CHAPTER 1

Measurement

Rajeev Kohli
Columbia University

Objective and Scope

This chapter introduces Steven's hierarchy of measurement scales and invariant transformations. It introduce scaling problems, starting with Thurstone's (1927) "law of comparative judgment," and provides an understanding of ordinal regression, nonmetric MDS, external analysis of preferences and internal analysis (unfolding) models. The material is presented in an informal manner, with a focus on providing an intuitive understanding of basic concepts. Exercises serve to further the understanding of the material and are pointers to further work in the area.

Types of Scales

In casual conversations, we talk about how hot or cold it is, how much we liked a movie, and how good a meal tasted. Mothers take the temperature of their sick babies, and nurses measure the weight and blood pressure of patients. Traffic cops record the speed of passing cars. And scientists in every field make all kinds of measurements using more or less sophisticated instruments. In each case, the purpose of measurement is to record certain types of information.

In the 1930's and 1940's, a psychologist by the name of S.S. Stevens observed that although there are a vast number of instruments that are used to measure all sorts of things, the information content of measurements must be independent of the units of the measurements themselves. Thus, the information about someone's height should not depend on whether we measure it in inches, feet or meters. Similarly, the information about the speed of a car or shooting star should not depend on whether we measure it in miles per hour of meters per second. The question is: what is this information?

Steven's answer was that the information in a measurement is not the value we record, but the properties of the values that do not change when we use another unit of measurement. He observed that the most common types of measurements fall into four classes. He called these classes ratio scales, interval (or cardinal) scales, ordinal scales, and nominal scales.

Ratio scales describe measurements for which we can compute ratios. That is, the units used for measurements must be such that the ratio of a pair of measurements does not depend on the units used. Height, weight and frequency are examples of quantities measured on ratio scales. For example, we can calculate the ratio of 2 ft. and 4 ft., and the value of this ratio does not change if we measure the same heights in inches, centimeters or any other unit of length. More generally, if x and y are two scales of measurement, then $x_1/x_2 = y_1/y_2$ for any pair of measurements that have values x_1 and x_2 on one scale, and y_2 and y_2 on the other scale. Let $y = f(x)$. Then

$$\frac{y_1}{y_2} = \frac{f(x_1)}{f(x_2)} = \frac{x_1}{x_2}.$$

We can re-write this expression as

$$\frac{y_1}{x_1} = \frac{y_2}{x_2}.$$

As x_1 and x_2 are any pair of arbitrary values of x, this implies that only those transformations $y = f(x)$ are permitted for which y/x is a constant for all values of x. That is, y is a permissible transformation of a ratio scaled variable x if

$$y = bx, \ b \neq 0.$$

A a necessary condition for a measurement to be on a ratio scale is that it allow multiplicative transformation. It is also sufficient, because if $x_1/x_2 = k$ for any two values of x, then $y_1/y_2 = k$ when $f = bx$, for any positive value of b. We say that a ratio scale measurement is *invariant to multiplicative transformations.*

A measurement is *interval scaled* if the *differences* in the values of measurements are ratio scaled. Thus, if

$$\frac{x_1 - x_2}{x_3 - x_4} = k,$$

for measurements on a scale x, then y is an equivalent scale of measurement if

$$\frac{y_1 - y_2}{y_3 - y_4} = k.$$

The transformation $y = a + bx, b \neq 0$, is permissible because

$$\frac{y_1 - y_2}{y_3 - y_4} = \frac{(a + bx_1) - (a + bx_2)}{(a + bx_3) - (a + bx_4)} = \frac{x_1 - x_2}{x_3 - x_4} = k.$$

It also the only transformation leaves the ratio of differences unchanged. To see this, we rewrite the above expression as

$$\frac{y_1 - y_2}{x_1 - x_2} = \frac{y_3 - y_4}{x_3 - x_4},$$

for $x_i, y_i, i = 1, \ldots, 4$. Consider $x_2 = x_1 + \Delta x_1$ and $x_4 = x_3 + \Delta x_3$, with corresponding values $y_2 = y_1 + \Delta y_1$ and $y_4 = y_3 + \Delta y_3$. Then

$$\frac{\Delta y_1}{\Delta x_1} = \frac{\Delta y_3}{\Delta x_3}.$$

Taking the limits $\Delta x_1 \to 0, \Delta x_3 \to 0$, gives

$$\left.\frac{dy}{dx}\right|_{x_1} = \left.\frac{dy}{dx}\right|_{x_3}.$$

As x_1 and x_3 are any arbitrary values of x, we have

$$\frac{dy}{dx} = \text{constant},$$

which implies that y is of the form $y = a + bx, b \neq 0$. That is, y and x are equivalent interval scales if y is a linear function $y = a + bx, \ b \neq 0$. We say that y and x are related by an affine transformation, or that they are invariant to affine transformations. Observe that a ratio scale permits a subset of the transformations allowed for an interval scale, which is no accident because if you can take ratios of measurements you can also take the ratios of differences of measurements. So a ratio scale has more information than an interval scale but fewer permissible transformations than an interval scale.

REMARK. You are probably familiar with the Fahrenheit (F), Celsius (C) and Kelvin (K) temperature scales. The first two are interval scales. The Kelvin scale is a ratio scale. Nothing can have a lower temperature than 0 K, which is the point at which all motion ceases. Which scale we use depends on why we use it. For everyday use — taking body temperature, or measuring ambient temperature — we have a sense of how the temperature on a scale is related to well being, or hot and cold weather. However, scientists sometimes do need to measure temperature on a ratio scale, and then we need very precise definitions. In 2011, more than a century after the Belfast-born engineer Lord Kelvin wrote of the need for an "absolute thermometric scale", the General Conference on Weights and Measures proposed to define the kelvin as the unit of thermodynamic temperature; its magnitude is set by fixing the numerical value of the Boltzmann constant to be equal to exactly 1.38065×10^{-23} when it is expressed in the unit $\mathrm{s}^{-2} \cdot \mathrm{m}^2 \cdot \mathrm{kg}\ \mathrm{K}^{-1}$. One consequence of this change is that the new definition makes the definition of the kelvin depend on the definitions of second (s), the metre (m), and the kilogram (kg).

Ordinal and *nominal* scales have progressively lesser information than an interval scale. They permit a larger set of transformations than an interval scale. Measurements on an ordinal scale allow rankings and on a nominal scale allows discrimination but not rankings. Ordinal measurements are invariant to an *increasing, monotonic transformation*: x and y are equivalent ordinal measures if $x_1 > x_2$ implies $y_1 > y_2$, and vice versa. If x is a continuous ordinal measure, then so is and function $y = f(x)$ provided $dy/dx > 0$. Note that this includes as a special case the affine transformation $y = a + bx$.

Measurements on a nominal scale are invariant to transformation that preserve differences in values: x and y are equivalent scales if $x_1 = x_2$ implies $y_1 = y_2$ and $x_1 \neq x_2$ implies $y_1 \neq y_2$. Thus, any one-to-one mapping from y to x is a permissible transformation for a nominal scale.

These four types of measurements are the most common, but there are others. An ordered semi-metric allows an ordering of differences of measurements; it lies "between" an interval scale and an ordinal scale. A partially-ordered scale allows weak orderings $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$; it lies between ordinal and nominal scales.

EXERCISE 1. Let p_i denote the probability that item j, with utility u_j, is selected from a set of n items. Identify the measurement scales for the utilities in each of the following cases:

$$(a)\ p_i = \frac{u_i}{\sum_{j=1}^n u_j}$$

$$(b)\ p_i = \frac{e^{u_i}}{\sum_{j=1}^n e^{u_j}}$$

$$(c)\ p_i = \frac{e^{u_i/\sigma(u_i)}}{\sum_{j=1}^n e^{u_j/\sigma(u_j)}},\ \text{where } \sigma(u_j) = \sigma(u), 1 \le j \le n,$$

and $\sigma(ku) = k\sigma(u)$.

Some Implications

Central tendency. Let $x_1, x_2, x_3, \ldots, x_n$ denote n observations. Let us consider the measurement scale necessary for each of the following five measures of central tendency.

(a) The *mode* is the most frequent observation. It can be computed for measurements with at least nominal information.

(b) The *median* is the middle number; it can be calculated if the measurement scale is at least ordinal.

(c) The *arithmetic mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$ requires at least interval-scaled data.

(d) The *geometric mean*

$$G = \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}},$$

can be calculated only if the data are ratio scaled. As an example, Nash's bargaining model predicts that if $u_1(x)$ and $u_2(x)$ are the utility functions for two individuals over an amount x of a resource, and if x_{1d} and x_{2d} are the disagreement points for the two individuals, then the outcome of a bargain between them is a division of x that maximizes the geometric mean $G = \sqrt{u_1 u_2}$, where $u_1 = u_1(x) - u_1(x_{1d}), u_2 = u_2(x) - u_2(x_{1d})$. Note that for the differences u_1 and u_2 to be ratio scaled, one requires interval-scaled utility functions $u_1(x)$ and $u_2(x)$. As another example, suppose \$1 has value $x_i = 1 + r_i$ after a year if invested in stock i. Then the average value per dollar invested in n stocks is their geometric mean.

(e) The *harmonic mean*

$$H = 1 \Bigg/ \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{x_i}\right),$$

requires observations on a ratio scale. For example, suppose a driver travels a distance d at a speed of x_1 mph. On the way back, he covers the same distance at a speed of x_2 mph. Let t denote the total travel time and s the average driving speed. Then

$$t = \frac{d}{x_1} + \frac{d}{x_2} = \frac{2d}{H(x_1, x_2)}$$

and

$$s = \frac{2d}{t} = H(x_1, x_2).$$

Corrrelation. The measure of association between a pair of variables, x and y, depends on their measurement scale. The Pearson correlation is an appropriate measure of linear association if x and y are both interval or ratio measures, but not if they are ordinal measures; the appropriate measures in the latter case are Spearman rank order correlation and Kendall's tau. If x and y are nominal variables, Cramer's V and ϕ, which are based on the chi-square value for a contingency table, can be used; one can also use λ, Goodman and Kruskal's τ and the uncertainty coefficient, all of which rely on a measure of proportional reduction in error.

Linear regression. Let

$$y = \alpha + \beta x + \epsilon,$$

where ϵ is an error term. If y is a ratio-scaled variable, then α is unique up to a multiplicative constant that depends on the scale of y; ϵ has zero mean; and β is unique up to a multiplicative constant, but its value depends on the scales for both y and x. For example, suppose y is sales and x is price. Then α is the expected (value of) sales when price is zero, and its value depends only on the units of sales measurement (kgs. or lbs. or oz.); β is the expected change in sales for a unit change in price, and its value depends on the scales used to measure not only y but also x (Dollars, Dinars). Sometimes, we know *a priori* the value of the intercept and can impose it as a constraint. For example, if y is sales and x is hours since the opening of a store on a given day (9:00 am, say), then we can set $\alpha = 0$. Now suppose y is ratio scaled and x is interval scaled. That is, we can replace x by $x^{'} = a + bx$. An example is when we want to include a trend in a sales model. That is, y is unit sales, and x is the time (say week) in which a certain volume of a product is sold. There is no reason why we should measure trend in weeks starting with 1; we could set the first week of sales data to any value a; we could also specify the trend in weeks or days. If we substitute $x = (x^{'} - a)/b$ in the regression equation, we get

$$\begin{aligned} y &= \alpha + \beta \frac{x^{'} - a}{b} + \epsilon, \\ &= (\alpha - \frac{a\beta}{b}) + \frac{\beta}{b} x^{'} + \epsilon \\ &= \alpha^{'} + \beta^{'} x^{'} + \epsilon. \end{aligned}$$

Thus, if y is ratio scaled and x is interval scaled, then α is interval scaled and β is ratio scaled.

Exercise 2. What are the scale properties of α and β in the equation

$$y = \alpha + \beta x + \epsilon,$$

if (a) x is ratio scaled and y is interval scaled, and (b) both x and y are interval scaled.

Exercise 3. If you ask consumers leaving a supermarket what prices they paid for the products they purchased, you will find that their recall is quite inaccurate. Suppose the reported price P_i for product i depends on the actual price p_i and the price p of a (customer specific) reference brand. We hypothesize the following relationship:

$$P_i = \alpha p_i + (1-\alpha)\frac{p_i}{p} + \epsilon_i,$$

where $0 \le \alpha \le 1$. Suppose we collect appropriate data from buyers exiting a supermarket, run a constrained regression and get a statistically significant estimate of $\hat{\alpha} = 1/2$. What conclusion(s) can we draw?

Scaling

Scaling refers to a collection of methods for drawing inferences about interval or ratio scales from ordinal or nominal data. The literature on the subject is now so large and varied that we can cover only the basics. We will examine Thurstone scaling; conjoint analysis and ordinal regression; and metric and nonmetric multidimensional scaling.

Thurstone Scaling. Suppose u is an interval-scaled variable. We take the difference $y = u_i - u_j$ of a pair of values u_i, u_j, add some noise ϵ, and say that stimulus i is larger than stimulus j if $y + \epsilon > 0$, otherwise stimulus j is larger than stimulus i. Suppose we repeat this process n times, the only difference from one time to another being that we add a new error term to y. If ϵ are independent, identically distributed draws from a standard normal distribution $N(0,1)$, then $y + \epsilon$ has a normal distribution with mean y and unit variance. The probability $p = p(y + \epsilon > 0)$ then depend on the difference y. For example, if $y = 0$, then $p = p(\epsilon > 0) = 1/2$; if $y = -1.96$, then $p = p(\epsilon > 1.96) = 0.05$, because 1.96 is the value of the standard normal variate for which the area to the right hand side of the normal curve is 5%. Conversely, suppose we know the value of p. Because there is a one-to-one mapping from y to p, we can infer the value of y : if $p = 1/2$, then $y = 0$; if $p = 0.05$, then $y = -1.96$.

Let's see how we can use this to go from ordinal comparisons to interval-scaled utilities. Let u denote utility ("liking for a product"), which we assume is unobserved but interval scaled. Suppose you have a choice between going to an Italian restaurant (i) or a Japanese restaurant (j) for lunch. Let $\hat{p}$ denote the proportion (fraction) of visits you make to the Japanese restaurant over a period of a year. Suppose $\hat{p} = 0.05$. If I assume that the error ϵ is normally distributed with mean zero and variance 1, then I can say that $\hat{y} = \hat{u}_i - u_j = -1.96$.

We can generalize the above development. Let u_i and u_j denote the utilities of products i and j. The utilities can vary from one time to another, in a random fashion. Suppose u_i is normally distributed with mean μ_i and variance σ_i^2. Let ρ denote the correlation between u_i and u_j; it is a measure of similarity or dissimilarity of products i and j. Let p_{ij} denote the proportion of times item i is preferred to item j; let z_{ij} denote the value of the unit normal distribution for which the area to the right of the curve is p_{ij}. Then

$$\mu_i - \mu_j = z_{ij}\sqrt{\mathrm{var}(u_i - u_j)} = z_{ij} \cdot \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Thurstone proposed this "law of comparative judgment" in 1927.[1] It suggests that we can make inferences about a preference scale by observing uncertain choices.

Thurstone noted that his model could not be used in the general form above: it is one equation with more than one unknown variable. We need judgments for more than one pair of stimuli and ensure that there are at least as many equations as the number of parameters. However, there are special cases of the model in which repeated judgments for one pair of stimuli (e.g., a pair of sounds evaluated for their loudness) can be sufficient. If we assume that $\sigma_i = \sigma, \rho = 0$, we have

$$\mu_i - \mu_j = z_{ij} \cdot \sigma\sqrt{2}.$$

As the unobserved scale values are interval scaled, we can divide both sides by a constant, say $\sigma\sqrt{2}$ (which is unknown but still a constant). This gives

$$\mu_{ij} = \frac{\mu_i - \mu_j}{\sigma\sqrt{2}} = z_{ij}.$$

We can also set — again because of the interval scale — $\mu_j = 0$; then the value of the standard normal variate z_{ij} gives the scale value for item i. This is Thurstone's famous "Case V." It is the only case for which the proportion p depends monotonically on inter-mean distances, and this may have been the source of the idea of psychological distance and so ultimately of multidimensional scaling.[2] Later, people began to work with other distributions than the normal. Logit models, for example, use the double exponential as the error distribution.

Conjoint Analysis. The origins of conjoint analysis are in a research stream originated by psychologists (e.g., Suppes, Krantz and Tversky), who proposed methods for testing unobserved relationships among psychological variables using only the ordinal properties of measurements. For example, suppose we ride a bicycle up a hill. The ease (E) with which we go up depends on, say, our riding ability (A), strength (S) and fitness (F). We may think of these three variables as unobserved psychological constructs, possibly consider them to be interval scaled. But it is not clear if we can assume any specific, functional relationship between the constructs and their measurements. On the other hand, we may reasonably assume that there is at least an ordinal relationship between the constructs and their measurements. The question is: can we conclude, based on these ordinal measurements, which of the possible unobserved relationships among the three variables are consistent with data we might collect? It turns out that different models — e.g., $E = A(F + S)$ or $E = EFS$ or $E = F(S + A)$ — make different ordinal predictions under certain conditions. Sometimes, the ordinal predictions are the same for two of the three models but not for a third. Axiomatic conjoint measurement aims to test alternative composition models by examining only these ordinal consequences.

In marketing, the term "conjoint analysis" refers to a set of related techniques for collecting data and estimating consumer preference functions defined over multiple attributes. Instead of testing for alternative model structures, one wishes to estimate the parameters of one or another preference model using data on individual preferences. The data can be ordinal. One seeks estimates of the unobserved

[1] Thurstone, L.L. (1927), "A Law of Comparative Judgment," *Psychological Review,* 34, 273-286; reprinted in the same journal in 1994, Volume 101 (2), pp. 266-270.

[2] See footnote 4 in Luce, R.D. (1994), "Thurstone and Sensory Scaling: Then and Now," *Psychological Review,* 101 (2), 271-277.

preference function using these data. Sometimes, the data are choices from sets of items; here, too, one wishes to estimate the parameters of an unobserved preference function that, in some sense, the best predictor of the observed choices. Typically, conjoint analysis focuses on individual-level estimates of preference functions. Sometimes, however, one is limited by the amount of data that can be obtained from a respondent, and so segment-level estimates are obtained, using so-called “latent-class’ methods. The segments in these cases are not observed, but are inferred from the data. Such segments are called benefit segments, because they represent groups of consumers with similar preferences.

A distinguishing feature of conjoint analysis is its use of experimental design for generating product product profiles. This allows for orthogonal parameter estimates in a regression context (but not necessarily in other contexts). It also reduces the number of profile evaluations per respondents. The first is useful because one obtains “uncorrelated” parameter estimates, the second because respondent fatigue and involvement are important factors in data collection. The appendix to the chapter describes the basics of fractional factorial designs, which are used in collecting conjoint data.[3]

Let x denote an attribute and y a preference score. Then

$$y = \beta_0 + \beta_1 x$$

is called a “vector” model; and

$$y = \beta_0 + \beta_1 (x - x^*)^2$$

is called an ideal-point model. A vector model is often used to represent preferences for an attribute for which more is always better (or worse). Examples are the safety or gas milage of a car. An ideal-point model is better suited for representing preferences over an attribute for which there is some (unobserved, person specific) ideal level of an attribute that an individual desires. Example are the amount of sugar in, and the temperature of, a beverage. We can write the ideal-point model as

$$y = \beta_0 + \beta_1 x^2 - 2\beta_1 x x^* + \beta_1 {x^*}^2$$

or

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

where

$$\alpha_0 = \beta_0 + \beta_1 {x^*}^2,\ \alpha_1 = -2\beta_1 x^*;\ \text{and}\ \alpha_2 = \beta_1.$$

If we obtain estimates of α_0, α_1, and α_2, then we can estimate the values of β_0, β_1, and the ideal point $\hat{x}^*$ as follows:

$$\hat{\beta}_1 = \hat{\alpha}_2, \hat{x}^* = -\frac{\hat{\alpha}_1}{2\hat{\alpha}_2},\ \text{and}\ \hat{\beta}_0 = \hat{\alpha}_0 - \hat{\beta}_1 \hat{x^*}^2 = \hat{\alpha}_0 + \frac{\hat{\alpha}^2}{4\hat{\alpha}_2}.$$

Thus, to estimate an ideal point model of preferences, it is sufficient to estimate a preference model with both a linear term x and a quadratic term x^2. We can thus construct preference models with both vector and ideal-point types of attributes in them.

The most common type of conjoint model is called a “part-worths” model, in which each attribute has a discrete number of values, called its levels. For example,

[3] For recent work on interactive methods of data collection, please see Toubia, Simester and Hauser (2002).

30 mpg, 25 mpg and 20 mpg are three levels of a discretized attribute, gas milage. Let n denote the number of attributes and n_k the number of levels of attribute $k = 1, \ldots, n$. Let i denote an alternative (stimulus) with one level of each attribute. Then a part-worths model has the form

$$y_i = \beta_0 + \sum_{k=1}^{n} \sum_{j=1}^{n_k - 1} \beta_{jk} x_{ijk}$$

where $x_{ijk} = 1$ if level j of attribute k appears in alternative i; otherwise, $x_{ijk} = 0$. Note that the level $j = n_k$ is represented by setting $x_{ijk} = 0$ for all levels $j = 1, \ldots, n_k - 1$ of attribute k. To illustrate, consider $n = 2$ attributes, each at three levels. Suppose an alternative, say $i = 1$, has level $j = 2$ of attribute 1, and level $j = 3$ of attribute 2. Then we have

$$y_1 = \beta_0 + \beta_{21} + \beta_{32},$$

where β_{21} is the term corresponding to level 2 of attribute 1 and β_{32} is the term corresponding to level 3 of attribute 2. Similarly,

$$y_2 = \beta_0 + \beta_{11},$$

if alternative $i = 2$ has level 1 of attribute 1 and level 3 of attribute 2. The β_{jk} terms are called “part worths.” The alternatives are called profiles, or in the case where these are products, product profiles. We will use the terms alternatives and profiles interchangeably.

Suppose y is interval scaled. Then β_{jk} are ratio-scaled values (why?). Suppose also that we ask a person to rank order a set of alternatives, assigning a lower rank r to a less-preferred alternative; that is, $r = 1$ for the least-preferred alternative, and $r = n$ for the most preferred alternative. We wish to use the ordering of profiles to estimate the parameters β_{jk}, for all the levels $j = 1, \ldots, n_k - 1$, of all the attributes $k = 1, \ldots, n$. There are several ways of doing so. We consider a method called LINMAP; another method is discussed in Appendix 2.

Suppose we have obtained a preference ranking from a person over a set of m alternatives, each of which is described using n continuous attributes, $x_1, \ldots, x_n$. Let x_{ik} denote the value of attribute k in profile i, for all attributes $k = 1, \ldots, n$, and all profiles $i = 1, \ldots, m$. We assume that there is an unobserved, interval-scaled preference function (i.e., cardinal utility function)

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

We construct the set Ω of all pairs (i, j) so that $y_i \geq y_j$. For example, if there are $m = 3$ alternatives, and if $y_1 > y_2 > y_3$, then $\Omega = \{(1, 2), (1, 3), (2, 3)$. Alternatively, we might direct collect preference data on pairs of alternatives. We then solve the problem

$$\begin{aligned} \text{Minimize} \quad & \sum_{(i,j)\in\Omega} e_{ij} \\ \text{subject to} \quad & y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}, \quad i = 1, \ldots, m \\ & y_i - y_j + e_{ij} \geq 0, \quad \text{for all } (i,j) \in \Omega \\ & \sum_{(i,j)\in\Omega} y_i - y_j = 1 \\ & y_i \geq 0, i = 1, \ldots, m. \end{aligned}$$

The constraint

$$\sum_{(i,j)\in\Omega} y_i - y_j = 1$$

is a normalization of the parameter estimates; it also ensures that the LP does not obtain the trivial solution $\beta_1 = \cdots = \beta_n = 0$. One can replace it by any other convenient normalization constraint on the parameter values; for example, we can use $\beta_1 + \cdots + \beta_n = 1$.

We can simplify the above formulation by replacing the first two constraints by

$$\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_n(x_{in} - x_{jn}) + e_{ij} \geq 0, \quad \text{for all } (i,j) \in \Omega,$$

and the third constraint by

$$\sum_{(i,j)\in\Omega} \beta_1(x_{i1} - x_{j1}) + \cdots + \beta_n(x_{in} - x_{j1}) = 1.$$

We thus rewrite the inference problem as follows:

$$\begin{aligned} \text{Minimize} \quad & \sum_{(i,j)\in\Omega} e_{ij} \\ \text{subject to} \quad & \sum_{k=1}^{n} \beta_k(x_{ik} - x_{jk}) + e_{ij} \geq 0, \quad \text{for all } (i,j) \in \Omega \\ & \sum_{(i,j)\in\Omega} \sum_{k=1}^{n} \beta_k(x_{ik} - x_{jk}) = 1 \\ & y_i \geq 0, i = 1, \ldots, m. \end{aligned}$$

This is a linear-programming (LP) problem. It can be solved quite easily (i.e., in a computationally efficient manner) to obtain estimates of the β parameters using standard algorithms for solving LP problems.

Exercise 4. Formulate a version of LINMAP that estimates the part worths from a rank ordering of m product profiles described over n discrete attributes, where attribute k has n_k levels.

Exercise 5. Let C denote a choice set with $|C|$ profiles, each defined over m continuous attributes. Assume that there is an unobserved, cardinal utility function

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

Suppose a person selects an alternative $i \in C$ if it has the highest utility y_i among all items in C. Let $C_1, \ldots, C_p$ denote p choice sets. Suppose we record a consumer's choices over each of the p choice sets. Formulate a version of the LINMAP model that estimates the parameters of the above utility function.

Conjoint choice experiments refers to a particular method for generating choice sets. First, profiles are generated using an experimental plan. Second, each of these profiles enters as a binary (present/absent) factor in another experimental design, the treatments of which define choice set containing the subset of "present" items. The respondents' task is to select one item from each choice set. The parameter estimates are typically obtained using a logit/probit model, which Prof. Ansari will discuss later in the course. The parameter estimates are then obtained by maximizing a likelihood function, which gives the joint probability of all choices being made across all the choice sets.

Proponents of conjoint choice experiments highlight their greater realism. Critics of the method point to the increased complexity of judgments and greater data requirement for estimation. Greater complexity can occur when there are choice sets with many (say, 7+) product profiles and each profile is described over a large (say, 10+) product attributes; such situations frequently occur in applied settings. The increase in the amount of data refers to the need for a larger number of choice-set evaluations to obtain the same number of pairwise comparisons as from a ranking of a set of profiles. Conjoint studies for consumer goods typically collects data from $400-500$ participants when the results need to be projected to the entire US. Commercial studies are seldom performed with fewer than 100 respondents.

Remark. The linear preference model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots,$$

has some interesting special cases. Suppose $x_i = 0, 1, 2, 3, \dots, n_i - 1$, is a discrete variable, $n_i \geq 2$. Consider

$$y = \frac{x_1}{n_1} + \frac{x_2}{n_1 n_2} + \frac{x_3}{n_1 n_2 n_3} + \dots.$$

Then y represents lexicographic preferences over the attributes. Note that for $n_i = n \geq 2,\quad 1 \leq i \leq n$, we have

$$y = \frac{x_1}{n} + \frac{x_2}{n^2} + \frac{x_3}{n^3} + \dots,\ 0 \leq x_i \leq n-1,$$

which is just a fractional number between 0 and 1 in radix (base) n. For example, if each attribute has 10 possible values, then

$$y = \frac{x_1}{10} + \frac{x_2}{100} + \frac{x_3}{1000} + \dots + \frac{x_n}{10^n},\ 0 \leq x_i \leq 9,$$

is a lexicographic utility function, where attribute i is preferred to attribute $i+1$, and a higher value of x_i corresponds to a preferred attribute level. This is an "ordinary" number between 0 and 1. To see that this is a lexicographic function, note that if we replace $0, 1, 2, \dots, 9$ by the first ten letters $a, b, c, \dots, j$, of the English alphabet, then each alternative is a word in a dictionary of n-letter words formed using these ten letters of the alphabet; and a lexicographic arrangement of the words is a lexicographic arrangement of the alternatives.

Another special case concerns certain types of "logical" preferences. Let x_i denote a $0-1$ variable representing the presence or absence of a "level" (value) of some attribute. An alternative is defined as some combination of the levels of the attributes; for example, consider two attributes, each at two levels. we associate the $0-1$ variables x_1 and x_2 with the levels of attribute 1; and x_3 and x_4 with the levels of attribute 2. Each sequence $x_1 - x_2 - x_3 - x_4$ defines an alternative, where $x_1 + x_2 = 1$ and $x_3 + x_4 = 1$. In general, suppose we represent alternatives using n such $0-1$ variables, $x_1, x_2, x_3, \ldots, x_n$. Consider the preference function

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

Suppose we restrict β_i to $0-1$ values for all $1 \leq i \leq n$. Let $1 \leq k \leq n$ be an integer. Consider

$$z = \begin{cases} 1 & \text{if } y > k; \\ 0 & \text{otherwise;} \end{cases}$$

where $z = 1$ indicate that an alternative is acceptable to a person. Then $k = 1$ corresponds to a disjunctive evaluation of an alternative; and $k = n$ corresponds to a conjunctive evaluation of an alternative. More generally, each k describes a subset-conjunctive rule. Such rules are often used in medical diagnosis. For example, if you are often thirsty ($x_1 = 1$), urinate frequently ($x_2 = 1$), have blurry vision ($x_3 = 1$) and get tired quickly ($x_4 = 1$), then you have diabetes. There is no definitive test for the disease, but the conjunctive pattern is so often seen that it is commonly used for diagnosing the disease. The measurement properties of y for both the lexicographic and subset-conjunctive cases do not correspond to any of the four scales we have discussed. They lie between ordinal and interval scales. In the lexicographic model, as the number of attributes go to infinity, the set of y values corresponds to the set of all real numbers between zero and one; so the scale properties depend on the number of attributes, taking interval values in the limiting case.

Multidimensional scaling. We begin by considering the problem of finding the location of m points in n-dimensional space. We are to find these locations from an ordering of distances between pairs of points in the space. That is, if d_{ij} is the distance between points i and j, and d_{kl} is the distance between points k and l, then we know which of the pair of distances is larger, but not by how much larger. For example, suppose the points represent brands and a consumer believes that brands i and j are more similar than brands k and l. We can then envision a perceptual map in in which brands i and j are closer than brands k and l. The problem in which we infer the relative locations of points on a map given ordinal similarity judgments is called nonmetric MDS. Later, we will consider the problem where the distances are known, but not the point locations; this method is called metric MDS.

Nonmetric MDS. Let d_{ij} denote the distance between a pair of points, i and j, in n-dimensional space. The Euclidean distance is

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2}$$

Suppose the distance between the pair i, j is smaller than the distance between k, l. That is, $d_{ij} < d_{kl}$. We write this as the ordered pair (ij, kl). Let Ω denote a set of all ordered pairs for which we have observations. This may be obtained, for example, by asking a subject to rank pairs of brands in a product category in decreasing order of similarity or dissimilarity. The pairswise comparisons can then be generated from this ranking data.

Consider the following constraint for each pair of points (ij, kl) :

$$d_{ij} + e_{ij,kl} \geq d_{kl} \text{ for all } (ij, kl) \in \Omega.$$

The point locations $x_1, \ldots, x_m$, are completely consistent with the ordinal data if the constraints are all satisfied when we set $e_{ij,kl} = 0$ for all $(ij, kl) \in \Omega$. We wish to find such locations, if these exist; otherwise, we wish to find their locations so that the total error — the sum of all $e_{ij,kl}$ values — is as small as possible. That is, we wish to solve the following problem:

$$\text{Minimize} \quad \sum_{(ij,kl)\in\Omega} e_{ij,kl}$$

subject to:

$$d_{ij} + e_{ij,kl} \geq d_{kl} \text{ for all } (ij, kl) \in \Omega.$$
$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2}$$
$$\sum_{ij,kl} d_{ij} - d_{kl} = 1$$
$$d_{ij} \geq 0.$$

As before, we use the constraint $\sum_{ij,kl} d_{ij} - d_{kl} = 1$ as a normalization of distances, and to rule out the trivial solution in which all points have the same location on each dimension. One can also work with the squared distances instead of distances:

$$\text{Minimize} \quad \sum_{(ij,kl)\in\Omega} E_{ij,kl}$$

subject to:

$$d_{ij}^2 + E_{ij,kl} \geq d_{kl}^2, \text{ for all } (ij, kl) \in \Omega.$$
$$d_{ij}^2 = (x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2$$
$$\sum_{ij,kl} d_{ij}^2 - d_{kl}^2 = 1.$$

We can rewrite this problem as

$$\text{Minimize} \quad \sum_{(ij,kl)\in\Omega} E_{ij,kl}$$

subject to:

$$E_{ij,kl} \geq \sum_{t=1}^{n} (x_{kt} - x_{lt})^2 - (x_{it} - x_{jt})^2, \text{ for all } (ij, kl) \in \Omega$$

$$\sum_{(ij,kl)\in\Omega} \sum_{t=1}^{n} (x_{it} - x_{jt})^2 - (x_{kt} - x_{lt})^2 = 1.$$

This is a non-linear optimization problem. You can use the PROC-NLP procedure in SAS to obtain solutions. An explicit algorithm, due to Kruskal, is described in Appendix 3. Typically, one seeks a solution in a small number of dimensions; scree plots, which show the poorness of fit against the number of dimensions, are commonly used to look for guidance in selecting the number of dimensions. The MDS solution gives only the relative locations of points; the locations can be moved to a different origin, and the distances can be stretched by multiplying each with a common constant. The directions of the axes are arbitrary, and so one needs other methods to give meaning to perceptual maps. If one has a rank ordering of the brands in terms of preferences, one can use the (vector and ideal-point) preference models described above to represent individual preferences on the map, or to show individual ideal points. Such an analysis is called external analysis, because the preference data are independent of (external to) the preference representation. An alternative method, called internal analysis (also called an unfolding model) uses preference data from multiple individuals to simultaneously find locations of alternatives and ideal points in a perceptual map. We briefly discuss this in Appendix 3.

Metric MDS. Suppose we know, or can estimate, distances (similarities) between pairs of stimuli. For example, we may ask subjects to indicate how similar pairs of items are on a rating scale and threat the ratings as ratio measurements; or we may compute the proportion of people who say two items are similar as a ratio-scaled measure of similarity. How do we find the locations of stimuli in a metric space that fit, as closely as possible, the known distances? This problem is called metric MDS. We describe below an inference method developed by Torgerson in 1952. Suppose we know the distances

$$d_{ij} = \sqrt{\sum_t (x_{it} - x_{jt}^2)}$$

between all pairs of points i, j in a n-dimensional space. Without loss of generality, assume that the point locations are mean-centered on each dimension:

$$\bar{x}_t = \sum_i x_{it} = 0.$$

Construct a square matrix D with m rows and columns; each row and column corresponds to a point in the n-dimensional space. The elements of D are the squared distance d_{ij}^2. Compute the row mean $d_{i.}^2$ for row i, the column mean $d_{.j}^2$ for column j, and the grand mean $d_{..}^2$. Construct the $n \times n$ matrix $\Delta = \{\delta_{ij}\}$ where

$$\delta_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2).$$

Torgerson showed that

$$\delta_{ij} = \sum_t x_{it} x_{jt},$$

and so

$$\Delta = XX^{'},$$

where X is the $n \times t$ matrix of stimulus coordinates. If Δ is positive semi-definite of rank r[4], then

$$\Delta = V \Lambda V^{'} = XX^{'}$$

where Λ is a diagonal matrix, denoted

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r).$$

[4] If Δ is not positive semi-definite, one adds a constant to each of the pairwise distances until it becomes so; this is called the “additive constant” problem.

The solution is given by

$$X = V\Lambda^{1/2},$$

where

$$\Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}, \ldots, \sqrt{\lambda_r}).$$

This way of decomposing Δ is called a spectral decomposition[5]; the Euclidean distances between the points are given by

$$d_{ij} = \sum_t \lambda_k (x_{it} - x_{jt})^2.$$

The λ_t are called eigenvalues, and one retains only those largest of these to get a solution in a few dimensions. Recall that the locations have been normalized to have zero mean; they also can be rotated, because this has no effect on distances.

After that, the notable work in the area is the development of INDSCAL by Carroll and Chang (1970), which allows individual differences in metric MDS by allowing person-specific weights for distances along different dimensions. That is, each person p has a squared distance

$$d_{ijp}^2 = \sqrt{\sum_t w_{tp}^2 (x_{it} - x_{jt}^2)}$$

between all pairs of points i, j in a n-dimensional space, where w_{tp} is the weight for dimension t and person p; and x_{it} is the common location of alternative i on dimension t. Thus, we have a common perceptual space, but we permit differential stretching of the dimensions across people. In matrix notation,

$$\Delta_p = X W_p^2 X^{'}$$

where $W_p^2 = W_p W_p^{'}$ is a diagonal matrix with w_{tp}^2 as the tth diagonal element. The solution is $X_p = XW_p$. As the weights are associated with dimensions, this model does not allow arbitrary orientation of the axes. Carroll and Chang (1972), Harshman (1972) and Tucker (1972) proposed three-mode MDS, in which the individual-differences model is extended to allow individual rotations of axes. In matrix notation,

$$X_p = X W_p T_p.$$

The rotation of the axes is done for each individual by the transformation matrix T_p.

Closing Remarks. The landmark paper on non-metric MDS is Shepherd (1962). Soon after that, Joseph Kruskal (1964), working at Bell Labs with Shepherd, developed an algorithm for non-metric MDS and opened the gates to a long and continuing line of work on multidimensional scaling.[6]

[5] i.e., a singular-value decomposition of a square matrix.

[6] A little later, Guttman (1968) proposed a different approach to nonmetric scaling.

Appendix 1

Fractional factorial designs. Let m denote the number of attributes and n_k number of levels of attribute $k, 1 \leq k \leq m$. Without loss of generality, let level n_k of attribute k denote a "reference" level. Let $x_{jk} = 1$ be an indicator variable associated with level j of attribute k; the reference level is then associated with $x_{jk} = 0$ for all $1 \leq j \leq n_k - 1$, $1 \leq k \leq m$. A "part-worths" model of preferences is:

$$u(x_1, x_2, x_3, \ldots, x_m) = \beta_0 + \sum_{k=1}^{m} \sum_{j=1}^{n_k - 1} \beta_{jk} x_{jk},$$

The model has $1 + \sum_k (n_k - 1)$ parameters, and so it requires at least as many observations (typically at least three times as many observations) to estimate. For the immediate discussion, suppose $n_k = 2, 1 \leq k \leq m$. Then we require at least $m+1$ data points, and prefer to have at least $3(m+1)$ observations, to estimate the part worths. On the other hand, there are $\prod_k n_k \geq 2^m$ distinct product profiles, a number that increases exponentially with m. Which of these product profiles should one use to get consumer preference judgments? One answer to the question is to select a subset of product profiles that are both believable and that form an orthogonal subset; i.e., a subset of product profiles that allow estimates of β_{jk} that are not correlated. This can be achieved by using fractional factorial plans. The simplest way of generating the plans when all attributes are binary is to start with a small main-effects plan and confound the highest-order interaction terms with the main effects. To illustrate, let one level of each attribute be represented by the symbol '-' and the other by the symbol '+'. Consider the four product profiles constructed by combining two attributes:

	Attribute	
Profile	A	B
1	$-$	$-$
2	$-$	$+$
3	$+$	$-$
4	$+$	$+$

We add a third binary attribute thus:

	Attribute		
Profile	A	B	C
1	$-$	$-$	$+$
2	$-$	$+$	$-$
3	$+$	$-$	$-$
4	$+$	$+$	$+$

You can probably see the relation between the + and - signs in the first two columns and those in the third column: they follow the composition rule (denoted $\oplus$):

$$\begin{array}{ccc} + \oplus + & \to & + \\ + \oplus - & \to & - \\ - \oplus + & \to & - \\ - \oplus - & \to & + \end{array}$$

Notice that if we were to apply this composition rule to any two of the three columns, we would produce the third column; so we cannot use the rule to produce a new column. We write this as $ABC = I, A^2 = B^2 = C^2 = I$, and use the rules for "ordinary multiplication" to write for example

$$A^2BC = IBC = BC = A,$$

and similarly

$$B = AC, C = AB.$$

This sort of rule works for two-factor design only. If we have a full factorial design with 3 binary factors

	Attribute		
Profile	A	B	C
1	$-$	$-$	$-$
2	$-$	$-$	$+$
3	$-$	$+$	$-$
4	$-$	$+$	$+$
5	$+$	$-$	$-$
6	$+$	$-$	$+$
7	$+$	$+$	$-$
8	$+$	$+$	$+$

We start by adding a fourth factor using the following composition rule $\oplus$ for any row:

# − terms		
Odd	→	$-$
Even	→	$+$

	Attribute			
Profile	A	B	C	D
1	$-$	$-$	$-$	$-$
2	$-$	$-$	$+$	$+$
3	$-$	$+$	$-$	$+$
4	$-$	$+$	$+$	$-$
5	$+$	$-$	$-$	$+$
6	$+$	$-$	$+$	$-$
7	$+$	$+$	$-$	$-$
8	$+$	$+$	$+$	$+$

and write $D = ABC$, From this we obtain

$$ABCD = I,$$

and thus the pattern of confounding:

$$A = BCD, B = ACD, C = ABD, D = ABC, AB = CD, AC = BD, AD = BC.$$

This is called a "Resolution IV" design, because there are four factors on the left side of $ABCD = I$. It means that the third order interactions are confounded with main effects and the second order interactions effects are confounded with each other. So if we start with a full factorial design using three factors, A, B, C and add a fourth factor D using the composition rule $\oplus$, we get a fractional factorial design in which factor A is confounded with the BCD interaction effect (in general every main effect is confounded with a third order interaction effect); and the second order interaction effects are confounded with each other. You can figure out which ones are confounded with which from the above expressions. Now suppose we were willing to assume that the second order interaction effects are not significant. Then we could use the following mapping to introduce new factors into the design without increasing the number of treatments as follows.

$$E = AB = CD, F = AC = BD, G = AD = BC.$$

So we will have a design with $2^3 = 8$ treatments and 7 factors, $A - G$. This type of design is said to be saturated because we have 8 observations and 8 parameter estimates (seven "treatment" effects and a mean).

Unfortunately, the simple set of rules for generating orthogonal designs does not extend to more than two factors. Generating experimental designs with more than binary factors requires an understanding of Galois theory and is beyond our present interest. Tables of designs are available (e.g., Addleman 1962, Plackett and Burman 1946), and are used to generate designs in programs like JMP (a SAS product).

Four other features of these rules for generating experimental designs are worth noting. First, the + and - coding for attribute levels are arbitrary; one typically plays around with them to make sure that (1) we don't have any dominated product profiles and (2) there are no infeasible product profiles (e.g., a very cheap top of the line computer). The intention is to construct product profiles that are both realistic and require tradeoffs among attributes. Second, the part worths correspond to treatment effects and are orthogonal (but only if you use effects coding of variables; dummy variable coding introduces some correlation). A necessary and sufficient condition for orthogonality is that if, in a design with N treatments (profiles), level j of one factor (attribute) appears in N_j treatments and level s of another factor appears in N_s treatments, then the two levels appear together in

$$N_{js} = \frac{N_j \cdot N_s}{N} \text{ treatments.}$$

Exercise 1. Is the following design orthogonal? Explain.

	Attribute			
Profile	A	B	C	D
1	$+$	$-$	$-$	$-$
2	$-$	$-$	$+$	$+$
3	$-$	$+$	$-$	$-$
4	$-$	$+$	$-$	$-$
5	$-$	$-$	$+$	$+$
6	$+$	$-$	$+$	$+$
7	$+$	$+$	$-$	$+$
8	$+$	$+$	$+$	$-$

Third, we typically start by confounding (aliasing) the highest order interaction effects, and progressively introduce new attributes by confounding lower-order interaction effects. Fourth, it is sometimes useful to begin with a large factorial arrangement and then split the treatments into orthogonal subsets of profiles according to a fractional factorial. This can be done by identifying some factors in the confounding scheme (e.g., factor G above) with a "block," and then assigning one person with the subset of profiles associated with the + levels of G and another person with the profiles associated with the - level of G. Adding more blocking variables introduces more blocks and fewer profiles per person. However, note that the effects (part worths) are orthogonally estimable within person only if there are more profiles evaluated by a person than the number of parameters in the model. Such a method also allows one to partial out within-person response bias in ratings.

Exercise 2. Use blocking to construct two orthogonal blocks of four profiles each.

	Attribute		
Profile	A	B	C
1	−	−	−
2	−	−	+
3	−	+	−
4	−	+	+
5	+	−	−
6	+	−	+
7	+	+	−
8	+	+	+

Appendix 2

Monotone regression. Consider the regression

$$r = a + bx + e,$$

where e is an error term, r is a rating for a product on a $1-5$ (Likert) scale, x is its price, and a, b are the unknown model parameters. We run a regression and find that the a and b values are not significant. Then it strikes us that r may not be an interval scale — we have many more data points than the 5 categories in which we can "bin" the responses.[7] What should we do?

One answer, which Joseph Kruskal gave in the 1960's, is to assume that there is an unobserved, interval scale y, which is related to r by some unknown isotonic function $y = f(r)$ and to x by the relation

$$y = \alpha + \beta x + \epsilon.$$

Suppose we try out some isotonic transformations $y = f(r)$; e.g., $y = \sqrt{r}, y = r^2, y = e^r, y = \log r$. For each such transformation, we get a value y, which we assume is interval scaled, and we can run a linear regression of y on x. We run all such (hypothetical) regressions and then select that one for which the R^2 value is the largest. That is, we find a monotonically non-decreasing transformation of the ordered categories y for which we get the best fit between the unobserved interval-scale response and price. We can never do worse in terms of goodness-of-fit than if we run a regression of r on x, because one possible transformation is $y = f(r) = r$. We can potentially do a lot better if, for example, the "true" relation between y and r is nonlinear, say $y = \sqrt{r}$. The problem is that there are an infinite number of permissible transformations, $f(r)$; we cannot possibly try out all of them. But we can device an algorithm that successively changes the values of y — maintaining the monotonicity requirement with respect to r — to get better and better fit. We stop when no further improvement is possible; i.e., when proportion of error variance on the unobserved scale

$$E^2 = \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

is as small as possible (the summation runs over all possible values of i). As usual, $\hat{y}$ means the predicted values of the (unobserved) y values.[8] Kruskal's (1964) formulation is slightly different: he minimizes the "stress," S, where

$$S^2 = \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}$$

in which the denominator is the squared deviations of the predicted values from their mean value. This reformulation allows for an efficient algorithm to solve the following problem:

[7] Alternatively, we might suspect that there is a linear relation between $f(r)$ and x, but we do not know $f(\cdot)$: we'd like to do some exploratory analysis to figure out the possibilities.

[8] You can see that $1 - E^2$ is analogous to R^2 as a measure of fit.

$$\min_{y_i} S = \sqrt{\frac{\sum_i (y_i - \hat{y})^2}{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}}$$

$$\text{s.t.}$$

$$y_i \leq y_j \quad \text{if} \quad r_i \leq r_j, \quad 1 \leq i < j \leq n.$$

This is a nonlinear optimization problem. We can simplify it a little by noting that we can always set $\alpha = 0$ without affecting the monotonicity constraint or the fit. Similarly, we can normalize the y values so that

$$\bar{\hat{y}} = 0 \quad \text{and} \quad \frac{1}{n-1} \sum_i (\hat{y}_i - \bar{\hat{y}})^2 = 1.$$

For example, suppose $n = 3$ and $\hat{y}_1 = 6, \hat{y}_2 = 4, \hat{y}_3 = 2$. If we subtract the average value 4 from each observation, we get the (revised) values (which we still call $\hat{y}$): $\hat{y}_1 = \hat{2}, \hat{y}_2 = 0, \hat{y}_3 = -2$. These values still satisfy the monotonicity condition, and also the condition $\bar{\hat{y}} = 0$. Then we divide each value of $\hat{y}$ by

$$\sqrt{\frac{1}{3-1}\Big\{(2-0)^2 + (0-0)^2 + (-2-0)^2\Big\}} = 2$$

i.e., set $\hat{y}_1 = 1, \hat{y}_2 = 0, \hat{y}_3 = -1$ so that

$$\frac{1}{3-1} \sum_{i=1}^{3} (\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{2}\{(1-0)^2 + (0-0)^2 + (-1-0)^2\} = 1.$$

Using standardized values of y, the optimization problem becomes:

$$\min_{\hat{\beta}, y_i} S = \sqrt{\sum_i (y_i - \hat{y}_i)^2} = \sqrt{\sum_i (y_i - \hat{\beta} x_i)^2}$$

$$\text{s.t.}$$

$$y_i \leq y_j \quad \text{if} \quad r_i \leq r_j, \quad 1 \leq i < j \leq n.$$

We can design an iterative algorithm to find the value of β that minimize S. Begin with some values y. For example, set $y = r$, treating them as if these are interval-scale values. Run an OLS regression of y on x (with zero intercept) to get $\hat{\beta}$ and $\hat{y} = \hat{\beta} x$. Arrange $\hat{y}_i = \hat{\beta} x_i$ in an increasing sequence of the r_i values. Check if the $\hat{y}$ values satisfy the monotonicity condition: $y_i \leq y_j$ if $r_i \leq r_j$. If the answer is "yes," set $y_i = \hat{y}_i$ for all $1 \leq i \leq n$ and take it as the final solution value. Otherwise, set the y values to be as close to $\hat{y}$ as possible while satisfying the monotonicity condition. If i, j are two indices for which $r_i < r_j$ but $\hat{y}_i > \hat{y}_j$, set all values $y_i, y_{i+1}, \ldots, y_j$, to the average value of $\hat{y}_i, \hat{y}_{i+1}, \ldots, \hat{y}_j$, where $r_i < r_{i+1} < r_{i+2} < \cdots < r_j$. For example, consider the following data (for expository ease, we will ignore the standardization of the $\hat{y}_i$ values):

i	1	2	3	4	5	6	7
r_i	1	2	3	4	5	6	7
$\hat{y}_i$	1	3	4	5	2	7	6

As $\hat{y}_3 > \hat{y}_5$ $(4 > 2)$, we set $y_3 = y_4 = y_5 = (4+5+2)/3 = 3.67$

i	1	2	3	4	5	6	7
r_i	1	2	3	4	5	6	7
$\hat{y}_i$	1	3	4	5	2	7	6
y_i	1	3	3.67	3.67	3.67	7	6

Also, $\hat{y}_6 > \hat{y}_7$ $(7 > 6)$, and we set $y_6 = y_7 = (7 + 6/2) = 6.5$

i	1	2	3	4	5	6	7
r_i	1	2	3	4	5	6	7
$\hat{y}_i$	1	3	4	5	2	7	6
y_i	1	3	3.67	3.67	3.67	6.5	6.5

At this point, we have our first revised estimates of $\hat{y}$ and the corresponding y values that satisfy the monotonicity condition. We can compute the value of S. We repeat the above process with these new y values, and stop when the value of S cannot be made smaller (or does not decrease by more than a small, specific amount, say δ); the associated β value is our solution.

The method can now be generalized to accommodate multiple predictor variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots,$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots.$$

There are some further improvements to make the iterative algorithm go faster, and checks to make sure that the solution does not diverge. Also, we do not have to start with $y = r$; any arbitrary or random starting value can be used. We generally run the procedure with many different starting values to make sure that the solution we get is not a local optimum.

Monotone regression can be used for preference scaling. For example, suppose $x_1, x_2, x_3, \ldots, x_m$ denote the locations of m brands on a *perceptual map.* If we know a person's preference ranking of the brands, we can use ordinal regression to estimate the preference model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

where y is an unobserved preference (utility) score. The above formulation is called a vector model (of preferences). We can use this model if all the x's are continuous, which is the situation with perceptual maps. But it is also possible to have the x's denote discrete, dummy variables, which is the typical situation in conjoint analysis, or to have combinations of continuous and discrete variables. There is one other case of particular interest. There are some experiential attributes as "sweetness" and "coldness" (say of a beverage) for which a person's preferences might diminish as the value of an alternative deviates from an (unknown) ideal value. Let x_i^* denote the ideal point for variable x_i. Suppose this is the only attribute under consideration. Then a convenient form for representing ideal-point preferences is

$$y = \beta_0 + \beta_1 (x - x^*)^2,$$

or

$$y = \beta_0 + \beta_1 x^2 + 2\beta_1 x^* x + \beta_1 x^{*2}.$$

Let

$$\alpha_0 = \beta_0 + \beta_1 {x^*}^2, \quad \alpha_1 = 2\beta_1 x^*, \quad \alpha_2 = \beta_1.$$

We use ordinal regression to estimate $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$, in the preference equation

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2,$$

and infer

$$\hat{\beta}_1 = \hat{\alpha}_2, \ \hat{x}^* = \frac{\hat{\alpha}_1}{2\hat{\beta}_1}.$$

For a single attribute, the estimation problem becomes

$$\min_{\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, y_i} S = \sqrt{\sum_i (y_i - \hat{y}_i)^2} = \sqrt{\sum_i \left(y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i + \hat{\alpha}_2 x_i^2)\right)^2}$$

$$\text{s.t.}$$

$$y_i \leq y_j \ \text{ if } \ r_i \leq r_j, \quad 1 \leq i < j \leq n.$$

We have assumed so far that the x_i values are known. Suppose these were unknown, and we had preference rankings of the same set of stimuli for w people. Let p denote an individual. Then we can formulate the following problem, still in a single dimension:

$$\min_{\hat{\alpha}_0^p, \hat{\alpha}_1^p, \hat{\alpha}_2^p, x_i, y_i^p} S = \sqrt{\sum_p \sum_i (y_i^p - \hat{y}_i^p)^2}$$

$$\text{s.t.}$$

$$y_i^p \leq y_j^p \ \text{ if } \ r_i^p \leq r_j^p, \quad 1 \leq i < j \leq n, \quad 1 \leq p \leq w.$$

This is called an *unfolding model* or a model of internal analysis, because the ideal points and stimuli locations are internal to the model. We can extend these preference models to consider multiple attributes: all that changes is the number of dimensions over which the distance is computed, and the number of parameters, which are multiplied by the number of dimensions in the model.

Appendix 3

A Solution Procedure for Nonmetric MDS. Let $x_0, x_1, x_2, \ldots, x_s$, denote $s \geq 2$ points on the line and let $n = \binom{s}{2}$ denote the number of pairs of points; we denote the distances $y_1, y_2, y_3, \ldots, y_n$.

————————x_0——————————x_1——x_2————————

For example, the are $s = 3$ points x_0, x_1, x_2 on the above line and $n = \binom{3}{2} = 3$ pairs of points, $(x_0, x_1), (x_0, x_2), (x_1, x_2)$; we call the distances between the pairs of points

$$y_1 = |x_0 - x_1|, \quad y_2 = |x_0 - x_2|, \quad y_3 = |x_1 - x_2|.$$

Suppose we do not know the distance, but only an ordering $r_1 < r_2 < r_3 \cdots < r_n$ of the distances. Then we can perform a monotone regression in which y correspond to distances between pairs of points on a line. As before, we are interested in finding the best-fitting transformation $y = f(r)$ for which $y_i < y_j$ if $r_i < r_j$. But there are two differences. First, the relation between x and y is of the form

$$y_i = |x_k - x_l| = +\sqrt{(x_k - x_l)^2}.$$

Second, we do not know the x values, we want to find them. This seem odd, because we are used to thinking about regression problems in which x is a known predictor variable. It is still a predictor of y, except that it is unknown. What we have to do is start with some sequence of values x_i and then minimize the stress over both $y = f(r)$ and x. We normalize $\hat{y}$ values as before so that

$$\bar{\hat{y}} = 0, \quad \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 = 1.$$

Thus, we have

$$\min_{x_i, y_i} S = \sqrt{\sum_i (y_i - \hat{y})^2}$$

s.t.

$$y_i \leq y_j \;\text{ if }\; r_i \leq r_j, \quad 1 \leq i < j \leq n.$$

We substitute for y in terms of x:

$$\min_{x_k, \hat{x}_k} S = \sqrt{\sum_{k=1}^{s-1} \sum_{l=k+1}^{s} (x_k - x_l)^2 - (\hat{x}_k - \hat{x}_l)^2}$$

s.t.

$$y_i \leq y_j \;\text{ if }\; r_i \leq r_j, \quad 1 \leq i < j \leq n$$

where for simplicity of notation we write the constraint in terms of y, although in fact we should substitute $y_i = |x_k - x_l|$ and $y_j = |x_g - x_h|$. We will keep this in mind.

The only novel thing in this monotone regression problem is that the x's are unknowns, not the β values. Once again we use an iterative algorithm to find the best values of the points on the line. To start things off, we select some arbitrary point locations on the line. These are our initial estimates $\hat{x}_k, 1 \leq k \leq s$. We use these to calculate the estimates of pairwise distances $\hat{y}$; i.e., $\hat{y}_i = |\hat{x}_k - \hat{x}_l|$. We set $y = \hat{y}$ and check if these values satisfy the monotonicity condition: $y_i \leq y_j$ if $r_i \leq r_j$. If the answer is "yes," we can stop, taking $\hat{x}_k$ as our solution values. Otherwise, we change the values of y so that these satisfy the monotonicity condition: if i, j are two indices for which $r_i < r_j$ but $y_i > y_j$, (remember, these are the revised y values) we set $y_i = y_j$ and all intervening values of y in the ordering $r_i < r_{i+1} < r_{i+2} < \cdots < r_j$. We then compute S. Next, we change the x values to improve S. To do this, we use a "method of steepest descent" or "method of gradients," which changes the x_k values so that S decreases most quickly. This direction is called the (negative) gradient and is determined by evaluating the partial derivatives of the function S. The (negative) gradient is

$$\big(g_1, g_2, g_3, \ldots, g_s\big) = \bigg(-\frac{\partial S}{\partial x_1}, -\frac{\partial S}{\partial x_2}, -\frac{\partial S}{\partial x_3}, \ldots, -\frac{\partial S}{\partial x_s} \bigg).$$

Then the new values of x_k are

$$x_k^{'} = x_k + \frac{g_k}{G}\alpha$$

where α is the step size and

$$G = \sqrt{\sum_k g_k^2} \Bigg/ \sqrt{\sum_k x_k^2}.$$

The step size can be constant, but for computational reasons it often varies form one iteration to another, with an initial value of about 0.2. After arriving at a new, slightly better point, we again determine the gradient, which is different at different points, and move along it. We continue doing this until no improvement in the S value is possible; this is our solution, and the partial derivatives are all zero at the point. There are details to fill out, for example we have to decide the value of α at each iteration, we have to decide how small the partial derivatives can be for us to treat them as effectively zero, how to avoid local optima. We will ignore these; a SAS manual will tell you about how to deal with them.

This approach can be generalized to accommodate multiple perceptual variables. The difference is that the distance is

$$y_i = |x_k - x_l| = +\sqrt{\sum_j (x_{kj} - x_{lj})^2}\ ,$$

where the summation j runs over all dimensions in a problem. Everything now has to be done in terms of the $s \times t$ coordinates on the t dimensions, including the calculation of distances and the computation of the gradients. Nothing else changes. This method is called *nonmetric* multidimensional scaling (MDS) because we use ordinal data, not metric (interval or ratio) data, to locate the points in space.