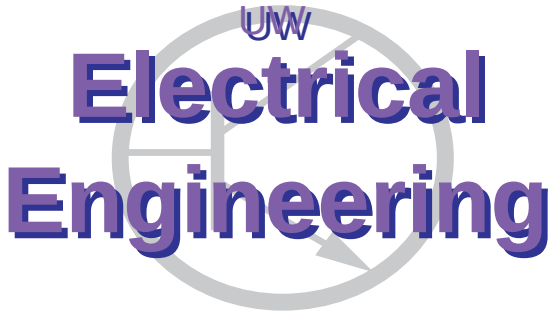
How to Analyze Paired Comparison Data

Kristi Tsukida and Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
{gupta}@ee.washington.edu



UWEE Technical Report
Number UWEETR-2011-0004
May 2011

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Abstract

Thurstone's Law of Comparative Judgment provides a method to convert subjective paired comparisons into one-dimensional quality scores. Applications include judging quality of different image reconstructions, or different products, or different web search results, etc. This tutorial covers the popular Thurstone Case V model and the Bradley-Terry logistic variant. We describe three approaches to model-fitting: standard least-squares, maximum likelihood, and Bayesian approaches. This tutorial assumes basic knowledge of random variables and probability distributions.

Contents

1 Why Paired Comparisons?

When comparing different options, one often wishes to assign a single quality score to each option. For example, you may want to score the quality of different image processing algorithms, or the skill of different chess players, or the seriousness of different crimes. To illustrate, we place three images of an apple on a quality scale in Figure 1. For a quality scale, the relative difference between any two quality scores measures

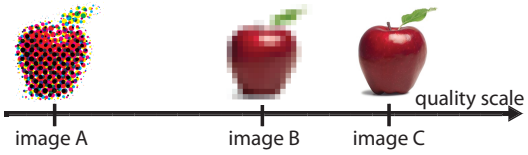


Figure 1: Image quality scale

how much better one image looks over another image, but defining the placement of zero, and the unit of measurement is arbitrary. Scales of measurement with these two degrees of freedom (zero placement and measurement unit) are known as *interval scales*[1] because the interval between any two scale values has meaning, but the numerical value of any single score is arbitrary. Equivalent interval scales can be defined with different zeros and units. For example, the Fahrenheit and Celsius temperature scales are equivalent interval scales with different zero placements, and different definitions of the amount of heat represented by "1 degree." You can convert between any two equivalent interval scales by shifting and multiplicatively scaling the scale values.

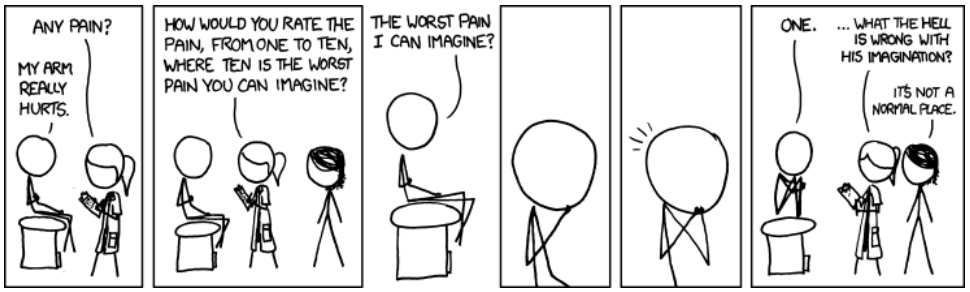


Figure 2: Pain Rating (xkcd cartoon by Randall Munroe [2])

How do you gather data to score a set of options? It is tempting to simply ask a bunch of people to score each of your options: "On a scale of 1 to 10, what is the quality of this image?" (Or "How would you rate the pain, from one to ten, where ten is the worst pain you can imagine?" as depicted in Figure 2.) However, people may mean different things by the same score (a 3 to one person may be different than a 3 for another person). It may be hard to determine specifically what "1" and "10" mean (how bad does an image have to be to get a 1?). It may be inflexible (what if you want to give something a 15?). Further, you may care more about scoring the options in the context of the set, rather than on an absolute scale (you want to know "How much better does this image look than the other options?" rather than "Does this image look good?"). Because of these issues, gathering paired comparisons may be more useful than directly asking for quality scores.

In a paired comparison experiment, you ask, "Is A better than B?" Generally ties are not allowed (or they may be counted as half a vote for each option). Ideally you would get comparisons for all possible pairs of options you are judging, but this is not necessary to estimate the scores, and for a large number of

[1] An alternate scale of measurement is a *ratio scale*, where the zero value is fixed because it has special meaning e.g. age, where 0 years is when you are born or Kelvins where $0°K$ is absolute zero. In a ratio scale we can always compare to 0, so 20 years is twice as old as 10 years, whereas on a interval scale $20°F$ is not twice as hot as $10°F$. The classification of measurement scales is discussed by Stevens [1].

options, may simply be infeasible. There may also be issues of order presentation (which option is presented first could affect the preference) but in the rest of this tutorial we assume that this issue can be ignored.

1.1 Roadmap

Now that we have described the problem setup, in the next section we define our experimental data (2) and Thurstone's statistical model of judgments, which provides a method of estimating the quality score difference for two options using Thurstone's Law of Comparative Judgment. We will then extend the analysis to estimating scores for more than two options using a least squares method (Section 4), a maximum likelihood method (Section 5), and using the expected value of the score (Section 6). We also discuss Bayesian methods (7). We illustrate the different approaches with simulations (Section 8). This document ends with an Appendix of proofs and Matlab code to implement the described functions.

2 Paired Comparison Data

The result of a paired comparison experiment is a count matrix, C, of the number of times that each option was preferred over every other option,

$$C_{i,j} = \begin{cases} \#\text{ of times option } i \text{ preferred over option } j, & i \neq j \\ 0, & i = j \end{cases} \tag{2}$$

We will assume that each paired comparison is independent, and that we don't need to know the order that any of the comparisons occurred. Generally, different pairs may have different total number of comparisons.

3 Models for Comparative Judgment

There are two common models for analyzing paired comparison data (2). We first discuss Thurstone's model, and then the Bradley-Terry model.

3.1 Thurstone's Model

In 1927, Louis Leon Thurstone pioneered psychometrics by using Gaussian distributions to analyze paired comparisons [3, 4]. Thurstone's model assumes that an option's quality is a Gaussian random variable.[2] This models the fact that different people may have different opinions on the quality of an option. Each option's *quality score* is taken to be the mean quality of the corresponding Gaussian.

Consider the basic case of two options, where we let the Gaussian random variables A and B represent the quality of option A and option B respectively,

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2), \qquad B \sim \mathcal{N}(\mu_B, \sigma_B^2).$$

Their probability density functions (PDFs) are

$$p_A(a) = \tfrac{1}{\sigma_A}\,\phi\Big(\tfrac{a-\mu_A}{\sigma_A}\Big), \qquad p_B(b) = \tfrac{1}{\sigma_B}\,\phi\Big(\tfrac{b-\mu_B}{\sigma_B}\Big),$$

where ϕ is the standard normal PDF (zero mean, unit variance),

$$\phi(x) = \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

As shown in Figure 3, option A is placed on the quality scale at μ_A, and option B is placed on the quality scale at μ_B. Thurstone's model says that when a person judges whether option A is better than option B,

[2] In some literature, the distribution of quality values is known as the *discriminal process* and the variance is the *discriminal dispersion*.

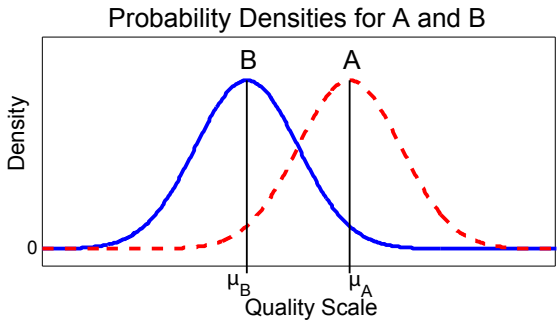


Figure 3: Probability density functions for A and B, the random quality for two options. The x-axis represents the quality scale, where each option is placed on the quality scale at its mean.

they draw a realization from A's quality distribution and a realization from B's quality distribution, and then choose the option with the higher quality. Equivalently, they choose option A over option B if their draw from the random quality difference $A - B$ is greater than zero,

$$P(A > B) = P(A - B > 0).$$

Since $A - B$ is the difference of two Gaussians, $A - B$ is a Gaussian random variable,

$$\begin{aligned} A - B &\sim \mathcal{N}(\mu_{AB}, \sigma_{AB}) \\ \mu_{AB} &= \mu_A - \mu_B \\ \sigma_{AB}^2 &= \sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B. \end{aligned} \tag{3}$$

where μ_{AB} is the *mean quality difference* of $A - B$, σ_{AB} is the standard deviation of the random quality difference $A - B$, and ρ_{AB} is the correlation between A and B.

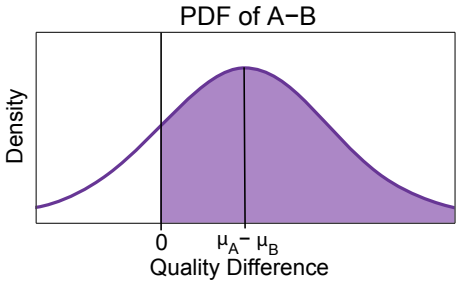


Figure 4: The random quality difference $A - B$ is Gaussian with mean $\mu_A - \mu_B$. $P(A > B)$ is the shaded area under the PDF curve of $A - B$.

Therefore the probability of choosing option A over option B (shown in Figure 4) is

$$\begin{aligned} P(A > B) &= P(A - B > 0) \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} e^{-(x-\mu_{AB})^2/(2\sigma_{AB}^2)} dx \\ &= \int_{-\mu_{AB}}^\infty \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} e^{-x^2/(2\sigma_{AB}^2)} dx \end{aligned}$$

By the symmetry of the Gaussian,

$$
\begin{aligned}
&= \int_{-\infty}^{\mu_{AB}} \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} e^{-x^2/(2\sigma_{AB}^2)} dx \\
&= \int_{-\infty}^{\mu_{AB}} \frac{1}{\sigma_{AB}} \phi(\tfrac{x}{\sigma_{AB}}) dx \\
&= \int_{-\infty}^{\frac{\mu_{AB}}{\sigma_{AB}}} \phi(t) dt \\
&= \Phi\left(\frac{\mu_{AB}}{\sigma_{AB}}\right), \qquad (4)
\end{aligned}
$$

where $\Phi(z)$ is the standard normal cumulative distribution function (CDF)

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt = \int_{-\infty}^{z} \phi(t)\, dt.$$

By inverting (4), we can calculate the mean quality difference μ_{AB} as

$$\mu_{AB} = \sigma_{AB}\, \Phi^{-1}\left(P(A > B)\right),$$

where $\Phi^{-1}(x)$ is the inverse CDF of the standard normal (also known as the *probit*). The inverse CDF of the standard normal is also commonly known as the *z-score* or *standard score* since it gives the number of standard deviations that x is from the mean. Although traditionally, getting the z-score required large lookup tables, modern computers can calculate the inverse CDF function precisely.

Thurstone proposed estimating $P(A > B)$ by the empirical proportion of people preferring A over B, $C_{A,B}/(C_{A,B} + C_{B,A})$. Assuming we know (or can estimate) the standard deviation σ_{AB}, the estimator for the quality difference $\hat{\mu}_{AB}$ is

$$\hat{\mu}_{AB} = \sigma_{AB}\, \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right). \qquad (5)$$

The estimate (5) is known as Thurstone's Law of Comparative Judgment.

3.2 Thurstone's Case V Model

The general model represented by (3) requires the correlation ρ_{AB} and the standard deviation σ_{AB} (or σ_A and σ_B) to be estimated. In his original paper [3] Thurstone made a number of model simplifications for tractability.[3] The simplest and most popular simplification is the Case V model, which assumes that each option has equal variance and zero correlation (or less restrictively, equal correlations instead of zero correlations [5]):

$$\sigma_A^2 = \sigma_B^2$$
$$\rho_{AB} = 0.$$

Without loss of generality, set the variances to one half $\sigma_A^2 = \sigma_B^2 = \frac{1}{2}$ so the variance of $A - B$ is one,

$$\sigma_{AB}^2 = \sigma_A^2 + \sigma_B^2 = 1.$$

[3] Case I assumes that the correlation ρ_{AB} is constant throughout all comparisons. Case II adds the assumption that the general model can be applied to judgments from a group of observers (as opposed to multiple judgments from the same observer). Case III additionally assumes that A and B are uncorrelated so that $\rho_{AB} = 0$. Case IV additionally assumes that the variances approximately equal, $\sigma_A = \sigma_B + \epsilon$, where ϵ is small. Case V additionally assumes that the variances are exactly equal, $\sigma_A = \sigma_B$.

This sets the scale unit for the interval scale (removing one degree of freedom) so that a quality scale difference of 1 implies that the mean of $A - B$ is one standard deviation of $A - B$. This simplifies Thurstone's Law given in (5) for Case V to

$$\hat{\mu}_{AB} = \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right). \tag{7}$$

For the rest of this tutorial, we will use these Case V assumptions and refer to this Case V model as *Thurstone's model*.

Another common approach is to assume that $\sigma_A^2 = \sigma_B^2 = 1$, so that the interval scale is measured in the number of standard deviations of the random quality (instead of the number of standard deviations of the random quality difference). Then, $\sigma_{AB} = \sqrt{2}$, and Thurstone's Law for Case V is $\hat{\mu}_{AB} = \sqrt{2}\Phi^{-1}(\frac{C_{A,B}}{C_{A,B}+C_{B,A}})$. This makes the equations less convenient, but then the quality scale differences are easily interpreted as the number of standard deviations of the random qualities. If you want the quality scale values in terms of the quality standard deviations, you can multiply the quality scale values in this tutorial by $\sqrt{2}$ (which is ok since the interval scale may be redefined by multiplicatively scaling the scale values).

3.3 Prior Knowledge

Prior knowledge is easily incorporated into the model by adding values to the count matrix according to what you believe the proportion of counts should be *a priori*. Create a matrix B of the proportion of times you believe one option would be preferred over the other and add a weighted version to the collected data,

$$\tilde{C} = C + \alpha B.$$

Then use the new $\tilde{C}$ matrix as if it was the data you collected.

Even if you don't know what the proportions should be, you can add a constant value to all the counts to smooth the counts (e.g. add one to all the counts to achieve Laplace smoothing). Using a prior B can help regularize the estimates and solve the 0-1 problem, as we discuss in Section 4.2.

3.4 The Bradley-Terry model

Ralph Bradley and Milton Terry [6] introduced an alternate model for paired comparisons, also known as the Bradley-Terry-Luce model (BTL) for Ducan Luce's extension to multiple variables in [7].

The original Bradley and Terry papers [6, 8, 9] develop the Bradley-Terry model as giving each option a rating, π_i which satisfies

$$P(\text{choose A over B}) = \frac{\pi_A}{\pi_A + \pi_B}. \tag{9}$$

Luce formulated the "Choice Axiom", extending the Bradley Terry model to accommodate comparisons of more than 2 objects, e.g. for 3 options, each rating π_i must satisfy,

$$P(\text{choose A out of A, B, and C}) = \frac{\pi_A}{\pi_A + \pi_B + \pi_C}.$$

By changing variables $\pi_i = \exp(\mu_i/s)$ (where s is a scale parameter), (9) can be rewritten as

$$\begin{aligned} P(\text{choose A over B}) = P(A > B) = P(A - B > 0) &= \frac{\exp(\mu_A/s)}{\exp(\mu_A/s) + \exp(\mu_B/s)} \\ &= \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\mu_A - \mu_B}{2s}\right). \end{aligned} \tag{10}$$

Equation (10) is $1 - F_{A-B}(0)$, where F is the CDF of the random variable $A - B$. Therefore, it is consistent with (10) to assume that $A - B$ is a logistic random variable with mean $\mu_A - \mu_B$ and scale parameter s.

Bradley [10] Luce [7] and others noted the similarity to Thurstone's model in that the Bradley-Terry model assumes the random quality difference $A - B$ has a logistic distribution where the Thurston assumes

that the random quality difference $A - B$ is Gaussian. Noting the similarity, in 1959 [7] Luce asked what distribution of A and B yield a logistic $A - B$ (thus satisfying the "Choice Axiom"). Block and Marschak's 1960 proof[11] and Holman and Marley's simplified proof (printed in [12]) show that if A and B have Gumbel distributions of qualities then $A - B$ is logistic. We offer a simple alternate proof in Appendix B.

Unlike the Gaussian CDF which requires evaluating the erf function, the logistic CDF has a closed-form expression. We can estimate the quality difference $\mu_{AB} = \mu_A - \mu_B$ by inverting (10) and estimating $P(A > B)$ as the empirical count proportion $\frac{C_{A,B}}{C_{A,B}+C_{B,A}}$. The inverse logistic CDF (also known as the *logit*) has a closed-form expression (since $\tanh^{-1}(x) = \frac{1}{2}[\ln(1+x) - \ln(1-x)]$), so the BTL quality difference estimate is

$$\hat{\mu}_{AB} = s\left(\ln\left(\frac{C_{A,B}}{C_{A,B}+C_{B,A}}\right) - \ln\left(1 - \frac{C_{A,B}}{C_{A,B}+C_{B,A}}\right)\right). \tag{12}$$

To compare the BTL model scale differences with the Thurstone model ones from (5), equate the variance by setting $s = \frac{\sqrt{3}}{\pi}$. Empirically as shown in Figure 5, the logistic CDF is very similar to the Gaussian CDF, so that using Thurstone's model or the BTL model produces very similar results. Some people prefer BTL for computational simplicity (you don't have to compute the erf function for the inverse Gaussian CDF), although with modern computers and algorithms, computing the inverse Gaussian CDF is simple, so the computational aspect is not an issue.

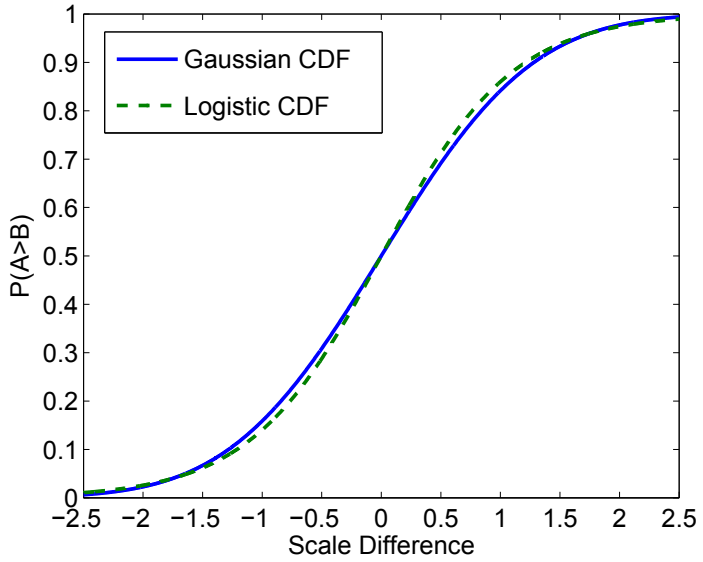


Figure 5: Gaussian vs Logistic CDF

The logistic CDF has a fatter tail and is slightly more sloped at the inflection point than a Gaussian with the same mean and variance. This means that the BTL model will estimate slightly smaller scale differences for proportions near $\frac{1}{2}$ and slightly larger scale differences for proportions near 0 or 1 when compared with Thurstone's model. (For example, if $P(A > B) = 0.7$, Thurstone's model would estimate the quality scale difference to be slightly bigger than a half, but the BTL model would estimate the quality scale difference to be slightly less than a half.)

4 Model Fitting

Thurstone's model provides a method of estimating the scale difference for any single pair of options by estimating $P(A > B)$ by the empirical proportion of people preferring A to B. However, when considering more

than two options, we generally can't find exact scale values which satisfy all of the scale difference estimates. (The scale differences form an over-determined system, for example consider we can't find μ_A, μ_B, μ_C such that $\mu_A - \mu_B = 1$, $\mu_B - \mu_C = 2$ and $\mu_A - \mu_C = 5$.) In this section we detail three different approaches to estimating the quality scores given more than two options using the Thurstone model (the same approaches can be applied to the BTL model).

4.1 Thurstone-Mosteller Least Squares Method

To determine the quality scores for a set of m options, Thurstone offered a solution, which Mosteller later showed was the solution to a least squares optimization problem [5]. Define the vector of quality scores $\mu = [\mu_1, \mu_2, ..., \mu_m]$, and let D be an $m \times m$ matrix where $D_{i,j} = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j}+C_{j,i}}\right)$ is the (Case V) Thurstone's Law estimate (7) for the quality difference between option i and option j. (You may also use the BTL model by forming D using the logit from (12).) The least squares estimate for the quality scores μ minimizes the squared error between the quality scores and the Thurstone's Law pairwise estimates:

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^m} \sum_{i,j} (D_{i,j} - (\mu_i - \mu_j))^2.$$

This least squares problem has a simple closed-form solution which can be derived from the D matrix. If we set $\hat{\mu}_1 = 0$, the least squares solution is

$$\hat{\mu}_j = \sum_{i=1}^{m} \frac{D_{i,1}}{m} - \sum_{i=1}^{m} \frac{D_{i,j}}{m}.$$

Instead of assuming $\hat{\mu}_1 = 0$, another common approach is to assume that the mean of all the $\hat{\mu}_i$ is zero. In this case, the least squares solution is

$$\hat{\mu}_j = -\sum_{i=1}^{m} \frac{D_{i,j}}{m}.$$

4.2 Least Squares Disadvantages

When estimating quality differences by this least squares method, we will have problems when $C_{i,j}$ is zero or $n_{i,j}$: the proportion $\frac{C_{i,j}}{C_{i,j}+C_{j,i}}$ will be 0 or 1, so that $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$, causing the estimates for μ_i and μ_j to be ∞ or $-\infty$.

There are a couple of ways to deal with this problem. One solution[13] to the "0/1 problem" is to simply ignore the 0/1 entries and use an incomplete matrix solution [14, 15]. We argue this is too heavy-handed a fix in that it ignores important information that the one option is strongly preferred to the other option.

A second solution is to "fix" the 0/1 proportions[13] by adding and subtracting a count or fractional count to the unanimous pairs:

$$\tilde{C}_{ij} = \begin{cases} \frac{1}{2} & \text{if } C_{i,j} = 0 \text{ and } i \neq j \\ n_{i,j} - \frac{1}{2} & \text{if } C_{i,j} = n_{i,j} \text{ and } i \neq j \\ C_{i,j} & \text{otherwise} \end{cases} \tag{15}$$

where $n_{i,j} \triangleq C_{i,j} + C_{j,i}$ is the total number of comparisons of the (i,j) pair. We will refer to this modified data matrix as the *0/1 fixed data*. This can be viewed as correcting for the discrete nature of the count data. A related solution is to add a count or fractional count to both the 0 and $n_{i,j}$ entries (or add counts to all entries) of the count matrix; this is equivalent to assuming some prior data (see Section 3.3). These fixes do change the count matrix, but in a conservative way that biases the counts toward less confidence, and this fix is not as big a change as simply ignoring 0/1 entries.

A third solution is to estimate the means by the maximum likelihood estimate, which we detail in Section 5.2. This is the solution we recommend because it does not require ignoring or altering the data (potentially adding noise).

Another disadvantage to the least squares estimation is that it only looks at the proportion $\frac{C_{i,j}}{C_{i,j}+C_{j,i}}$, ignoring the total number of judgments, which contains information on the accuracy of the data.

4.3 Incomplete Matrix Solution

Morrissey [14] and Gulliksen [15] independently formulated an incomplete matrix solution to estimate the quality scores from a subset of the paired comparison data (ignoring missing data or pairs with 0 or 1 proportions). Use the same matrix D of Thurstone's Law estimates, $D_{i,j} = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j}+C_{j,i}}\right)$. The incomplete matrix solution is formulated as the least squares solution to a system of equations using only the valid data entries, where the last equation constrains the μ_i to have zero mean:

$$\begin{bmatrix} D_{1,2} \\ D_{1,3} \\ D_{2,3} \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}.$$

Let $d = [D_{1,2}, D_{1,3}, D_{2,3}, ..., 0]$ be a vector of all the valid entries in D with a zero in the last entry, and X be the length$(d)\times m$ matrix so that

$$d = X\mu.$$

As long as $X^T X$ is invertible, the incomplete least squares solution can be found by using the Moore-Penrose pseudoinverse of X

$$\hat{\mu} = (X^T X)^{-1} X^T d.$$

Morrissey's and Gulliksen's solutions are equivalent, but Gulliksen directly constructs the $X^T X$ matrix and $X^T d$ vector, so it may be faster or more stable when implemented.

5 Maximum Likelihood Scale Values

First, we show that if there are just two options, the maximum likelihood estimate of the quality score difference is given by Thurstone's law (5). Then we give the maximum likelihood estimate for multiple options (which is not equivalent to the least-squares solutions in the previous section).

5.1 Maximum Likelihood for Two Options

Let the paired judgment data for two options, A and B, be $a = C_{A,B}$ and $b = C_{B,A}$. The probability of this data, given the probability of choosing A over B, has a binomial distribution,

$$P(a, b|P(A > B)) = \binom{a+b}{a} P(A > B)^a (1 - P(A > B))^b.$$

Since we can calculate $P(A > B)$ given the quality scale difference μ_{AB}, define the likelihood of the μ_{AB} as

$$\begin{aligned} L(\mu_{AB}) = P(a, b|\mu_{AB}) &= \binom{a+b}{a} P(A > B)^a P(B > A)^b \\ &= \binom{a+b}{a} \Phi(\mu_{AB})^a \Phi(-\mu_{AB})^b, \end{aligned} \tag{17}$$

where we have used the identity $P(B > A) = 1 - \Phi(\mu_{AB}) = \Phi(-\mu_{AB})$. The maximum likelihood quality difference is

$$\begin{aligned}\hat{\mu}_{AB} &= \arg\max_{\mu_{AB}} L(\mu_{AB}) \\ &= \arg\max_{\mu_{AB}} \Phi(\mu_{AB})^a \Phi(-\mu_{AB})^b.\end{aligned}$$

We may equivalently maximize the log-likelihood

$$\mathcal{L}(\mu_{AB}) = \log P(a, b|\mu_{AB}) = \log \binom{a+b}{a} + a \log\left(\Phi(\mu_{AB})\right) + b \log\left(\Phi(-\mu_{AB})\right), \tag{19}$$

yielding the optimization

$$\begin{aligned}\hat{\mu}_{AB} &= \arg\max_{\mu_{AB}} \mathcal{L}(\mu_{AB}) \\ &= \arg\max_{\mu_{AB}} a \log\left(\Phi(\mu_{AB})\right) + b \log\left(\Phi(-\mu_{AB})\right).\end{aligned}$$

This may be solved by setting the derivative of the objective to zero,

$$\begin{aligned}0 &= \frac{a}{\Phi(\mu_{AB})}\phi(\mu_{AB}) - \frac{b}{\Phi(-\mu_{AB})}\phi(-\mu_{AB}) \\ \frac{a}{\Phi(\mu_{AB})} &= \frac{b}{1 - \Phi(\mu_{AB})} \\ \hat{\mu}_{AB} &= \Phi^{-1}\left(\frac{a}{a+b}\right) = \Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right),\end{aligned}$$

which verifies that Thurstone's Law yields the maximum likelihood solution if there are only $m = 2$ options.

5.2 Maximum Likelihood for Multiple Options

Extending the two option maximum likelihood estimation described in Section 5, to a comparison of m options, there is no longer a closed-form solution. Instead, one must solve a convex optimization problem.

Let μ be the vector of quality scale values $\mu = [\mu_1, \mu_2, ..., \mu_m]$. Define the log-likelihood of μ given our count data, C, as in (19),

$$\mathcal{L}(\mu|C) \triangleq \log P(C|\mu) = \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)).$$

To find the maximum likelihood solution quality scale values, one must solve

$$\begin{aligned}\arg\max_{\mu} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) \\ \text{subject to} \quad & \sum_i \mu_i = 0.\end{aligned} \tag{21}$$

To find a unique solution, include the constraint that the mean of all the quality scale values is zero as in (21), or set one of the quality scale values to zero $\mu_1 = 0$.

We show in the appendix that (21) is a convex optimization problem.

5.3 Maximum A Posteriori Estimation

One can also form the maximum a posteriori (MAP) estimate, by including a prior on the scale values $p(\mu)$

$$\begin{aligned} \arg\max_{\mu} \quad & \mathcal{L}(\mu|C) + \log(p(\mu)) \\ \text{subject to} \quad & \sum_i \mu_i = 0. \end{aligned}$$

If there is little information about the true scale values that can be used to choose a prior, then we suggest a Gaussian prior that assumes the different scale values are drawn independently and identically from a standard normal will reduce the estimation variance and often provide better estimates. In that case the MAP estimate solves:

$$\begin{aligned} \arg\max_{\mu} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) - \sum_i \frac{\mu_i^2}{2} \\ \text{subject to} \quad & \sum_i \mu_i = 0. \end{aligned} \tag{23}$$

This choice of prior performs a Ridge regularization on the scale values [16] and remains a convex optimization problem.

5.4 Advantages of Maximum Likelihood Estimation

Maximum likelihood estimation is an optimal approach to estimation problems in the sense that it produces the solution which makes the data most likely. Additionally, this maximum likelihood solution does not suffer from the 0/1 problem of the least squares methods because the maximum likelihood method does not use the inverse CDF. The 0 entries in the count matrix do not contribute to the likelihood and μ_i are constrained by the other terms in the log-likelihood to keep them from being driven to ∞.

Another advantage to the maximum likelihood estimation is that it takes into account the variance in estimating the data based on the total number of judgments made for each pair. If there is a pair which has a large number of judgments, the maximum likelihood solution will trust that data more. In comparison, the least squares estimation only cares about the proportion of judgments $\frac{C_{i,j}}{C_{i,j}+C_{j,i}}$, ignoring the total number of judgments.

6 Expected Quality Scale Difference

Instead of using the maximum likelihood quality difference, one can estimate the quality difference to be the *expected quality difference* where the expectation is taken with respect to the likelihood (or with respect to the posterior mean if there is a prior). On average, we expect this solution to perform better since this approach uses the full likelihood information rather than just the maximum of the likelihood.

We only consider the two option case here (multiple options are possible, but requires m-dimensional numerical integration).

6.1 Expected Quality Estimate

Treat the unknown quality score difference as a random variable U. The likelihood of U is the probability of observing a people preferring option A, and b people preferring option B, as given in (17),

$$P(a, b|U = u) = \binom{a+b}{a} P(A > B|U = u)^a \, P(B > A|U = u)^b.$$

Using Bayes rule, the posterior can be written in terms of the likelihood as[4]

$$\begin{aligned} p(U=u|a,b) &= \frac{P(a,b|U=u)p(U=u)}{P(a,b)} \\ &= \frac{1}{\gamma} P(A>B|U=u)^a \, P(B>A|U=u)^b \, p(U=u) \\ &= \frac{1}{\gamma} \Phi(u)^a \, (1-\Phi(u))^b \, p(U=u), \end{aligned}$$

where γ is a normalization constant

$$\gamma = \int_{\infty}^{\infty} \Phi(u)^a \, (1-\Phi(u))^b \, p(U=u) du.$$

If we assume a uniform prior for U over the range $[-t, t]$, then the expected quality scale difference is

$$E[U|a,b] = \frac{1}{2t\gamma} \int_{-t}^{t} u \, \Phi(u)^a (1-\Phi(u))^b \, du. \tag{24}$$

Alternatively, we could assume a Gaussian prior for U. Using the standard normal, the expected quality scale difference is

$$E[U|a,b] = \frac{1}{\gamma} \int_{-\infty}^{\infty} u \, \Phi(u)^a (1-\Phi(u))^b \, \phi(u) du$$

Performing a change of variables $x = \Phi(u)$,

$$= \frac{1}{\gamma} \int_{0}^{1} \Phi^{-1}(x) \, x^a (1-x)^b \, dx.$$

6.2 Computation of Expected Quality Estimate

Unfortunately, no closed form solution exists, so the integral and the normalizer constant must be numerically computed. Numerical integration is slow, and may be prone to precision errors depending on the method of integration. Matlab may be used to attempt to approximate the integral (trapz, quad, quadgk), but Matlab is limited to machine precision (32 or 64 bits) and may not evaluate the integral accurately enough. (Matlab may return 0 when the actual solution should be a very small non-zero number). Maple and Mathematica have more sophisticated numerical integration routines and arbitrary precision calculations, so they may yield better results.

7 Bayesian Estimation

In the previous section, we considered the quality difference, U, to be a random variable, and estimated it by taking its expectation. In this section, we instead consider the probability that option i is chosen over option j to be a random variable, $X_{i,j}$. We consider the result of performing Bayesian estimation, and the relation to priors and smoothing.

The count data is generated from a binomial distribution where the parameter $x_{i,j}$, is an observation of $X_{i,j}$:

$$\begin{aligned} C_{i,j}, C_{j,i} &\sim \text{Binom}(C_{i,j}, C_{j,i} \,|\, x_{i,j}) \\ p(C_{i,j}, C_{j,i}|x_{i,j}) &\propto {x_{i,j}}^{C_{i,j}} (1-x_{i,j})^{C_{j,i}}. \end{aligned}$$

[4] Although at first glance $P(u|a,b)$ looks similar to a beta distribution, it is parameterized by u, instead of $p = \Phi(u)$, so it is not the same. The normalizer must be calculated numerically.

After observing the data C and assuming a uniform prior probability on $x_{i,j}$, the posterior probability of $x_{i,j}$ has a beta distribution with parameters $C + 1$:

$$X_{i,j} \sim \text{Beta}(x_{i,j}|C_{i,j}+1, C_{j,i}+1). \tag{25}$$
$$p(x_{i,j}|C_{i,j}, C_{j,i}) \propto {x_{i,j}}^{C_{i,j}}(1-x_{i,j})^{C_{j,i}}.$$

The maximum a posteriori estimate of $x_{i,j}$ is $\hat{x}_{i,j} = \dfrac{C_{i,j}}{C_{i,j}+C_{j,i}}$ (the mode of the beta distribution (25)). Then calculate the quality scale difference by setting $\hat{x}_{i,j} = \Phi(u_i - u_j)$ and inverting, which results in Thurstone's law:

$$u_i - u_j = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j}+C_{j,i}}\right).$$

Next, instead of estimating $x_{i,j}$ by its posterior mode, consider estimating $x_{i,j}$ by its posterior mean $E[X_{i,j}] = \dfrac{C_{i,j}+1}{C_{i,j}+C_{j,i}+2}$. Then the resulting estimate of the quality scale difference is

$$u_i - u_j = \Phi^{-1}\left(\frac{C_{i,j}+1}{C_{i,j}+C_{j,i}+2}\right).$$

This result is equivalent to the Thurstone's law estimate if one puts a prior of 1 on all the counts, meaning that *a priori* you believe that all of the choices are possible. This may also be interpreted as Laplace smoothing the count data.

8 Illustrative Experiments

We illustrate the different estimation approaches with some simulations. We make n quality observations for each of pair of m options (simulating surveying n people for their preferences about all possible pairs of m options). We first generate the true means $\mu_1, \mu_2, ..., \mu_m$ for the m quality score distributions. Then, for each pairwise comparison for each person, the perceived quality scores for the ith option are drawn IID from $\mathcal{N}(\mu_i, \sigma^2 = \frac{1}{2})$ as in Thurstone's Case V. The count data is collected, and $\hat{\mu}$ is estimated from the data.

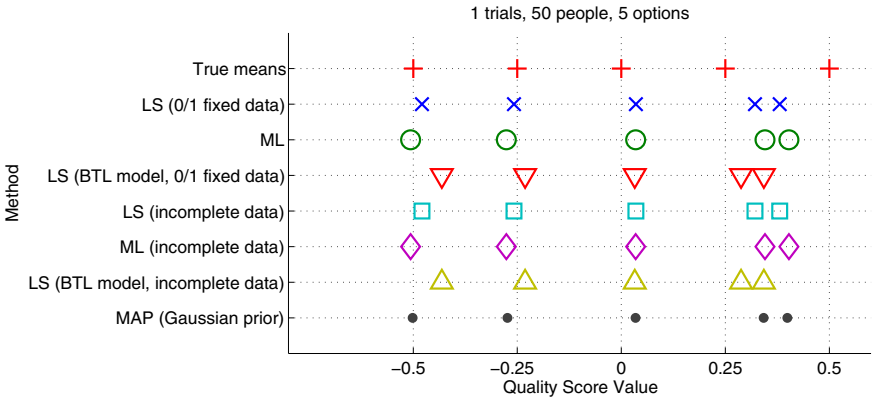


Figure 6: Results of one trial comparing five options with true qualities $(-.5, -.25, 0, .25, .5)$. Plot compares estimates from least-squares estimate assuming the Thurstone model ("LS"), the maximum likelihood estimate assuming the Thurstone model ("ML", "MAP"), and the least-squares estimate assuming the BTL model ("LS (BTL model)").

Fig. 6 illustrates the results of one run of the simulation, with $n = 50$ people surveyed about all pairs of $m = 5$ options. The true mean values are marked on top, and are at $(-.5, -.25, 0, .25, .5)$. The mean

quality estimates are shown on the next rows for the least-squares estimate assuming the Thurstone model, the maximum likelihood estimate assuming the Thurstone model, and the least-squares estimate assuming the BTL model.

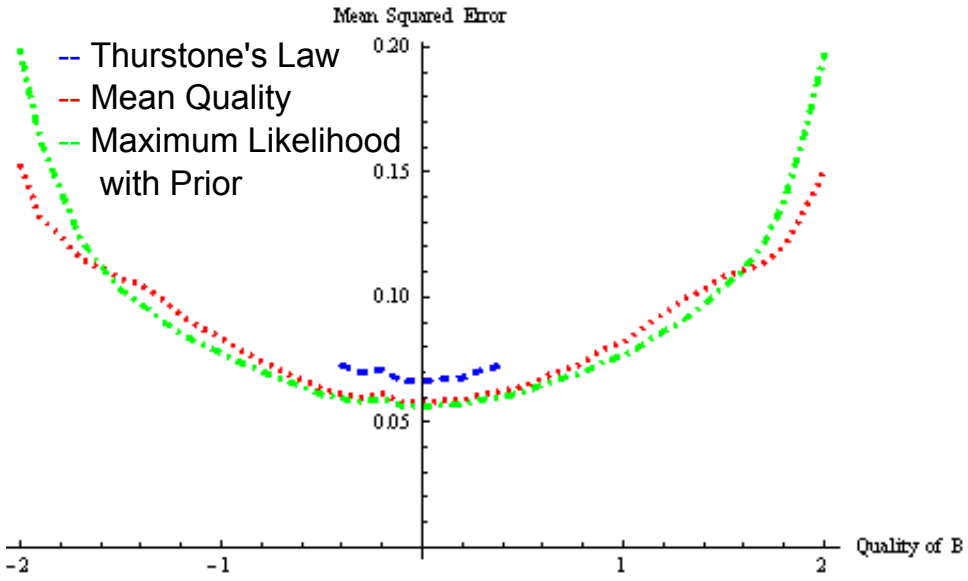


Figure 7: Result of 10,000 trials with two options. In each trial we simulate 25 people judging each paired comparison. The mean quality of option A is fixed at 0, and the mean quality of option B is varied along the x axis.

Fig. 7 shows results for a single pair of options ($m = 2$), averaged over 10,000 runs of the simulation for varying true quality-differences. If the true quality difference (shown on the x-axis) is large, then the 0/1 problem occurs (see Sec. 4.2), and the Thurstone's Law estimate given by (5) is that the quality difference is "infinite." For this case of $m = 2$ options, recall that (5) is also the maximum likelihood estimate. The green line is the maximum likelihood estimate given a prior of 1 count to both options (this could also be called a maximum a posterior estimate). This is always well-defined and performs better than Thurstone's Law for all quality differences. The red line shows the mean quality estimate as given by (24). This is a more robust estimate, and as shown in Fig. 7, will perform better than the maximum likelihood with prior when the true quality difference is large, but may perform worse when the true quality difference is small.

Fig. 8 shows the simulation results when there are $m = 10$ options. These results were averaged over 1000 runs of the simulation. For each run, the true mean quality of each of the ten options was chosen uniformly on $[-x, x]$. Fig. 8 compares the Bradley-Terry model (labeled "BTL") with the Thurstone model (all results not labeled "BTL"). We also compare the different methods of solving the 0/1 problem (Section 4.2): the least squares methods ("LS") where any 0/1 proportions were"fixed" according to (15) (denoted by "0/1 fix"), Morrissey and Gulliksen's incomplete matrix solution [14, 15] (denoted by "incomplete"), maximum likelihood ("ML"), and the maximum a posteriori estimate where the prior is independently and identically a standard normal on each of the quality scores ("MAP", from (23)).

We show two different metrics in Fig. 8:

Interval Mean Squared Error: the average squared error in the quality scale difference for each option

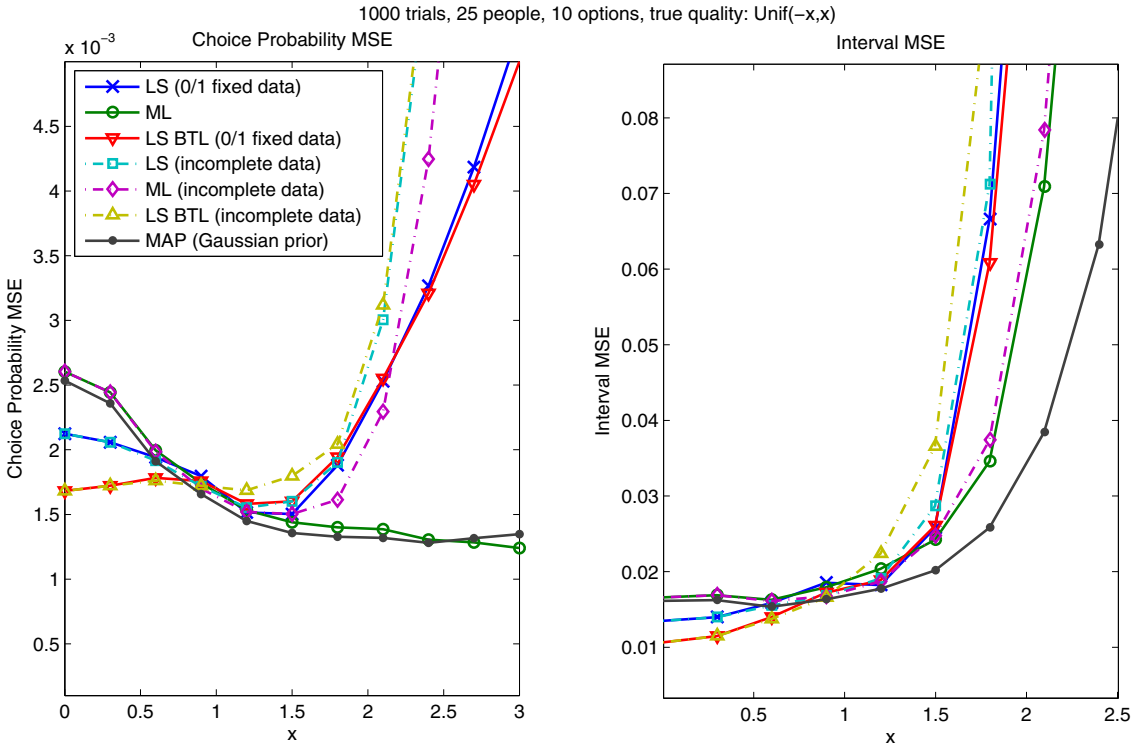


Figure 8: Result of 1,000 trials comparing ten options. In each trial 25 people judge each paired comparison.

pair.

$$S_{i,j} = \mu_i - \mu_j = \text{ ground truth quality difference}$$
$$S^*_{i,j} = \mu^*_i - \mu^*_j = \text{ estimated quality difference}$$
$$\text{Interval MSE} = \sum_{i \neq j} \frac{(S_{i,j} - S^*_{i,j})^2}{m(m-1)}$$

Probability Mean Squared Error: the average squared error in the estimated choice probability for each option pair.

$$P_{i,j} = P(Q_i > Q_j) = \Phi(\mu_i - \mu_j) = \text{ ground truth choice probability}$$
$$P^*_{i,j} = \Phi(\mu^*_i - \mu^*_j) = \text{ estimated choice probability}$$
$$\text{Choice Probability MSE} = \sum_{i \neq j} \frac{(P_{i,j} - P^*_{i,j})^2}{m(m-1)}$$

When the true qualities are close together, the BTL logistic model performs slightly better than the Thurstone's Gaussian model because the logistic CDF has a steeper slope when the probability is $\frac{1}{2}$ so it estimates slightly lower values. The least squares methods perform worse as the true means become more separated, but the maximum likelihood methods perform better.

9 Summary

Fitting Thurstone's model or the BTL model to paired comparison data can be a useful tool to analyze the relative qualities of a set of options. Various estimation methods can be used to fit each model. If the

true quality differences are believed to be separated by at least a standard deviation, then we suggest using the maximum a posteriori estimate. The maximum a posteriori is advantageous because we have seen it consistently produce good results, it is an optimal approach to estimation, the data does not need to be modified to avoid 0/1 problems, and it can be solved efficiently as a convex optimization problem.

Appendix

Appendix A: The Maximum Likelihood Is A Convex Optimization Problem

We prove that solving for the maximum likelihood scale values $\hat{\mu}$ in (21) is a convex optimization problem.

Definition 1 *Log-concave.* A function $f : \mathbb{R}^n \to \mathbb{R}$ is log-concave if $f(x) > 0$ for all $x \in \text{dom } f$ and $\log(f(x))$ is concave (or equivalently if $-\log(f(x))$ is convex).

Proposition 1 *$f(x)$ is log-concave iff $f(x) > 0$ and $(f'(x))^2 \geq f''(x)f(x)$*

PROOF This follows from the condition that $h(x) = -\log(f(x))$ is convex iff the second derivative of $h(x)$ is greater than or equal to zero. ■

Lemma 1 *The cumulative distribution function of a log-concave differentiable probability density is log-concave.*

PROOF Lemma 1 is proved by Prekopa [17] using measure theory, and by Bagnoli and Bergstrom [18] using the Cauchy mean value theorem. This is a simple alternate proof. (This proof approach appears in [19] as exercise 3.55.)

Let $g(t) = \exp(-h(t))$ be a differentiable log-concave probability density function and let the cumulative distribution be

$$f(x) = \int_{-\infty}^{x} g(t)\,dt = \int_{-\infty}^{x} e^{-h(t)}\,dt.$$

We prove that f is log-concave because it satisfies Proposition 1.
The derivatives of f are

$$\begin{aligned} f'(x) &= g(x) = e^{-h(x)} \\ f''(x) &= g'(x) = -h'(x)e^{-h(x)} = -h'(x)g(x). \end{aligned}$$

For $h'(x) \geq 0$, since $f(x) > 0$ and $g(x) > 0$,

$$\begin{aligned} f(x)f''(x) &= -f(x)h'(x)g(x) \leq 0 \\ (f'(x))^2 &= g(x)^2 > 0. \end{aligned}$$

Therefore, by Proposition 1, if $h'(x) \geq 0$, then $f(x)$ is log-concave. To show Lemma 1 also holds for $h'(x) < 0$, we note that since h is convex,

$$h(t) \geq h(x) + h'(x)(t - x).$$

Taking the negative, exponent and integrating both sides,

$$\begin{aligned} \int_{-\infty}^{x} e^{-h(t)}\,dt &\leq \int_{-\infty}^{x} e^{-h(x)-h'(x)(t-x)}\,dt \\ &= e^{-h(x)+xh'(x)} \int_{-\infty}^{x} e^{-th'(x)}\,dt \\ &= e^{-h(x)+xh'(x)} \frac{e^{-xh'(x)}}{-h'(x)} \\ &= \frac{e^{-h(x)}}{-h'(x)}. \end{aligned}$$

Multiplying both sides by $-h'(x)e^{-h(x)}$,

$$-h'(x)e^{-h(x)} \int_{-\infty}^{x} e^{-h(t)}\, dt \leq e^{-2h(x)}$$
$$f''(x)f(x) \leq (f'(x))^2.$$

■

Theorem 1 *Solving for the maximum likelihood scale values $\hat{\mu}$ in (21)*

$$\begin{aligned} \arg\max_{\mu} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) \\ \text{subject to} \quad & \sum_{i} \mu_i = 0 \end{aligned} \tag{21}$$

is a convex optimization problem.

Proof The maximum likelihood solution (21) is equivalent to solving for the $\hat{\mu}_{ij}$ in

$$\begin{aligned} \arg\max_{\mu_{ij}} \quad & \sum_{i,j} C_{i,j} \log(\Phi(\mu_{ij})) \\ \text{subject to} \quad & \mu_{ij} + \mu_{jk} = \mu_{ik} \quad \forall\, i,j,k \in \{1,...,m\}. \end{aligned}$$

The standard Gaussian PDF $\phi(x)$ is log-concave since the second derivative of its log is $\frac{d^2}{dx^2} \log \phi(x) = -1$. So, by Lemma 1 the standard Gaussian CDF $\Phi(x)$ is also log-concave.

The likelihood $\sum_{i,j} C_{i,j} \log(\Phi(\mu_{ij}))$ is concave since $\log(\Phi(x))$ is concave and concavity is preserved under addition and positive scaling. ■

Appendix B: The Difference Of Two Gumbel Random Variables Is A Logistic Random Variable

We show that the difference of two Gumbel random variables is a logistic random variable. We begin with some preliminaries.

A Gumbel random variable has the cumulative distribution function (CDF)

$$F(x;\mu,\beta) = e^{-e^{-(x-\mu)/\beta}}$$

and probability density function (PDF)

$$f(x;\mu,\beta) = \frac{1}{\beta} e^{-(x-\mu)/\beta} e^{-e^{-(x-\mu)/\beta}}.$$

The standard Gumbel PDF is

$$f(z;0,1) = e^{-z} e^{-e^{-z}}.$$

Any Gumbel PDF can be expressed in terms of the standard Gumbel PDF

$$f(x;\mu,\beta) = \frac{1}{\beta} f\left(\frac{x-\mu}{\beta}; 0, 1\right).$$

Also, since f is a PDF,

$$\beta = \int_{-\infty}^{\infty} e^{-x/\beta} e^{-e^{-x/\beta}} dx,$$

and

$$1 = \int_{-\infty}^{\infty} e^{-x} e^{-e^{-x}} dx.$$

A logistic random variable has the cumulative distribution function (CDF)

$$\begin{aligned} F(x;\mu,s) &= \frac{1}{1-e^{-(x-\mu)/s}} \\ &= \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{x-\mu}{2s}\right) \end{aligned}$$

and probability density function (PDF)

$$\begin{aligned} f(x;\mu,s) &= \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2} \\ &= \frac{1}{4s}\operatorname{sech}^2\left(\frac{x-\mu}{2s}\right). \end{aligned}$$

Any logistic distribution may be expressed in terms of the standard logistic distribution,

$$f(x;\mu,s) = \frac{1}{s} f\left(\frac{x-\mu}{s};0,1\right).$$

To show the equality of the two above definitions of the logistic PDF, let $z = (x-\mu)/s$ (for simplicity), then

$$\begin{aligned} f(z) &= \frac{1}{s}\frac{e^{-z}}{(1+e^{-z})^2} \\ &= \frac{1}{s}\frac{e^{-z}}{(1+2e^{-z}+e^{-2z})} \\ &= \frac{1}{s}\frac{1}{(e^{z}+2+e^{-z})} \\ &= \frac{1}{4s}\left(\frac{2}{e^{z/2}+e^{-z/2}}\right)^2 \\ &= \frac{1}{4s\cosh^2(z/2)} \\ &= \frac{1}{4s}\operatorname{sech}^2(z/2). \end{aligned}$$

Theorem 2 *If X and Y are independent standard Gumbel random variables with respective PDFs*

$$\begin{aligned} f(x;0,1) &= e^{-x}e^{-e^{-x}}, \\ g(y;0,1) &= e^{-y}e^{-e^{-y}}, \end{aligned}$$

then $X-Y$ is a logistic random variable with mean $\mu = 0$ and scale parameter $s = 1$.

Proof The PDF of $X-Y$ can be computed as the convolution $f(t;0,1) * g(-t;0,1)$:

$$\begin{aligned} h(t) &= \int_{-\infty}^{\infty} f(\tau;0,1)g(\tau-t;0,1)d\tau \qquad (38) \\ &= \int_{-\infty}^{\infty} f(t+\tau;0,1)g(\tau;0,1)d\tau \\ &= \int_{-\infty}^{\infty} e^{-(t+\tau)}e^{-e^{-(t+\tau)}}e^{-\tau}e^{-e^{-\tau}}d\tau \\ &= e^{-t}\int_{-\infty}^{\infty} e^{-2\tau}e^{-(e^{-(t+\tau)}+e^{-\tau})}d\tau \\ &= e^{-t}\int_{-\infty}^{\infty} e^{-2\tau}e^{-(1+e^{-t})e^{-\tau}}d\tau. \end{aligned}$$

Change variables to $z = e^{-\tau} \implies dz = -e^{-\tau} d\tau$,

$$\begin{aligned} &= -e^{-t} \int_{\infty}^{0} z e^{-(1+e^{-t})z} dz \\ &= e^{-t} \int_{0}^{\infty} z e^{-(1+e^{-t})z} dz. \end{aligned}$$

Since $\int_0^\infty x e^{-ax} dx = \frac{1}{a^2}$,

$$= \frac{e^{-t}}{(1+e^{-t})^2}$$

Therefore $X - Y$ is a logistic random variable with $\mu = 0$ and $s = 1$. ∎

Theorem 3 *More generally, if X and Y are Gumbel independent random variables with equal scale parameters and respective PDFs*

$$\begin{aligned} f(x; \mu_x, \beta) &= e^{-(x-\mu_x)/\beta} e^{-e^{-(x-\mu_x)/\beta}} \\ g(y; \mu_y, \beta) &= e^{-(y-\mu_y)/\beta} e^{-e^{-(y-\mu_y)/\beta}}, \end{aligned}$$

then $X - Y$ is a logistic random variable with mean $\mu = \mu_x - \mu_y$ and scale parameter $s = \beta$.

Proof The PDF of $X - Y$ can be computed as

$$\begin{aligned} p(t) &= \int_{-\infty}^{\infty} f(\tau; \mu_x, \beta) g(\tau - t; \mu_y, \beta) d\tau \\ &= \int_{-\infty}^{\infty} f(t + \tau; \mu_x, \beta) g(\tau; \mu_y, \beta) d\tau \\ &= \int_{-\infty}^{\infty} \frac{1}{\beta} f\left(\frac{t + \tau - \mu_x}{\beta}; 0, 1\right) \frac{1}{\beta} g\left(\frac{\tau - \mu_y}{\beta}; 0, 1\right) d\tau. \end{aligned}$$

Let $z = \frac{\tau - \mu_y}{\beta} \implies dz = \frac{1}{\beta} d\tau$

$$= \frac{1}{\beta} \int_{-\infty}^{\infty} f\left(z + \frac{t - (\mu_x - \mu_y)}{\beta}; 0, 1\right) g\left(z; 0, 1\right) dz.$$

Using $h(t)$ as defined in (4) from the proof of Theorem 2,

$$\begin{aligned} &= \frac{1}{\beta} h\left(\frac{t - (\mu_x - \mu_y)}{\beta}\right) \\ &= \frac{e^{-\frac{t-(\mu_x-\mu_y)}{\beta}}}{\beta \left(1 + e^{-\frac{t-(\mu_x-\mu_y)}{\beta}}\right)^2}. \end{aligned}$$

Therefore $X - Y$ is a logistic random variable with $\mu = \mu_x - \mu_y$ and $s = \beta$. ∎

10 Code

```
function S = scale_ls(counts)
% Use the least squares (complete matrix) solution
% (Thurstone 1927, Mosteller 1951) to
% scale a paired comparison experiment using
% Thurstone's case V model (assuming sigma^2 = 0.5 for each
% quality's distribution)
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values
%
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
N = counts + counts';
P = counts ./ (N + (N==0)); % Avoid divide by zero
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

Z = norminv(P);
S = -mean(Z,1)';
```

```
function S = scale_ls_btl(counts)
% Use the least squares complete matrix solution to
% scale a paired comparison experiment using
% Bradley-Terry's logistic model
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values
%
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

s = sqrt(3) / pi; % logistic distribution parameter
[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
N = counts + counts';
P = counts ./ (N + (N==0)); % Avoid divide by zero
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

Z = s * log(P ./(1-P)); % logit z-scores
S = -mean(Z,1)';
```

```
function S = scale_inc(counts, threshold)
% Use the Morrisey-Gulliksen incomplete matrix solution to
% scale a paired comparison experiment using
% Thurstone's case V model (assuming sigma^2 = 0.5 for each
% quality's distribution so that any quality difference has
% unit variance)
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values
%   Scale values are set up to have mean of 0
```

```
%
% (This code follows Gulliksen's formulation given in Engeldrum's
% book, Psychometric Scaling)
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

if nargin < 2 || isempty(threshold)
    % default threshold on scale difference
    threshold = 2;
end

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
N = counts + counts';
P = counts ./ (N + (N==0)); % Avoid divide by zero
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

% Thurstone's law estimates of each pairwise quality difference
% (norminv calculates the z-scores or z-value)
Z = norminv(P);
% Note the diagonal entries are included since diag(Z)=0
valid = (abs(Z) < threshold);
Z(¬valid) = 0;

d = sum(Z, 1)'; % Vector of column sums
M = double(¬valid); % 0 for valid entries, 1 where |Z(i,j)| > threshold
M(eye(m)>0) = sum(valid); % Set the diagonal values

S = M \ d; % = inv(M) * d;
```

```
function S = scale_inc_btl(counts, threshold)
% Use the Morrisey-Gulliksen incomplete matrix solution to
% scale a paired comparison experiment using the Bradley-Terry-Luce
% (BTL) model (assuming that any quality difference has
% unit variance)
%
% counts is a n-by-n matrix where
%   counts(i,j) = # people who prefer option i over option j
% S is a length n vector of scale values
%   Scale values are set up to have mean of 0
%
% (This code follows Gulliksen's formulation given in Engeldrum's
% Psychometric Scaling book)
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

if nargin < 2 || isempty(threshold)
    % default threshold on scale difference
    threshold = 2;
end

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

% Empirical probabilities
N = counts + counts';
P = counts ./ (N + (N==0)); % Avoid divide by zero
P(eye(m)>0) = 0.5; % Set diagonals to have probability 0.5

s = sqrt(3) / pi; % set variance to 1
Z = s * log(P ./(1-P)); % logit (inverse logistic CDF)

% Note the diagonal entries are included since diag(Z)=0
valid = (abs(Z) < threshold);
```

```
Z(¬valid) = 0;

d = sum(Z, 1)'; % Vector of column sums
M = double(¬valid); % 0 for valid entries, 1 where |Z(i,j)| > threshold
M(eye(m)>0) = sum(valid); % Set the diagonal values

S = M \ d; % = inv(M) * d;
```

```
function S = scale_ml(counts)
% Use cvx to compute maximum likelihood scale values
% assuming Thurstone's case V model.
%
% S* = argmax P(counts|S) P(S)
%        S
%    = argmin −log P(counts|S)
%        S
%
% Assume that mean(S)=0.
%
% CVX can be obtained at http://cvxr.com/cvx/
%
% 2011−06−05 Kristi Tsukida <kristi.tsukida@gmail.com>

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

counts(eye(m)>0) = 0; % set diagonal to zero

previous_quiet = cvx_quiet(1);
cvx_begin
    variables S(m,1) t;
    SS = repmat(S,1,m);
    Δ = SS − SS'; % Δ(i,j) = S(i) − S(j)

    minimize( t );
    subject to
    −sum(sum(counts.*log_normcdf(Δ))) ≤ t
    sum(S)==0
cvx_end
cvx_quiet(previous_quiet);
```

```
function S = scale_map(counts)
% Use cvx to compute maximum a posterori scale values,
% assuming Thurstone's case V model.
%
% S* = argmax P(counts|S) P(S)
%        S
%    = argmin −log P(counts|S) − log P(S)
%        S
%
% Assume a Gaussian prior for P(S), and that mean(S)=0.
%
% CVX can be obtained at http://cvxr.com/cvx/
%
% 2011−06−05 Kristi Tsukida <kristi.tsukida@gmail.com>

% std dev for the prior on S
prior_sigma=1;

[m,mm] = size(counts);
assert(m == mm, 'counts must be a square matrix');

```

```
counts(eye(m)>0) = 0; % set diagonal to zero

previous_quiet = cvx_quiet(1);
cvx_begin
    variables S(m,1) t;
    SS = repmat(S,1,m);
    Δ = SS - SS'; % Δ(i,j) = S(i) - S(j)

    minimize( t );
    subject to
    -sum(sum(counts.*log_normcdf(Δ))) + sum(square(S))/(2*prior_sigma) ≤ t
    sum(S)==0
cvx_end
cvx_quiet(previous_quiet);
```

```
function S = scale_ex(a,b)
% Use the expected quality scale value to
% scale a paired comparison experiment using
% Thurstone's case V model (assuming sigma^2 = 0.5 for each
% quality's distribution)
%
% a is the number of times option A is preferred over option B
% b is the number of times option B is preferred over option A
%
% S is the scale values (assuming that S_A = 0)
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

lower_bound = -10;
upper_bound = 10;
tol = 1e-30;

% compute normalizer
%N=@(x) (normcdf(x).^a).*((1-normcdf(x)).^b);
N=@(x) exp(a * log(normcdf(x)) + b * log(1-normcdf(x)));
normalizer = quadl(N, lower_bound, upper_bound,tol);

%F=@(x) x.*(normcdf(x).^a).*((1-normcdf(x)).^b);
F=@(x) x .* exp(a * log(normcdf(x)) + b * log(1-normcdf(x)));
S_B = quadl(F, lower_bound, upper_bound,tol) / normalizer;

S = [0 S_B];
```

```
function fixed_counts = fix_counts(counts)
% "fixes" a count matrix so that any pair of comparisons
% with zero or N unanimous judgments are set to be
% .5 and N-.5 respectively,
% where N is the total number of judgments for the pair.
%
% counts is a count matrix where counts(i,j) is the
% number of times option i was preferred over option j.
%
% fixed_counts(i,j) = 0.5,          if counts(i,j) = 0 and i≠j
%                     N - 0.5,      if counts(i,j) = N and i≠j
%                     counts(i,j), otherwise
%
% 2011-06-05 Kristi Tsukida <kristi.tsukida@gmail.com>

N = counts+counts'; % Total number of comparisons for each pair
zero_counts = (counts == 0); % logical matrix: 1 if counts(i,j)=0
N_counts = zero_counts'; % logical matrix: 1 if counts(i,j)=N(i,j)

fixed_counts = counts;
```

```
fixed_counts(zero_counts) = 0.5;
fixed_counts(N_counts) = N(N_counts)-0.5;
fixed_counts(N==0) = 0; % If there were no comparisons, don't modify counts
fixed_counts(eye(size(counts))>0) = 0; % Set the diagonal to be zero
```

```
% This is a simple script which estimates scale values
% for a paired comparison experiment.
%
% The maximum likelihood and MAP estimation methods require cvx
% to be installed in the Matlab environment.
% http://cvxr.com/cvx/
%
% Kristi Tsukida <kristi.tsukida@gmail.com>
% June 5, 2011

% clear variables;
% close all;

% % Generate a count data matrix
% mu = [-1 -0.5 0 0.5 1];
% mu = mu - sum(mu); % force mu to be zero mean
% sigma = 1/sqrt(2); % std dev for each quality score
% num_judgments_per_pair = 100; % e.g. # people making judgments
% num_opt = 5; % options
%
% count_data = zeros(num_opt);
% for num=1:num_judgments_per_pair
%     % Each person generates a new quality score for each pair
%     quality = normrnd(mu(:)*ones(1,num_opt), sigma);
%     % Add a count for the higher quality option for each pair
%     count_data = count_data + bsxfun(@gt, quality, quality');
% end

% Paired comparison count data matrix.
% This count_data matrix was generated with the above code.
% count_data(i,j) is the number of times option i was
% preferred over option j.
count_data = [   0    27    24     4     0
                73     0    29    10     8
                76    71     0    37    16
                96    90    63     0    32
               100    92    84    68     0 ];

%=====================
% Fix 0/1 proportions in the data
% (for use with the least squares estimators)
%=====================
fixed_count_data = fix_counts(count_data);

%=====================
% Estimate scale values
%=====================

% Incomplete matrix, maximum likelihood, and MAP methods
% don't require the 0/1 fixed data.

% Z-score threshold for least squares incomplete matrix methods
thresh = 2;

% Least squares method, Thurstone model (Gaussian)
% Least squares method should use fixed count data to
% "solve" the 0/1 problem.
% (Not using the fixed data results in estimates with +Inf or -Inf
%  scale values for any 0/1 proportion entries)
```

```
S_ls_fixed = scale_ls(fixed_count_data);

% Incomplete Matrix solution for the Thurstone model (Gaussian)
S_inc = scale_inc(count_data, thresh);

% Least squares method, BTL model (Logistic)
% Least squares method should use fixed count data to
% "solve" the 0/1 problem.
% (Not using the fixed data results in estimates with +Inf or -Inf
%  scale values for any 0/1 proportion entries)
S_ls_btl_fixed = scale_ls_btl(fixed_count_data);

% Incomplete Matrix solution for the BTL model (Logistic)
S_inc_btl = scale_inc_btl(count_data, thresh);

% Maximum likelihood method
% (Requires cvx)
S_ml = scale_ml(count_data);

% Maximum a posteriori (MAP) method
% (Requires cvx)
S_map = scale_map(count_data);

%=====================
% Plot results
%=====================
one = ones(size(S_ls_fixed));

scatter(S_ls_fixed, -1*one, 'bx');
hold on;
scatter(S_inc, -2*one, 'bo');

scatter(S_ls_btl_fixed, -3*one, 'gx');
scatter(S_inc_btl, -4*one, 'go');

scatter(S_ml, -5*one, 'rx');
scatter(S_map, -6*one, 'r^');
hold off;

title('Scale values for different methods')
xlabel('Scale values')
ylabel('Methods')
methods={'Least Squares (with 0/1 fixed data)', ...
    'Incomplete Matrix Solution',...
    'BTL model, Least Squares (with 0/1 fixed data)', ...
    'BTL model, Incomplete Matrix Solution',...
    'Maximum Likelihood', ...
    'Maximum A Posteriori Likelihood'};
set(gca,  'YTick',-6:-1, 'YTickLabel', fliplr(methods));
ylim([-7,0]);
grid on;
```


Acknowledgements

This work was supported by the United States Office of Naval Research and the United States PECASE award. We would like to thank Bela Frigyik, Nathan Parish, Alex Marin, Amol Kapila, and Evan Hanusa for helpful discussions.

References

[1] S. S. Stevens, "On the Theory of Scales of Measurement." *Science (New York, N.Y.)*, vol. 103, no. 2684, pp. 677–80, Jun. 1946.

[2] R. Munroe, "Pain Rating," http://xkcd.com. [Online]. Available: http://xkcd.com/883

[3] L. L. Thurstone, "A law of comparative judgment," *Psychological review*, 1927.

[4] R. D. Luce, "Thurstone and sensory scaling: Then and now." *Psychological Review*, vol. 101, no. 2, pp. 271–277, 1994.

[5] F. Mosteller, "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, vol. 16, no. 1, pp. 3–9, Mar. 1951.

[6] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, p. 324, Dec. 1952.

[7] R. D. Luce, *Individual choice behavior; a theoretical analysis.* New York: Wiley, 1959.

[8] R. A. Bradley, "Rank Analysis of Incomplete Block Designs: II. Additional Tables for the Method of Paired Comparisons," *Biometrika*, vol. 41, no. 3, pp. 502–537, Dec. 1954.

[9] ——, "Rank Analysis of Incomplete Block Designs: III Some Large-Sample Results on Estimation and Power for a Method of Paired Comparisons," *Biometrika*, vol. 42, no. 3, pp. 450–470s, Dec. 1955.

[10] ——, "Some Statistical Methods in Taste Testing and Quality Evaluation," *Biometrics*, vol. 9, no. 1, pp. 22–38, Dec. 1953.

[11] H. Block and J. Marschak, *Random orderings and stochastic theories of responses.* Contributions to Probability and Statistics, 1960, vol. 97.

[12] R. D. Luce and P. Suppes, "Preference, utility, and subjective probability," *Handbook of mathematical psychology*, vol. 3, pp. 249–410, 1965.

[13] P. G. Engeldrum, "Psychometric scaling : a toolkit for imaging systems development," Winchester, Mass., 2000.

[14] J. H. Morrissey, "New method for the assignment of psychometric scale values from incomplete paired comparisons." *Journal of the Optical Society of America*, vol. 45, no. 5, pp. 373–8, May 1955.

[15] H. Gulliksen, "A least squares solution for paired comparisons with incomplete data." *Psychometrika*, vol. 21, pp. 125–134, 1956.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* New York: Springer-Verlag, 2001.

[17] A. Prekopa, "On logarithmic concave measures and functions," *Acta Sci. Math.(Szeged)*, vol. 34, pp. 335–343, 1973.

[18] M. Bagnoli and T. Bergstrom, "Log-concave probability and its applications," *Economic Theory*, vol. 26, no. 2, pp. 445–469, Aug. 1989.

[19] S. P. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge: Cambridge Univ. Press, 2010. [Online]. Available: http://www.stanford.edu/~boyd/cvxbook/