such a quadruple. The observer's task is to judge whether the upper or lower pair of scatterplots exhibits the larger difference. It will prove convenient to replace $(I_a, I_b; I_c, I_d)$ by the simpler notation $(a, b; c, d)$.

2.1. Choosing the quadruples

Over the course of the experiment, the observer sees many different quadruples. The experimenter could choose to present all possible quadruples $(a, b; c, d)$ for p stimuli to the observer or a random sample of all possible quadruples. In past work, experiments have used the set

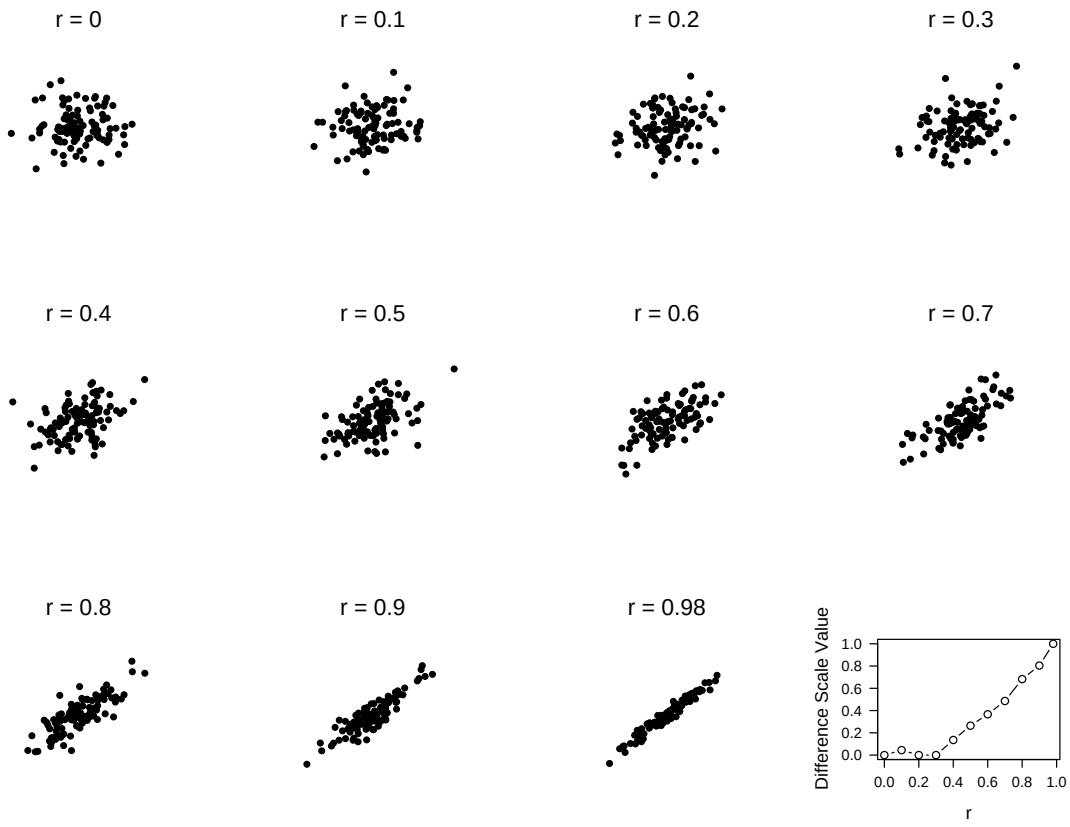

Figure 1: The first 11 graphs are scatterplots of 100 points each drawn from a bivariate Gaussian distribution with the correlation given above each graph. The twelfth graph is a perceptual scale for differences of correlation estimated from the judgments of a single observer. Notice that, for values of correlation less than about 0.4, the scatterplots are difficult to discriminate. The corresponding difference scale is nearly constant across this range.

of all possible non-overlapping quadruples $a < b < c < d$ for p stimuli and the resulting scales have proven to be readily interpretable. Moreover, Maloney and Yang (2003) report extensive evaluations of this subset of all possible quadruples. Consequently we will be primarily concerned with this set of quadruples.

In the example we present, we use only the non-overlapping quadruples. The physical scale values are correlations of bivariate Gaussian distributions, and the stimuli are scatterplots drawn from bivariate Gaussian distributions. The changes needed to employ the methods and diagnostic tests presented here with other subsets of possible quadruples are very slight, should the user prefer a different set of quadruples.

By restricting ourselves to non-overlapping quadruples, we avoid a possible artifact in the experimental design. Suppose we included quadruples such as $(a, b; a, c)$ with $a < b < c$

Figure 2: An example of a trial stimulus presentation from the difference scaling experiment for estimating correlation differences in the scatterplot experiment. The observer must judge whether the difference in perceived correlation is greater in the lower or upper pair of scatterplots.

where the same physical scale value appears twice or quadruples of the form $(b, c; a, d)$ with $a < b < c < d$, where one interval is contained in the interior of the other. Now consider an observer who is actually not capable of comparing intervals and whose behavior cannot therefore be captured by a difference scale. If this observer can correctly order the stimuli, then, on a trial of the form $(a, b; a, c)$, he can still get the right answer by noting that both intervals have a at one end so that $b < c$ implies that (a, b) must be less than (a, c). A similar heuristic applied to $(b, c; a, d)$ with $a < b < c < d$ would allow the observer to appear to be ordering intervals correctly when, in fact, he cannot compare intervals. In either case,we might conclude that the observer is to some extent ordering intervals correctly when in fact he is simply employing a heuristic based on stimulus order. Using only non-overlapping intervals effectively forces the observer to compare intervals. Moreover, as noted above, the use of such intervals has proven itself in previous computational and experimental situations. However, as we shall see, one possible diagnostic test, the three-point test, requires comparison of intervals

of the form $(a, b; a, c)$, and the experimenter interested in applying this test would have to include certain overlapping intervals, as we describe below in the discussion of this test.

If $p = 11$, as in the example, there are

$$\left(\begin{array}{c} 11 \\ 4 \end{array} \right) = \frac{11!}{4!\,7!} = 330 \tag{1}$$

different, non-overlapping quadruples. To control for positional effects, on half of the forced-choice trials, chosen at random, the pairs are presented in the order $(a, b; c, d)$ and on the other half, $(c, d; a, b)$. The order in which quadruples are represented is randomized. For $p = 11$, the observer may judge each of the 330 non-overlapping quadruples in randomized order or the experimenter may choose to have the observer judge each of the quadruples m times, completing $330m$ trials in total. Of course, the number of trials judged by the observer affects the accuracy of the estimated difference scale (See Maloney and Yang 2003). The time needed to judge all 330 trials in the example is roughly 10–12 minutes.

At the end of the experiment, the data are represented as an $n \times 5$ matrix or data frame in which each row corresponds to a trial: four columns give the indices, (a, b, c, d) of the stimuli from the ordered set of p, and one column indicates the response of the observer to the quadruple as a 0 or 1, indicating choice of the first or second pair. For example,

```
  resp S1 S2 S3 S4
1    0  4  8  2  3
2    1  2  3  6 11
3    1  2  6  7 10
4    0  4 11  1  2
5    0  9 11  7  8
6    0  7 10  1  3
```

gives the first 6 rows of the data frame for the observations that generated the scale shown in the lower right of Figure 1.

From these data, the experimenter estimates the perceptual scale values $\psi_1, \psi_2, \ldots, \psi_p$ corresponding to the stimuli, $I_1, \ldots, I_p$, as follows. Given a quadruple, $(a, b\,; c, d)$ from a single trial, we first assume that the observer judged I_a, I_b to be further apart than I_c, I_d precisely when,

$$|\psi_b - \psi_a| > |\psi_d - \psi_c| \tag{2}$$

that is, the difference scale predicts judgment of perceptual difference.

It is unlikely that human observers would be so reliable in judgment as to satisfy the criterion just given, particularly if the perceptual differences $|\psi_b - \psi_a|$ and $|\psi_d - \psi_c|$ are close. Maloney and Yang (2003) proposed a stochastic model of difference judgment that allows the observer to exhibit some stochastic variation in judgment. Let $L_{ab} = |\psi_b - \psi_a|$ denote the unsigned perceived length of the interval I_a, I_b . The proposed decision model is an equal-variance Gaussian signal detection model (Green and Swets 1974) where the signal is the difference in the lengths of the intervals,

$$\delta(a, b; c, d) = L_{cd} - L_{ab} = |\psi_d - \psi_c| - |\psi_b - \psi_a| \tag{3}$$

If δ is positive, the observer should choose the second interval as larger; when it is negative, the first. When δ is small, negative or positive, relative to the Gaussian standard deviation,

σ, we expect the observer, presented with the same stimuli, to give different, apparently inconsistent judgments. The decision variable employed by the observer is assumed to be

$$\Delta(a, b; c, d) = \delta(a, b; c, d) + \epsilon = L_{cd} - L_{ab} + \epsilon \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$: given the quadruple, $(a, b\,; c, d)$ the observer selects the pair I_c, I_d precisely when,

$$\Delta(a, b\,; c, d) > 0. \tag{5}$$

2.2. Direct maximization of likelihood

In each experimental condition the observer completes n trials, each based on a quadruple $\mathbf{q}^k = \left(a^k, b^k; c^k, d^k\right)$, $k = 1, n$. The observer's response is coded as $R^k = 0$ (the difference of the first pair is judged larger) or $R^k = 1$ (second pair judged larger). We denote the outcome "cd judged larger than ab" by $cd \succ ab$ for convenience. We fit the parameters $\mathbf{\Psi} = (\psi_1, \psi_2, \ldots, \psi_p)$ and σ by maximizing the likelihood,

$$L(\mathbf{\Psi}, \sigma) = \prod_{k=1}^{n} \Phi\left(\frac{\delta\left(\mathbf{q}^k\right)}{\sigma}\right)^{1-R_k} \left(1 - \Phi\left(\frac{\delta\left(\mathbf{q}^k\right)}{\sigma}\right)\right)^{R_k}, \tag{6}$$

where $\Phi(x)$ denotes the cumulative standard normal distribution and $\delta\left(\mathbf{q}^k\right) = \delta\left(a^k, b^k; c^k, d^k\right)$ as defined in Equation 4.

At first glance, it would appear that the stochastic difference scaling model just presented has $p + 1$, free parameters: $\psi_1, \ldots, \psi_p$ together with the standard deviation of the error term, σ. However, any linear transformation of the $\psi_1, \ldots, \psi_p$ together with a corresponding scaling by σ^{-1} results in a set of parameters that predicts exactly the same performance as the original parameters. Without any loss of generality, we can set $\psi_1 = 0$ and $\psi_p = 1$, leaving us with the $p - 1$ free parameters, $\psi_2, \ldots, \psi_{p-1}$ and σ. When scale values are normalized in this way, we describe them as standard scales.

Equation 6 can be recognized as the likelihood for a Bernoulli variable. Taking the negative logarithm allows the parameters to be estimated simply with a minimization function such as `optim` in R (for example Venables and Ripley 2002, p. 445).

Reparameterization

In practice, we define the minimization functions to work with the parameters on transformed scales: the $p - 2$ scale values by a logit transformation

$$\log\left(\frac{x}{1 - x}\right) \tag{7}$$

so they are constrained to be in the interval $(0, 1)$ and σ by the logarithm so that it remains positive. These transformations have no theoretical significance; they are used to avoid problems in numerical optimization. Maximum likelihood methods are invariant under such reparameterization (Mood, Graybill, and Boes 1974, pp. 284–285).

2.3. GLM method

In this section, we show how the above formulation can be framed as a GLM. A generalized linear model (GLM McCullagh and Nelder 1989) is described by

$$\eta(\mathsf{E}[Y]) = X\beta, \tag{8}$$

where η is a (link) function transforming the expected value of the elements of the response vector, Y, to the scale of a linear predictor given by the product of the model matrix, X, and a vector of coefficients, β, and the elements of Y are distributed as a member of the exponential family. In the present case, the responses of the observer can be considered as Bernoulli variables and, thus, can be modeled with the binomial distribution which conforms to this family. The canonical link function for the binomial case is the logistic transformation, Equation 7. However, other links are possible, and the inverse cumulative Gaussian, or quantile function, corresponds to Equation 6, described above and would be equivalent to a probit analysis.

In this section, we assume that we have re-ordered each quadruple $(a, b; c, d)$ so that $a < b < c < d$. With this ordering, we can omit the absolute value signs in Equation 3 which then becomes

$$\begin{aligned} \delta &= \psi_d - \psi_c - \psi_b + \psi_a \\ \Delta &= \delta + \epsilon. \end{aligned} \tag{9}$$

The observer bases his or her judgment on $\Delta = \delta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The observer therefore selects the second pair (c, d) with probability $\Phi(\delta)$, where Φ is the Gaussian distribution function.

The design matrix, X, can be constructed by considering the weights of the ψ_i as the covariates. On a given trial, the values in only four columns are non-zero, taking on the values $1, -1, -1, 1$ in that order. This yields an $n \times p$ matrix, X, where n is the number of quadruples tested and p is the number of physical levels evaluated over the experiment. For example, consider a set of 7 stimuli distributed along a physical scale and numbered 1–7. The five quadruples

$$\begin{array}{cccc} 2 & 4 & 5 & 6 \\ 1 & 2 & 3 & 7 \\ 1 & 5 & 6 & 7 \\ 1 & 2 & 4 & 6 \\ 3 & 5 & 6 & 7 \end{array}$$

yield the design matrix

$$X = \begin{pmatrix} 0 & 1 & 0 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & -1 & 1 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 & 1 \end{pmatrix}.$$

To render the model identifiable, however, we drop the first column, which has the effect of fixing $\beta_1 = 0$, yielding a model with $p - 1$ parameters to estimate as with the direct method. This yields the model,

$$\Phi^{-1}(\mathsf{E}[Y]) = \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p, \tag{10}$$

where X_j is the j^{th} column of X. Unlike the direct method, $\psi_p = \beta_p$ is unconstrained. Implicitly in the GLM model, $\sigma = 1$. In fact, $\hat{\beta}_p$ from the GLM fit equals $\hat{\sigma}^{-1}$ from the direct method, so that, all things being equal, normalizing the GLM coefficients by $\hat{\beta}_p$ should yield the same scale as obtained by the direct method.

We have compared solutions using direct optimization (`optim` in R) and GLM fits (`glm` function) and find good agreement. Differences arise occasionally due to the additional constraints that we have imposed when fitting by the direct method.

2.4. Robustness

In R, there is a choice between five built-in link functions for the binomial family, including the logit, probit and cauchit (based on the Cauchy distribution). As of R version 2.4.0, it has become simple for the user to define additional links. In many circumstances, the choice of link is not critical, since over the rising part of these functions, they are quite similar. The difference scaling procedure, however, generates many responses at the tails, i.e., easily discriminable differences and one might think that it would be more sensitive to the choice of link.

Maloney and Yang (2003) evaluated distributional robustness of the direct optimization method. They varied the distributions of the error term ϵ while continuing to fit the data with the constant variance Gaussian error assumption. They found that MLDS was remarkably resistant to failures of the distributional assumptions. Hence, the GLM approach using the `probit` link is likely to be adequate for most applications of MLDS.

2.5. Goodness of fit

Use of the GLM approach benefits from the availability of several diagnostic tests available in R for generalized linear models, and we report a measure "proportion of deviance accounted for" (DAF) that has been suggested (Wood 2006, p. 84) as well as the Akaike information criterion (AIC Akaike 1973). Some standard diagnostic measures are problematic or difficult to interpret for binary data, however, because the distribution of the deviance cannot be guaranteed to approximate a χ^2 distribution (Venables and Ripley 2002; Wood 2006). Here, we implement two diagnostic tests suggested by Wood (2006) based on a bootstrap analysis of the distribution and independence of the residuals in the example below. The principle on which these tests are based would be applicable to any GLM model.

2.6. Diagnostic tests of the measurement model

Even if the data passed an overall goodness of fit test, there still may be patterned failures in the data that would allow rejection of the difference scaling model. In this section, we describe two additional diagnostic tests based on the necessary and sufficient conditions that an observer must satisfy if we are to conclude that his judgments can be described by a difference scaling model (Krantz *et al.* 1971, Chapter 4, p. 147, Definition 1) discussed above. These tests are specific to difference scaling and they correspond to necessary conditions for the existence of a difference scale in the non-stochastic case in Definition 1 of Krantz et al.

The six-point test

The first condition is the six-point condition, illustrated in Figure 3. It is referred to as the "weak monotonicity" condition in Krantz *et al.* (1971, Chapter 4, p. 147 and Figure 1) in Definition 1, Axiom 4, p. 147. It is also known as the "sextuple condition" (Block and Marschak 1960). We describe the condition with an example. Suppose that there are two groups of three stimuli whose indices are $a < b < c$, and $a' < b' < c'$, respectively. Suppose that a non-stochastic observer considers the quadruple$(a, b; a', b')$ and judges that $ab \succ a'b'$, that the interval ab is larger than the interval a', b'. On some other trial, he considers $(b, c; b'c')$ and judges that $bc \succ b'c'$. Now, given the quadruple, $(a, c; a', c')$ there is only one possible response consistent with the difference scaling model: he or she must choose $ac \succ a'c'$. The reasoning behind this constraint is illustrated in the figure and it can be demonstrated directly from the model.

For the non-stochastic observer, even one violation of this six-point condition would allow us to conclude that there was no consistent assignment of scale values $\psi_1, \psi_2, \ldots, \psi_p$ in a difference scaling model that could predict his or her judgments in a difference scaling task.

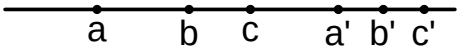

Figure 3: Six-point condition: Given stimuli $a < b < c$ and $a' < b < c'$ ordered along a scale, if $ab \succ a'b'$ and $bc \succ b'c'$, then $ac \succ a'c'$.

The six-point condition is a slightly disguised test of additivity of contiguous intervals in the difference scale. To see how it might fail, imagine that distances in the scale correspond to chordal distances along a circular segment as shown in the left side of Figure 4. Then the six-point condition in equality form implies that if $ab = a'b'$ and $bc = b'c'$, then $ac = a'c'$ where $=$ denotes subjective equality. If the six-point condition and other necessary conditions hold, then the chordal distances on the left side of Figure 4 can be represented along a line as in the previous Figure 3 (see Krantz *et al.* 1971, Chapter 4, for further discussion). On the right side of Figure 4, in contrast, judgments are based on chordal distances along an ellipse. The six-point condition fails and these judgments cannot be represented by a difference scale.

Human judgments in difference scaling tasks are not deterministic: if we present the same quadruple at two different times, the observer's judgments need not be the same. The MLDS model allows for this possibility. In MLDS decisions are based on a decision variable $\Delta(a, b; c, d)$ and, for any given six points a, b, c and a', b', c' there is a non-zero probability that the stochastic observer will violate the six-point condition. In particular, suppose that $\psi_b - \psi_a$ is only slightly greater than $\psi_{b'} - \psi_{a'}$, $\psi_c - \psi_b$ is only slightly greater than $\psi_{c'} - \psi_{b'}$, and $\psi_c - \psi_a$ is only slightly greater than $\psi_{c'} - \psi_{a'}$. Then we might expect that the observer's probability of judging $ab \succ a'b'$ is only slightly greater that 0.5 and similarly with the other two quadruples. Hence, he has an appreciable chance of judging that $ab \succ a'b'$ and $bc \succ b'c'$ but $ac \prec a'c'$ or $ab \prec a'b'$ and $bc \prec b'c'$ but $ac \succ a'c'$, either a violation of the six-point

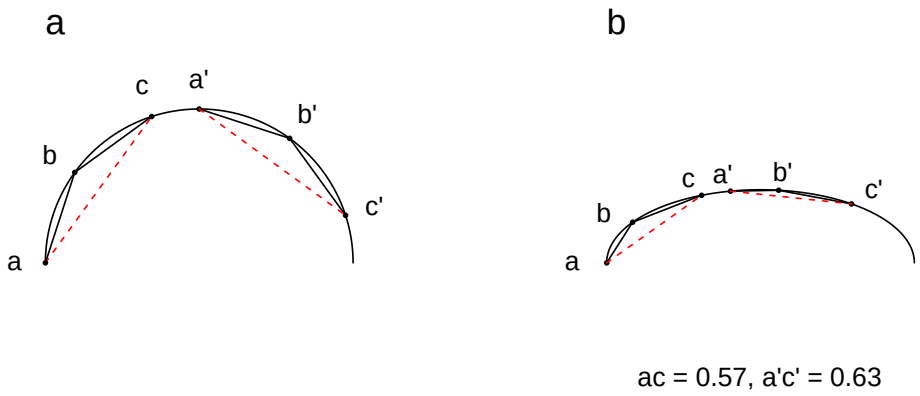

Figure 4: Six-point condition: Left. Given stimuli $a < b < c$ and $a' < b < c'$ ordered along a circular (constant curvature) segment, if the chordal distances $ab \approx a'b'$ and $bc \approx b'c'$, then $ac \approx a'c'$. Right. The six-point condition fails on a contour with non-constant curvature.

property.

Maloney and Yang (2003) proposed a method for testing the six-point property that takes into account the stochastic nature of the observer's judgment and uses a resampling procedure (Efron and Tibshirani 1993) to test the hypothesis that the MLDS model is an appropriate model of the observer's judgments.

Given the experimental design and all of the quadruples used, we can enumerate all six-point conditions present in the experiment, indexing them by $k = 1, n_6$. We count the number of times, V_k, that the observer has violated the k^{th} six-point condition during the course of the experiment and the number of times he has satisfied it, S_k. If we knew the probability that the observer should violate this six-point condition p_k, then we could compute the probability of the observed outcome by the binomial formula,

$$\Lambda_6^k = \left(\begin{array}{c} V_k + S_k \\ V_k \end{array} \right) p_k^{V_k} (1 - p_k)^{S_k}, \tag{11}$$

and we could compute the overall likelihood probability

$$\Lambda_6 = \prod_{k=1}^{N_6} \Lambda_6^k. \tag{12}$$

Under the hypothesis that the difference scale model is an accurate model of the observer's judgments, we have the fitted estimates of scale values $\hat{\psi}_1, \ldots, \hat{\psi}_p$ and $\hat{\sigma}$. We can compute estimates of the values $\hat{\Lambda}_6^k$ based on these scale values and compute an estimate of $\hat{\Lambda}_6 = \prod_{k=1}^{N_6} \hat{\Lambda}_6^k$. This is an estimate of the probability of the observed pattern of six-point violations and successes. We next simulate the observer N times with the fitted parameter values $\hat{\psi}_1, \ldots, \hat{\psi}_p$ and $\hat{\sigma}$ of the actual observer used for the simulated observer and perform the analysis above to get N bootstrap estimates $\hat{\Lambda}_6^*$ of $\hat{\Lambda}_6$. Under the hypothesis that MLDS

is an accurate model of the observer's judgments, $\hat{\Lambda}_6$ should be similar in value to $\hat{\Lambda}_6^*$ and we employ a resampling procedure to test the hypothesis at the 0.05 level by determining whether $\hat{\Lambda}_6$ falls below the 5th percentile of the bootstrap values $\hat{\Lambda}_6^*$ (See Maloney and Yang (2003) for details).

The three-point test

The second empirically-testable necessary condition for a difference scale representation of data is the three-point condition (Krantz *et al.* 1971, Chapter 4, p. 147, Definition 1, Axiom 3). Given three stimuli $a < b < c$, the non-stochastic observer must judge that $ac \succ ab$ and $ac \succ bc$: an interval must be greater than a proper interval contained within it. Often, in difference scaling applications, this three-point property is evidently satisfied and not formally tested. In some applications, the observer can simply be shown the test stimuli and asked to order them. If he or she can do so in agreement with the physical scale, further testing of the three-point condition can be omitted.

In the stochastic case, subjects may confuse some of the stimuli on some trials. In the example of Figure 1 many observers will confuse the stimuli with lowest correlation values. The three-point property can be stated in a form appropriate for the stochastic case as: if $a < b < c$ then the probability of judging ab as greater than ac is less than or equal to 0.5 and similarly for bc and ac.

To test the 3-point condition, we must include quadruples of the form $(a, b; a, c)$ with $a < b < c$. As noted above, the inclusion of such quadruples could introduce an artifact in the experimental design: the subject can correctly order the intervals based on consideration of only the order of the stimuli. If the experimenter excludes such quadruples then he cannot test the three-point condition, and in any application the experimenter must decide if a test of the three-point condition is warranted.

We do not provide a three-point test in the **MLDS** package. If the experimenter does choose to include quadruples of the form $(a, b; a, c)$ with $a < b < c$ ("3-point quadruples"), then it is very easy to design a three-point test patterned on the six-point test above. We use the fitted scale values and estimated σ to bootstrap an ideal observer matched to the actual. It is appropriate to exclude the 3-point quadruples in this initial fit of the scale. We then repeatedly compute the log likelihood $\hat{\Lambda}_3^*$ of the ideal observer's performance for the "three-point intervals" and then compare the actual log likelihood $\hat{\Lambda}_3$ to the distribution of bootstrap replications $\hat{\Lambda}_3^*$. We reject if it falls below the αth percentile of the bootstrap values for appropriate choice of α. If the observer's performance is consistent with the three-point condition, then the scale can be re-fitted using all data including the three-point quadruples.

Other necessary conditions and alternative axiomatizations

Krantz *et al.* (1971, Chapter 4, p. 147, Definition 1) list six conditions ("axioms") that are jointly necessary and sufficient for a data set to be representable by a difference scale in the non-stochastic case. The three-point and six-point diagnostic tests were based on two of these conditions (Definition 1, Axioms 3,4, respectively). Of the remaining necessary conditions, two effectively guarantee that the experimental design contains "enough" distinct quadruples, and that the observer can order intervals transitively (Axioms 1,2). Axiom 5 is satisfied if the values of the physical scale can be put into correspondence with an interval of the real numbers (evidently true in our example for correlation $-1 \leq r \leq 1$). Axiom 6 precludes the possibility

that an interval ab with $a < b$ contains an infinite number of non-overlapping intervals that are all equal. In the stochastic case, these conditions are either evidently satisfied or are replaced by consideration of the accuracy and stability of the estimated scale. Maloney and Yang (2003) have investigated accuracy, stability and robustness in some detail.

There are alternative sets of axiomatizations of difference scaling such as Suppes (1972) and, of course, all sets of necessary and sufficient conditions are mathematically equivalent. We have chosen those of Krantz *et al.* (1971) because they lead to simple psychophysical tasks. Observers are asked only to order intervals. Either they can do so without violating the six-point and three point conditions or they cannot and whether they can or cannot is an experimental question. Krantz *et al.* (1971, Chapter 4) contains useful discussion of the link between the axiomatization that they propose and the task imposed on observers.

The reader may wonder why the observer is asked to compare intervals and not just pairs of stimuli. Krantz *et al.* (1971, Chapter 4, pps. 137–142) contains an extended discussion of the advantages of asking observers to directly compare intervals. We note only that pairwise comparison of the stimuli (i.e. given a, b, judge whether $a < b$) does not provide enough information to determine spacing along the scale in the non-stochastic case. Any increasing transformation of a putative difference scale would predict the same ordering of the stimuli. In the stochastic case the observer may judge that $a < b$ on some trials and that $b < a$ on others, and the degree of inconsistency in judgment could potentially be used to estimate scale spacing using methods due to Thurstone (1927). Thurstone scales, however, have three major drawbacks. First, the scale depends crucially on the assumed distribution of judgment errors (it is not robust) while MLDS is remarkably robust (see Maloney and Yang 2003). Second, stimuli must be spaced closely enough so that the observer's judgments will frequently be inconsistent. This typically means that many closely-spaced stimuli must be scaled, and the number of trials needed to estimate the scale is much greater than in MLDS. The third drawback is the most serious. It is not obvious what the Thurstonian scale measures, at least not without further assumptions about how "just noticeable differences" add up to produce perceptual differences. The MLDS scale based on quadruples is immediately interpretable in terms of perceived differences in interval lengths since that is exactly what the observer is judging.

3. Package MLDS

The package **MLDS** includes the function `mlds` for estimating a perceptual difference scale by MLDS, and the function `simu.6pt` for performing the six-point test as described above.

3.1. Fitting with `mlds`

The first argument of `mlds` is the data frame containing the results from a difference scaling experiment. The function expects the data to be organized as $n \times 5$ columns. Each row represents one trial. The first column named `resp`, of either type logical or a vector of 0s and 1s, gives the responses of the observer. The next four, named `S1`, `S2`, `S3` and `S4` indicate the indices of the four stimuli comprising the quadruple for that trial.

Frequently, the data from an experiment are ordered to indicate the positioning of the stimuli in the experiment and not the physical ordering of the stimuli, as `mlds` expects. For example, a trial

```
resp   S1   S2   S3   S4
   1    7    9    2    4
```

might indicate that the stimulus pair $(7, 9)$ was presented below and pair $(2, 4)$ above and that the observer chose the second pair as showing the greater difference. The function `SwapOrder` will check for such inversions and outputs a new data frame with the orders sorted and the responses inverted in case of inversion. If the results of `SwapOrder` are stored in the original variable, the original ordering is lost to subsequent applications.

An object of class "`mlds.df`" is defined to be a data frame from a difference scaling experiment, as described above but with attributes, "`invord`" and "`stimulus`". Attribute "`invord`" is a logical vector indicating whether the order was reversed in the original experiment. `SwapOrder` checks whether its input is of class "`mlds.df`" and if so, uses the "`invord`" attribute to re-order the data. In this case, successive applications flip the ordering back and forth between that of the experimental and sorted state. The results from several experiments can be combined into a single object of class "`mlds.df`" using the "`mlds.df`" method `rbind`, which concatenates the "`invord`" attributes as well as the rows of the data frame.

The second argument of `mlds` indicates the physical stimulus levels to which the indices in the data refer. If `NULL`, (the default), then it is set to either a sequence from 1 to the maximum index level in the data or a numeric vector stored as the attribute "`stimulus`", if present.

`mlds` uses `glm` by default, but `optim` may be chosen with the argument `method`. If the default is chosen, then the data are transformed to the design matrix within `mlds` using the function `make.ix.mat`. `ix.mat2df` performs the inverse transformation. If `optim` is chosen, its method defaults to `BFGS` but may be modified with the argument `opt.meth`. Using `optim` requires initial estimates passed through the argument `opt.init`. The default link is `probit` but this may be modified to alternative links from the binomial family or a user-defined link (R 2.4.0 or greater) with the argument `lnk`. Additional named arguments will be passed along to `glm` or `optim` through the `...` argument.

`mlds` produces an object of class "`mlds`" which is a list of components: `pscale`, a numeric vector containing the estimated perceptual scale, `stimulus`, a numeric vector of the physical scale, `sigma`, the value of σ (always 1 for `method = "glm"`), `link`, a character string indicating the link used and `method`, a character string specifying the method used. For `method = "optim"`, there are also, the log likelihood, Hessian, data frame and convergence condition returned in components `logLik`, `hess`, `data` and `conv`, respectively. For `method = "glm"`, the component `obj` contains the "`glm`" object from the fit which is used by the "`mlds`" methods to extract the information provided in the additional components when `optim` is specified.

There are seven methods currently defined for objects of class "`mlds`": `print`, `summary`, `plot`, `fitted`, `predict`, `logLik` and `AIC`. The `plot` method generates a graph of the estimated perceptual scale as a function of the physical stimulus. The function `boot.mlds` is provided to estimate standard errors for each scale value using a resampling procedure (Efron and Tibshirani 1993). In short, the fitted probabilities are used to generate new responses to the trials with the function `rbinom`. The estimated scale values for the new responses provide a bootstrap sample. In a similar manner, the function `binom.diagnostics` allows running two diagnostics based on bootstrap replications of the residuals in order to evaluate the suitability of the model.

For `method = "glm"`, the model formula used is

```
resp ~ . - 1
```

There is no `update` method defined currently for `mlds`. However, for the default method, the "`glm`" object is stored in the returned object as an element named `obj`. This object may be updated if care is taken also to include the data in the call, since ordinarily the data frame passed to the `glm` call is only visible within `mlds`. For example, if `x.mlds` is an object of class "`mlds`" obtained with the default method, one can try

```
with(x.mlds, update(obj, .  ~ . + 1, data = obj$data))
```

to obtain a model with an intercept term.

3.2. Running a six-point test with `simu.6pt`

The initial step in performing a six-point test requires identifying all of the six-point conditions in the experimental data. The function `Get6pts` takes an object of class "`mlds`" and returns a list of three data frames, with an attribute "`indices`" which is a fourth data frame. The three data frames are named `A`, `B` and `E` (the last to avoid the R function names `C` and `D`). All of these data frames have the same number of rows, and corresponding rows of `A`, `B` and `E` provide six-point cases for evaluation. The data frame attribute provides indices from the original data set, that from which the "`mlds`" object was generated, to the rows at which each trial occurred. For example, for the difference scale fit to the data set `AutumnLab` included in the package, row 4 of the three data frames generated by `Get6pts` is

```
  resp S1 S2 S3 S4
A 0    1  2  4  5
B 1    2  3  5  9
E 1    1  3  4  9
```

`A` gives the comparisons $(a, b; a'b')$, `B` $(b, c; b', c')$ and `E` $(a, c; a'c')$. Note that whether or not this example corresponds to a violation of the six-point condition depends on the differences between the perceptual scale values to which these indices correspond. Row 4 of the attribute is

```
  A   B   E
116  23  25
```

which are the rows of `AutumnLab` from which the three trials of the six-point condition were extracted.

4. Example: Perception of correlation in scatterplots

The study of graphical perception for enhancing data presentation has been of interest to statisticians, at least, since the pioneering work of Cleveland and colleagues (Cleveland and McGill 1984a). Scatterplots have often been the subject of investigation, for example, to determine the characteristics that best reveal the underlying association in the data (Cleveland, Diaconis, and McGill 1982) or the visual parameters that create the most salient differences

between overlaid data sets (Cleveland and McGill 1984b). Only a few studies have examined the sensitivity of human observers to statistical differences in scatterplots (see, for example, Legge, Gu, and Luebker 1989). MLDS offers a promising method for approaching such questions.

4.1. A psychophysical experiment

Executing

```
R> runSampleExperiment(DisplayTrial = "DisplayOneTrial",
+    DefineStimuli = "DefineMyScale")
```

runs 330 trials of the difference scaling experiment and records the observer's responses interactively on each trial for the scatterplot example of Figure 1. The stimulus from an example trial is shown in Figure 2. The observer's task is to decide whether the difference in r is greater between the lower pair or the upper and to enter a `1` or `2`, respectively, from the keyboard. This function can be readily modified to any difference scaling application by defining the functions `DefineStimuli` and `DisplayTrial` that define the stimuli and display the quadruples, respectively, of non-overlapping intervals on each trial. After the observer has completed the experiment, an object of class "`mlds.df`" is returned which can be used for further analysis. To preserve its attributes, it should be saved with `save` or `dput`.

One of the authors ran the experiment on himself three times, with the results stored in objects `kk1`, `kk2` and `kk3`. Each of the runs of 330 trials required less than 12 minutes to complete. After loading the three data sets in memory, we merge them into one object of 990 trials with `rbind` and apply `SwapOrder` to put the stimulus in physical order.

```
R> data("kk1")
R> data("kk2")
R> data("kk3")
R> kk <- SwapOrder(rbind(kk1, kk2, kk3))
```

4.2. Estimating a perceptual scale

For comparison, we fit the data by `mlds` using both `glm` and `optim` methods. Using `method = "optim"` is usually slower.

```
R> kk.mlds <- mlds(kk)
R> summary(kk.mlds)

Method:        glm                Link:        probit

Perceptual Scale:
      0     0.1     0.2     0.3     0.4     0.5     0.6     0.7
 0.0000 -0.0454  0.0439 -0.0863  0.5682  1.4234  2.0695  2.6661
   0.8     0.9    0.98
 3.5527  4.4297  5.5739
```

```
sigma:        1
logLik:        -306
```

```
R> kkopt.mlds <- mlds(kk, method = "optim", opt.init = c(seq(0, 1,
+     length.out = 11), 0.2))
R> summary(kkopt.mlds)
```

```
Method:        optim                Link:        probit

Perceptual Scale:
       0      0.1      0.2      0.3      0.4      0.5      0.6
0.00e+00 4.70e-05 1.54e-02 1.19e-07 1.10e-01 2.61e-01 3.76e-01
     0.7      0.8      0.9     0.98
4.83e-01 6.40e-01 7.96e-01 1.00e+00

sigma:        0.175
logLik:        -307
```

Note the differences in the summaries. The `glm` method fixes $\sigma = 1$ and does not constrain the upper range of scale values. The `optim` method fixes the extreme values of the scale to 0 and 1. Differences in the log likelihood can result because with the `optim` method, we have constrained the estimated scale values to be in $(0, 1)$. These are usually quite small, as here, unless `optim` has found a local minimum. This also accounts for the slight discrepancy between the `optim` value of $\sigma = 0.175$ and the reciprocal of the maximum scale value using `method = "glm"`, 0.179.

We compare the two estimated scales in Figure 5 by first normalizing the scale estimated by `glm` by its maximum scale value. This is easily accomplished by setting the argument `standard.scale = TRUE` in the plot method. The `glm` and `optim` scales are shown as points and lines, respectively, and it is clear that any differences between the two are unimportant. Note that the resulting perceptual scale is almost flat for correlations up to 0.3. If we plot the estimated scale instead as a function of squared correlation r^2 we see that for correlations above 0.4, the observer's judgment effectively matches r^2, the variance accounted for. Below this value, the observer seems unable to discriminate scatterplots.

We have noticed that using `glm` can frequently generate a warning message

```
Warning message:
fitted probabilities numerically 0 or 1 occurred in: glm.fit(x = X, y = Y,
weights = weights, start = start, etastart = etastart,
```

There are several possible sources of this warning. The obvious possibility that the fit has perfectly separated the responses of the observer to scale differences can be discounted by examining the fitted probabilities as a function of the observer's responses. For example, Figure 6 demonstrates the overlap of responses for the `kk2` and `AutumnLab` data sets, both of which generate this warning. The warning can be generated as well if just some of the fitted probabilities are effectively a 0 or 1. We can imagine this occurring if some of the intervals that the observer must differentiate are so large that errors are never made. This indeed

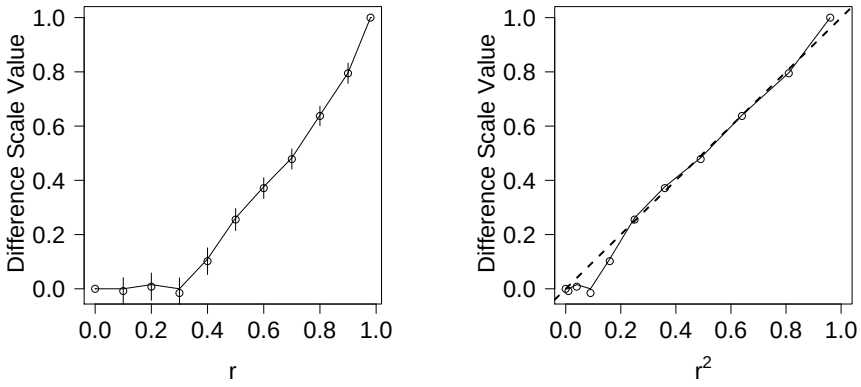

Figure 5: Left. Estimated difference scale for observer KK, from 990 judgments, distributed over 3 sessions for judging differences of correlation between scatterplots. The error bars correspond to twice the standard deviation of the bootstrap samples. Right. The same scale values plotted as a function of the squared correlation or variance accounted for. The diagonal line through the origin has unity slope.

may occur when comparing large and small suprathreshold differences, as here, and, in fact, can be taken as an indication that the observer does exploit an internal scale when making judgments. It is pertinent to note that this warning disappears for the fit to both of these data sets if the link is changed to either `logit` or `cauchit`, suggesting that the shape of the psychometric function could be at issue, here. These link functions differ primarily in the tails of the corresponding distributions.

A third possibility could arise from unpatterned response errors (the observer pressed the wrong key in recording his judgment on some trials). This can produce the Hauck-Donner

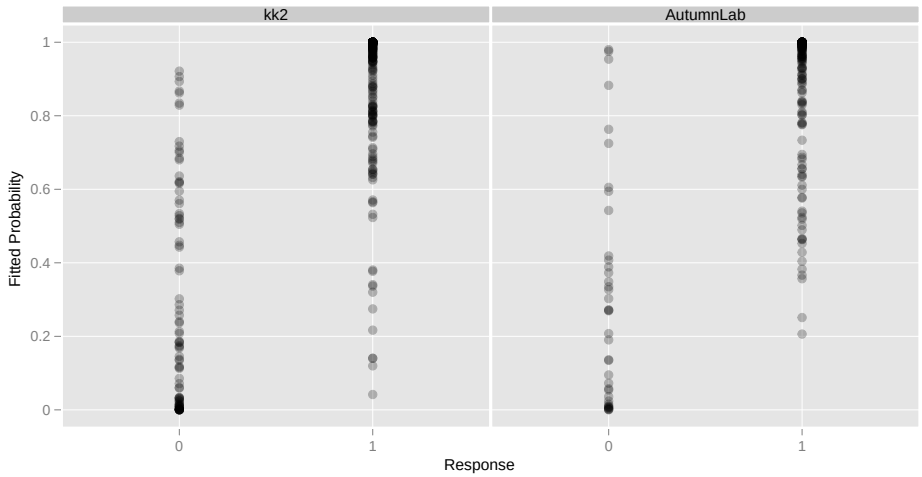

Figure 6: Fitted probabilities as a function of observer's response for the `kk2` and `AutumnLab` data sets.

phenomenon (Hauck and Donner 1977; Venables and Ripley 2002). In that case, the Wald approximation to the log-likelihood may be quite poor. The profile plots for the coefficients are curved in the above two data sets, and especially so for the larger coefficients. Under these circumstances, It may be preferable to use `optim` to estimate the scale by direct maximization of the likelihood (Venables and Ripley 2002), and a bootstrap approach to estimate standard errors. Collecting more data is recommended to address this warning, and we find that the warning disappears for `kk2` if we combine it with either of the other two replications. Despite the warning, the estimated scales using `glm` or `optim` are nearly the same.

4.3. Bootstrapping standard errors

Bootstrap standard errors of the scale values are obtained with the function `boot.mlds`. Running 10000 trials on the dataset `kk` required about 14 minutes on a 2.16 GHz Mac Intel Pro with 2 Gb of memory. A list is returned with 4 components, as indicated by applying `str`, below. The first, `boot.samp`, is a matrix of the bootstrapped scale values. The means and standard deviations are indicated in vectors `bt.mean` and `bt.sd`, respectively.

```
kk.bt <- boot.mlds(kk.mlds, 10000)
str(kk.bt)
```

```
List of 4
 $ boot.samp: num [1:11, 1:10000]  0.00000 -0.02075  0.00361 -0.01485  ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:11] "" "stim.2" "stim.3" "stim.4" ...
  .. ..$ : NULL
 $ bt.mean  : Named num [1:11]  0.00000 -0.00876  0.00710 -0.01630  ...
  ..- attr(*, "names")= chr [1:11] "" "stim.2" "stim.3" "stim.4" ...
 $ bt.sd    : Named num [1:11] 0.0000 0.0245 0.0252 0.0278 0.0246 ...
  ..- attr(*, "names")= chr [1:11] "" "stim.2" "stim.3" "stim.4" ...
 $ N        : num 10000
```

Twice the bootstrap errors are indicated as error bars in the left graph of Figure 5.

4.4. Goodness of fit

A first qualitative test of the validity of the estimated scale is to compare it with the actual stimuli to determine if it does, in fact, capture the perceptual variation displayed. For example, examination of the stimuli and of the estimated scale in Figure 1 confirms that the initial flat part of the scale corresponds to a range of stimuli that are difficult to distinguish and that subsequent stimuli do appear to increase in correlation, as the scale indicates. The quadratic dependence of the increasing part of the scale probably cannot be detected in this fashion. Estimated scales that do not display such *face validity* should certainly be re-examined.

Several approaches have been suggested to analyze the appropriateness of the model for binary data. For comparison, we will consider analyses of the `kk` data set with the probit (default, already calculated above), logit and cauchit links.

```
R> kk.mlds.logit <- mlds(kk, lnk = "logit")
R> kk.mlds.cauchit <- mlds(kk, lnk = "cauchit")
```

link	DAF	AIC	Pr(Runs)
probit	0.55	633	0.02
logit	0.57	610	0.22
cauchit	0.57	611	0.07

Table 1: Goodness of fit.

When `method = "glm"` is used, it is easy to extract the residual and null deviances from the model object component to calculate a deviance accounted for (DAF), analogous to an R^2 calculated for linear models (Wood 2006, p. 84). For example,

```
R> (kk.mlds$obj$null.deviance - deviance(kk.mlds$obj)) /
+    kk.mlds$obj$null.deviance
```

for the probit link. The results for the three link functions, given in Table 1 in the first column, indicate a negligible superiority of the other two links over the default.

The second column of Table 1 displays the AIC values for each link function, obtained with the `AIC` method applied to each model object. The logit link shows a more decided advantage over the probit, in this case.

Wood (2006) proposed two diagnostics for evaluating the suitability of the model fit to the data, each one based on the distribution of the deviance residuals of the fit. The first involves a comparison of the empirical distribution of the residuals to an envelope of the $1-\alpha$ proportion of the bootstrap-generated residuals. The second tests the dependence of the residuals by comparing the number of runs of positive and negative values in the sorted deviance residuals with the distribution of runs from the bootstrapped residuals.

We provide a function `binom.diagnostics` to implement both of these for objects of class "`mlds`". The function takes two arguments: `obj`, an "`mlds`" model object and `nsim`, the number of bootstrap simulations to run. For example,

```
R> kk.diag.prob <- binom.diagnostics(kk.mlds, 10000)
```

performs 10000 simulations. An object of class "`mlds.diag`" is returned that is a list of 5 components, illustrated below for the `probit` link.

```
R> str(kk.diag.prob)

List of 5
 $ NumRuns  : int [1:10000] 195 193 205 199 177 177 201 211 187 209 ...
 $ resid    : num [1:10000, 1:990] -5.25 -4.79 -4.77 -4.56 -4.52 ...
 $ Obs.resid: Named num [1:990]  0.458  0.164  0.413  1.155 -1.610 ...
  ..- attr(*, "names")= chr [1:990] "1" "2" "3" "4" ...
 $ ObsRuns  : int 173
 $ p        : num 0.0159
 - attr(*, "class")= chr [1:2] "mlds.diag" "list"
```

`NumRuns` is a vector of integer giving the number of runs in the sorted deviance residuals for each simulation. `resid` is a matrix of numeric, each row of which contains the sorted deviance

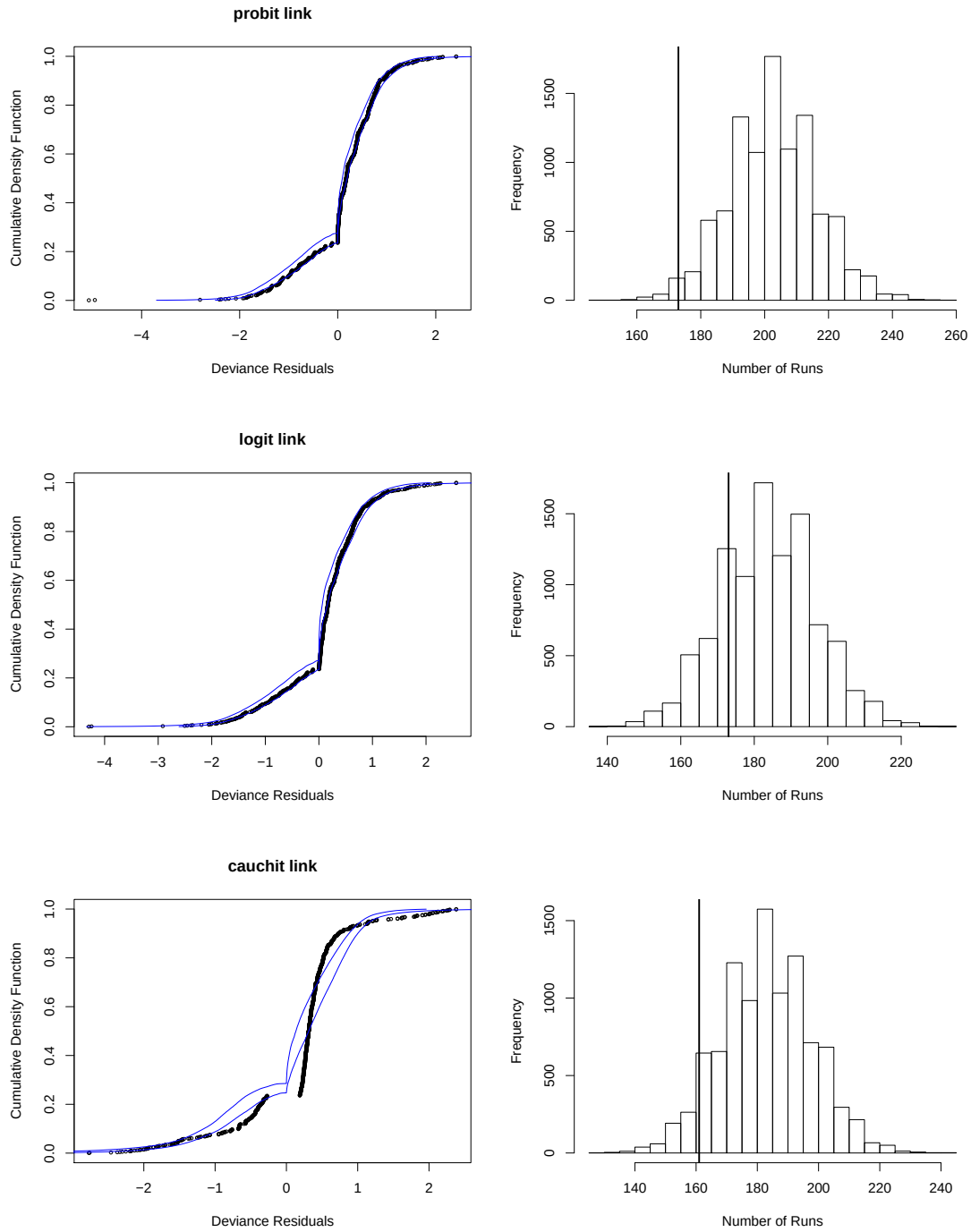

Figure 7: Diagnostic graphs produced by `plot.mlds.diag` for the models fit with the three link functions. The left graphs show the empirical cdf of the deviance residuals (black points) compared to the 95% bootstrapped envelope (blue lines). The right graphs display a histogram of the number of runs in the sign of the sorted deviance residuals from the bootstrap simulations. The observed value is indicated by a vertical line.

residuals from a simulation. `Obs.resid` is a vector of numeric providing the residuals from the obtained fit. `ObsRuns` is the number of observed runs in the sorted deviance residuals, and finally `p` gives the proportion of runs from the simulations less than the observed number. A `plot` method is supplied to visualize the results of the two analyses which are shown in Figure 7 for each of the link functions.

The distributions of the residuals for the probit and logit links seem reasonable, based on 10000 simulations. Tendencies toward deviation from the envelopes are small, in any case. These contrast with the cauchit link, that displays systematic deviations from the bootstrapped envelope.

The histograms indicate that there are too few runs in the residuals using the probit link. For the logit, the observed number falls well within the distribution of bootstrapped values, while the cauchit value, given its performance with the previous diagnostic, is debatable. The proportion of simulated runs less than the observed value for each link is given in the third column of Table 1.

Two points on the far left of the cdf's of Figure 7 stand out as having unusually large residual deviances. These points, as well as a third one, are flagged, also, by the diagnostics generated by the glm `plot` method. The three trials are simply extracted from the data set.

```
R> kk[residuals(kk.mlds$obj) < -2.5, ]
```

```
    resp S1 S2 S3 S4
295    0  1  2  3 10
857    0  1  2  4 10
939    0  1  2  9 11
```

Interestingly, if these points are removed from the data set, the value of `p` for the probit link increases to the value of 0.24, more in line with that obtained using the logit link. The number of runs does not change in the observed data, but the bootstrapped distribution of the number of runs shifts to a mean near 171.

Judging from the estimated scale as well as the stimuli, it seems surprising that the observer would have judged the correlation difference between 0 and 0.1 to be greater than that of 0.3 (or 0.4) and 0.9. We suspect that these correspond to non-perceptual errors on the part of the observer, such as fingerslips, lack of concentration or momentary confusion about the instructions. A few such errors nearly always occur in psychophysical experiments, even with practiced and experienced observers. In modeling data from detection experiments, it has proven useful to include a nuisance parameter to account for these occasional errors (Wichmann and Hill 2001).

The error rates are modeled by modifying the lower and upper asymptotes of the inverse link function. We can get a sense of the impact of adding these two nuisance parameters by using links `mafc.probit` and `probit.lambda` from the **psyphy** package, which permit specifying these asymptotes differently from 0 and 1 (Knoblauch 2007). Preliminary experimentation on the full data set indicates that the AIC is reduced by 48 if the lower estimate is set to about 0.06 and the upper to 0.007. These values also lower the number of runs in the distributions of bootstrapped residuals, so that the observed value yields $\texttt{p} = 0.7$.

4.5. The six-point test

Performing a six-point test on these data with 10000 simulations requires about 15 minutes on the same machine indicated above.

```
kk.6pt <- simu.6pt(kk.mlds, 10000)
str(kk.6pt}
```

```
List of 4
 $ boot.samp: num [1:10000] -488 -539 -531 -502 -447 ...
 $ lik6pt   : num -425
 $ p        : num 0.848
 $ N       :num 10000
```

Examination of the structure of the returned list with `str` shows the p-value and log-likelihood for the number of violations of the six-point test from the data and indicates that the observer did not make a significantly greater number of six-point violations than an ideal observer. Figure 8 shows a histogram of the log-likelihoods from such a simulation with the observed log-likehood indicated by a thick vertical line. These results support the appropriateness of the scale as a description of the observer's judgments.

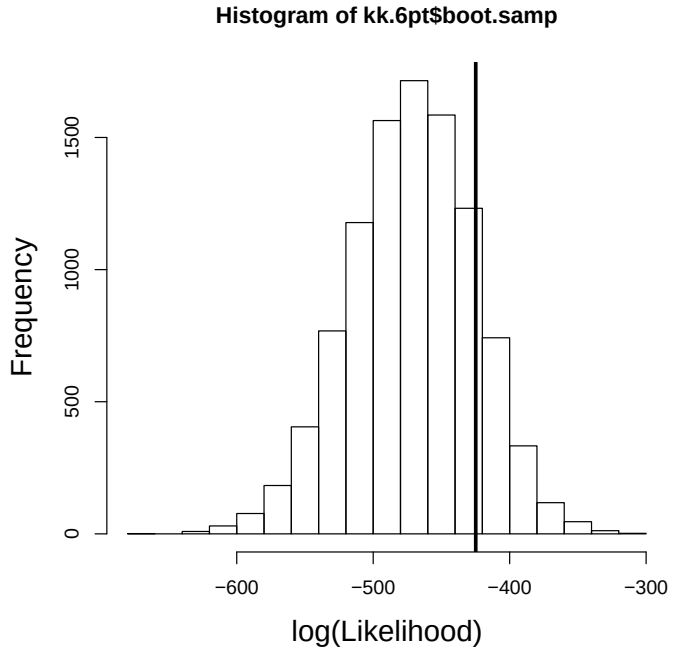

Figure 8: Histogram of log-likelihood values from the six-point test to data set `kk`. The thick vertical line indicates the observed six-point likelihood on the data set.

5. Future directions

There are several directions in which **MLDS** might be developed, four of which will be mentioned here. First, it would be interesting to allow a formula expression as an argument to `mlds`. The current fitting procedures ignore the ordering of the physical scale. With both methods, each scale value is treated as an independent covariate and we do not force scale values to be increasing. In particular applications, the experimenter may wish to fit observers' performance by difference scales drawn from a particular parametric family such as power functions. A formula interface would permit functions of just a few parameters to be evaluated as candidates for the difference scale. For example, Figure 5b suggests that a simple parametric power function of r^2 might be found to describe the observer's judgments with many fewer than the 10 parameters used in the fit.

Second, we do not know what would be the most efficient choice of stimuli along a continuum or of quadruples for a particular application of MLDS. These depend on the observer's actual scale and judgment uncertainty σ, but given pilot data for the observer or previous results for other observers, it would be interesting to work out methods for selecting stimuli and quadruples. For example, having seen the results for one observer in the scatterplot example, we might consider stimuli that are equally spaced along the scale r^2 and not the scale r in future experiments.

Third, we plan on developing a more systematic method of assessing the asymptotic probabilities of the inverse link function to take into account unsystematic errors by the observers. The difficulty is that these parameters are not part of the linear predictor. One possibility is to profile the nuisance parameters (as we did here) or, alternatively, to develop a method that switches back and forth between adjusting the nuisance parameters and the coefficients of the linear predictor.

Fourth, it would be useful to incorporate random effects that influence the scale when an observer repeats the experiment or to account for variations between individuals. Such heterogeneity is, indeed, apparent if we compare the three scales obtained on different days in the data set `kk`. For these data, there is only one random factor, the Run. It might be possible to treat this as an effect due to a randomized block (Venables and Ripley 2002, p. 295) The ratio of the scale values between any of the two repetitions is approximately constant across the physical scale, however, which suggests that the estimate of σ across runs, or equivalently the maximum scale value, would be a more likely candidate to explain such a source of variability, but as a multiplicative rather than as an additive effect.

Computational details

The bivariate point distributions were generated with the `mvrnorm` function from the package **MASS** 7.2-40 (Venables and Ripley 2002). The **ggplot** package 0.4.2 (Wickham 2007) was used to create Figure 6. An R script replicating the figures is available along with this paper

Acknowledgments

This research was funded in part by Grant EY08266 from the National Institute of Health (LTM).

References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In BN Petrov, F Csàki (eds.), "Second International Symposium on Inference Theory," pp. 267–281. Akadémia Kiadó, Budapest.

Block HD, Marschak J (1960). "Random Orderings and Stochastic Theories of Responses." In I Olkin, S Ghurye, W Hoeffding, W Madow, H Mann (eds.), "Contributions to Probability and Statistics," pp. 38–45. Stanford University Press, Stanford.

Boschman MC (2001). "**DifScal**: A Tool for Analyzing Difference Ratings on an Ordinal Category Scale." *Behavioral Research Methods, Instrumuments & Computers*, **33**(1), 10–20.

Charrier C, Maloney L, Cherifi H, Knoblauch K (2007). "Maximum Likelihood Difference Scaling of Image Quality in Compression-degraded Images." *Journal of the Optical Society of America A Optics, Image Science and Vision*, **24**, 3418–3426.

Cleveland WS, Diaconis P, McGill R (1982). "Variables on Scatterplots Look More Highly Correlated When the Scales are Increased." *Science*, **216**, 1138–1141.

Cleveland WS, McGill R (1984a). "Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods." *Journal of the American Statistical Association*, **79**, 531–554.

Cleveland WS, McGill R (1984b). "The Many Faces of a Scatterplot." *Journal of the American Statistical Association*, **79**, 807–822.

Efron B, Tibshirani RJ (1993). *An Introduction to the Bootstrap*. Chapman Hall, New York.

Green DM, Swets JA (1974). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hauck Jr WW, Donner A (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis." *Journal of the American American Statistical Association*, **72**, 851–853.

Knoblauch K (2007). ***psyphy**: Functions for Analyzing Psychophysical Data in R*. R package version 0.0-5, URL http://CRAN.R-project.org/package=psyphy.

Krantz DH, Luce RD, Suppes P, Tversky A (1971). *Foundations of Measurement (Vol. 1): Additive and Polynomial Representation*. Academic Press, New York.

Legge GE, Gu YC, Luebker A (1989). "Efficiency of Graphical Perception." *Perception & Psychophysics*, **46**, 365–374.

Maloney LT, Yang JN (2003). "Maximum Likelihood Difference Scaling." *Journal of Vision*, **3**(8), 573–585. URL http://www.journalofvision.org/3/8/5/.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall, London.

Mood A, Graybill FA, Boes DC (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York, 3rd edition edition.

Obein G, Knoblauch K, Viénot F (2004). "Difference Scaling of Gloss: Nonlinearity, Binocularity, and Constancy." *Journal of Vision*, **4**(9), 711–720. URL http://www.journalofvision.org/4/9/4/.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rhodes G, Maloney L, Turner J, Ewing L (2007). "Adaptive Face Coding and Discrimination Around the Average Face." *Vision Research*, **47**, 974–989.

Schneider B (1980a). "Individual Loudness Functions Determined from Direct Comparisons of Loudness Intervals." *Perception & Psychophysics*, **28**, 493–503.

Schneider B (1980b). "A Technique for the Nonmetric Analysis of Paired Comparisons of Psychological Intervals." *Psychometrika*, **45**, 357–372.

Schneider B, Parker S, Stein D (1974). "The Measurement of Loudness Using Direct Comparisons of Sensory Intervals." *Journal of Mathematical Psychology*, **11**, 259–273.

Suppes P (1972). "Finite Equal-Interval Measurement Structures." *Theoria*, **38**, 45–63.

Thurstone LL (1927). "A Law of Comparative Judgement." *Psychological Review*, **34**, 273–286.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edition.

Wichmann FA, Hill NJ (2001). "The Psychometric Function: I. Fitting, Sampling and Goodness of Fit." *Perception & Psychophysics*, **63**, 1293–1313.

Wickham H (2007). ***ggplot**: An Implementation of the Grammar of Graphics in R*. R package version 0.4.2, URL http://CRAN.R-project.org/package=ggplot.

Wood S (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton.

Yang J, Szeverenyi N, Ts'o D (2008). "Neural Resources Associated with Perceptual Judgment Across Sensory Modalities." *Cerebral Cortex*, **18**, 38–45.

Affiliation:

Kenneth Knoblauch
Inserm, U846, Bron, France
Institut Cellule Souche et Cerveau
Département Neurosciences Intégratives
Université de Lyon, Lyon 1
18 avenue du Doyen Lépine
69500 Bron, France

Telephone: +33/472913477
Fax: +33/472913461
E-mail: knoblauch@lyon.inserm.fr
URL: http://www.sbri.fr/

Laurence T. Maloney
Department of Psychology
Center for Neural Science
New York University
6 Washington Place, 8th Floor
New York, NY 10011, United States of America
Telephone: +1/212/9987851
E-mail: ltm1@nyu.edu
URL: http://www.psych.nyu.edu/maloney/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 25, Issue 2 *Submitted:* 2007-04-17
March 2008 *Accepted:* 2008-01-25