Estimation and optimal designing under latent variable models for paired comparisons studies via a multiplicative algorithm

Bernard Torsney

Abstract We consider

1. The problem of estimating the parameters of latent variable models such as the Bradley Terry or Thurstone Model by the method of maximum likelihood, given data from a paired comparisons experiment. The parameters of these models can be taken to be weights which are positive and sum to one.
2. The problem of determining approximate locally optimal designs for good estimation of these parameters; i.e of determining optimal design weights which are also positive and sum to one.

1 Paired Comparisons

1.1 Introduction

We have two alternative examples of a general problem, namely determining weights optimally. Much theory for this problem, e.g. optimality conditions and numerical techniques have been developed in the optimal design arena. So this can be transported to the estimation problem. We can extend techniques to this case. In section 1 we introduce the notion of paired comparisons studies and latent variable models. In section 2 the parameter estimation problem is outlined with optimality results and a general class of multiplicative algorithms outlined in sections 3 and 4 respectively. A specific algorithm is applied to the Bradley Terry log-likelihood in section 5 and locally optimal designing is considered in section 6.

We consider paired comparison experiments in which J treatments or products are compared in pairs. In a simple form a subject is presented with two treatments

Bernard Torsney
Department of Statistics, University of Glasgow, Glasgow G12 8QW,UK e-mail: ben@stats.gla.ac.uk

and asked to indicate which he/she prefers or considers better. In reality the subject will be an expert tester; for example, a food taster in examples arising in food technology. The link with optimal design theory (apart from the fact that a specialised design, paired comparisons, is under consideration) is that, the parameters of latent variable models for the resultant data are like weights. Hence the theory characterising and the methods developed for finding optimal design weights can be applied to characterising and finding the maximum likelihood estimators of these latent variable 'weights'.

1.2 The Data

In a simple experiment a set of such testers is available and each is presented with one pair from a set of J treatments, say $T_1, T_2, \ldots, T_J$. The number of comparisons, n_{ij} of T_i to T_j, we assume has been predetermined. Sufficient summary data comprises the set $\{O_{ij} : i = 1,\ldots,J; j = 1,\ldots,J; i < j \text{ or } i > j\}$, where O_{ij} is the observed frequency with which T_i is preferred to T_j. Of course $O_{ij} + O_{ji} = n_{ij}$

Bradley and El-Helbawy (1976) introduce an example involving 8 coffee types. 26 pairwise comparisons were made on each pair, i.e. $n_{ij} = 26$.
So $O_{ij} + O_{ji} = 26$ and $N = \sum_i \sum_j O_{ij} = 728$.

The coffees are the eight combinations arising from a 2^3 factorial structure, the factors being Brew Strength, Roast Colour, Coffee Brand. We are not exploiting this structure and leave them arbitrarily labelled.

1.3 Models

1.3.1 A General Model

In the absence of other information the most general model here is to propose

$$O_{ij} \sim Bi\big(n_{ij}, \theta_{ij}\big)$$

where, $\theta_{ij} = P(T_i \text{ is preferred to } T_j)$.

Apart from the constraint $O_{ij} + O_{ji} = n_{ij}$, independence between frequencies is an expected assumption. So, apart from the constraint $\theta_{ij} + \theta_{ji} = 1$, these define unrelated binomial parameters. The maximum likelihood estimator of θ_{ij} is O_{ij}/n_{ij} (the proportion of times T_i is preferred to T_j in these n_{ij} comparisons), and formal inferences can be based on the asymptotic properties of these.

1.3.2 Latent Variable Models

These are more restricted models in that they impose interrelations between the θ_{ij}. Assuming that $F(\cdot)$ is a symmetric distribution function, then

$$\theta_{ij} = F(\lambda_i - \lambda_j) = F\{loge(p_i/p_j)\}$$

where $p_i = exp(\lambda_i)$. The symmetry of $F(\cdot)$ ensures that

$$\theta_{ij} + \theta_{ji} = F(\lambda_i - \lambda_j) + F(\lambda_j - \lambda_i) = 1.$$

The p_i or λ_i can be viewed as indices or quality characteristics, one for each treatment. The implication of the model is that the difference in quality between two treatments has distribution function $F(\cdot)$.

Two primary examples of this model are the Bradley Terry and Thurstone models. Respectively these take $F(\cdot)$ to be the Logistic and the Normal distributions. In the Logistic case θ_{ij} has the simplistic form: $\theta_{ij} = p_i/(p_i + p_j)$; see Thurstone (1927), Bradley and Terry (1952), also Kuk (1995).

2 Parameter Estimation

The likelihood of the data is

$$L = \prod_{r<}\prod_{s} \left[F\{loge(p_r/p_s)\}\right]^{O_{rs}} \left[F\{loge(p_s/p_r)\}\right]^{O_{sr}}.$$

We focus on the parameters p_i and denote the likelihood by $L(p)$.
However we cannot estimate these as free parameters. This arises from the fact that we only have observations on comparisons between treatments, and is reflected in the property that θ_{ij} is invariant to proportional changes in p_i and p_j. In consequence the p_i are only unique up to a constant multiple; (likewise the λ_i up to a constant shift). In keeping with this they are positive as the relationship $p_i = exp(\lambda_i)$ implies $p_i > 0$. Mathematically speaking θ_{ij} and hence $L(p)$ is a homogeneous function of degree zero in the p_i i.e. $L(cp) = L(p)$, where c is a scalar constant. So $L(p)$ is constant on rays running out from the origin. It will therefore be maximised along one specific ray. We can identify this ray by finding a particular optimising p^*. This we can do by imposing a constraint on p. Possible constraints are $\sum_i p_i = 1$ or $\prod_i p_i = 1$, or $g(p) = 1$ where $g(p)$ is a surface which cuts each ray exactly once. In the case $J = 2$ a suitable $g(p)$ is defined by $p_2 = h(p_1)$, where $h(\cdot)$ is a decreasing function which cuts the two main axes, as in the case of $h(p_1) = 1 - p_1$, or has these as asymptotes, as in the case of $h(p_1) = 1/p_1$. In general a suitable choice of $g(p)$ is one which is positive and homogeneous of some degree h. Note that other alternatives are $\sum_i p_i = C$ or $\prod_i p_i = C$, where C is any positive constant; e.g. $C = J$ or $C = 100$.

The choice of $\prod_i p_i = 1$, being equivalent to $\sum_i ln(p_i) = 0$, confers on $\lambda_i = ln(p_i)$ the notion of a main effect. However we will opt for the choice of $\sum_i p_i = 1$, which conveys the notion of p_i as a weight. We wish to maximise the likelihood or log-likelihood subject to this constraint and to non-negativity too. This is an example of the following general problem:

Problem $(\mathfrak{P})$

Maximise $\phi(p)$ subjec to $p_i \geq 0,\ \sum_i p_i = 1$.

We wish to maximise $\phi(p)$ with respect to a probability distribution.
For the estimation problem we will take $\phi(p) = ln\{L(p)\}$.
There are many examples of this problem arising in various areas of statistics, especially in the area of optimal regression design. We can exploit optimality results and algorithms developed in this area. The feasible region is an open but bounded set. Thus there should always be a solution to this problem allowing for the possibility of an unbounded maximum, multiple solutions and solutions at vertices (i.e. $p_t = 1, p_i = 0,\ i \neq t$).

3 Optimality Conditions

We assume that $\phi(\cdot)$ is differentiable. Let

$$F_j = d_j - p^T d = d_j - \sum_i p_i d_i, \text{ where } d_j = \partial\phi/\partial p_j.$$

We call F_j the jth vertex directional derivative of $\phi(\cdot)$ at p.
Note that $\sum_j p_j F_j = 0$, so that, in general, some F_j are negative and some are positive.

Given $\phi(\cdot)$ is differentiable at p^*, then a necessary condition for $\phi(p^*)$ to be a local maximum of $\phi(\cdot)$ in the feasible region of Problem $(\mathfrak{P})$ is

$$F_j^* = 0 \text{ for } p_j^* > 0,$$
$$F_j^* \geq 0 \text{ for } p_j^* = 0.$$

If $\phi(\cdot)$ is concave on its feasible region, then these first order stationarity conditions are both necessary and sufficient. This is the general equivalence theorem in optimal design. See Whittle (1973), Kiefer (1974). In fact the second condition is redundant for this estimation problem, while, given homogeneity of degree zero of L(p), the first reduces to standard first order conditions: $d_j^* = 0$.

4 Algorithms

4.1 Multiplicative Algorithm

Problem ($\mathfrak{P}$) has a distinct set of constraints, namely the variables $p_1, p_2, \ldots, p_J$ must be nonnegative and sum to 1. Let $f(d,\delta)$ be a function satisfying (for $\delta > 0$):

- $f(d,\delta) > 0$,
- $\frac{\partial f(d,\delta)}{\partial d} > 0$ (for $\delta > 0$),
- $f(d,0) = \text{constant}$

(e.g. $f(d,\delta) = \Phi(\delta d)$ or $f(d,\delta) = d^{\delta}$ (if $d > 0$.))

An iteration which neatly submits to these and has some suitable properties is the multiplicative algorithm:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(d_j^{(r)})}{\sum_i p_i^{(r)} f(d_i^{(r)})}$$

where $d_j^{(r)} = \frac{\partial \phi}{\partial p_j}\bigg|_p = p^{(r)}$, while $f(d)$ is positive and strictly increasing in d and may depend on one or more free parameters.

4.2 Properties of the Algorithm

Under the conditions imposed on $f(\cdot,\cdot)$, the above iterations possess the following properties which are considered in more detail in Torsney (1988), Torsney and Alahmadi (1992) and Mandal and Torsney (2000):

1. $p^{(r)}$ is always feasible.
2. $F_\phi\{p^{(r)}, p^{(r+1)}\} \geq 0$, with equality when the d_j's corresponding to nonzero p_j's have a common value $d (= \sum_i p_i d_i\ )$, in which case $p^{(r)} = p^{(r+1)}$.
   So an iterate $p^{(r)}$ is a fixed point of the iteration if derivatives $d_j^{(r)}$ corresponding to nonzero $p_j^{(r)}$ are equal; i.e. if corresponding vertex directional derivatives $F_j^{(r)}$ are zero.
3. If $\delta = 0$ there is no change in $p^{(r)}$, given $f(d,\delta) = constant$
4. So the algorithm should be monotonic for small positive δ .

5 Fitting Bradley Terry Models

Our criterion is

$$\phi(p) = ln\{L(p)\}.$$

Since $L(p)$ is a homogeneous function of degree zero $\sum_i p_i d_i = 0$. In fact $d_j = F_j$. So there are always positive and negative d_j unless all are zero. We require a function $f(d,\delta)$ which is defined for positive and negative d, where we take d to represent a partial derivative. Noting that all p_j^* must be positive a suitable choice should be governed by the fact that at the optimum $d_j^* = 0$, $j = 1,2,\ldots,J$.

We opt for $f(d,\delta) = \Phi(\delta d)$, so that iterations prove to be

$$p_j^{(r+1)} = \frac{p_j^{(r)}\Phi(\delta d_j^{(r)})}{\sum_i p_i^{(r)}\Phi(\delta d_i^{(r)})}$$

Coffee Example.
In this case $J = 8$ coffee types were compared yielding a total of $N = 728$ observations; i.e. $\sum\sum O_{ij} = 728$. A suitable δ is $\delta = 1/N$. In effect we are standardising the sample size to 1, through replacing observed by relative frequencies in the log-likelihood, and then taking $\delta = 1$.

Torsney (2004) reported the following results. Starting from $p_j^{(0)} = 1/J$, the numbers of iterations needed to achieve $\max|d_j| = \max|F_j| \leq 10^{-n}$, for $n = 0,1,\ldots,7$ respectively are 17, 21, 25, 32, 38, 45, 51, 59. The optimal p^* is (0.190257, 0.122731, 0.155456, 0.106993, 0.091339, 0.149406, 0.080953, 0.102865).
Iterations were monotonic.

6 Local Optimal Designing

We have not introduced any design variables. However we can pose the question: how many comparisons n_{ij} there should be between T_i and T_j? This of course is an exact design problem. The easier approximate design problem poses the question: what proportion λ_{ij} of such comparisons there should be?

This depends on our model. We focus on the Bradley Terry Model. The parameters are now $p_1, p_2, \ldots, p_J$. We wish good estimation of these. The information matrix is

$$M(\lambda) = \sum_{i<j}\sum \lambda_{ij} w_{ij} v_{ij} v_{ij}^T$$

where $v_{ij} = (e_i - e_j), e_i$ being the ith unit vector $w_{ij} = 1/(p_i + p_j)^2$.

We note the following properties:

1. $M(\lambda)$ has the form of the information matrix of a weighted linear model with weights w_{ij}. This happens with a wide range of generalised linear models.
2. $M(\lambda)$ depends on the p_i's (but only through the w_{ij}'s).
   We need provisional values for them. A conventional choice is $p_j = 1/J$.
   However we have maximum likelihood estimates This does not seem to have been considered in the literature before.
3. $M(\lambda)$ is singular. This is another manifestation of the fact that we only have observations on comparisons between treatments. We can only estimate differences

between treatments. This has implications for choice of design criteria. We must restrict consideration to good estimation of such differences (or other contrasts). This issue too appears to have been ignored in the literature.

Two feasible classes are:

$$\begin{aligned} D_L - \text{criteria}: &\quad \Psi(M) = -\log\det(LM^+L^T), \\ A_L - \text{criteria}: &\quad \Psi(M) = -trace(LM^+L^T). \end{aligned}$$

Here M^+ denotes the Moore-Penrose inverse of M and L defines a set of $(k-1)$ linearly independent differences between the p_i parameters.
The D_L-criterion would be invariant to any such choice of L.

In general a locally optimal design problem is, for given p, to choose λ optimally subject to $\lambda_{ij} \geq 0, \sum\sum_{i<j}\lambda_{ij} = 1$, i.e. solve Problem ($\mathfrak{P}$) for $\phi(\lambda) = \Psi\{M(\lambda)\}$ for some $\Psi\{\cdot\}$.

We need derivatives with respect to λ_{ij}, which we denote by d_{ij}, for optimality checking and numerical purposes. We have:

$$\begin{aligned} &\text{for the } D_L\text{-criterion,} &\quad d_{ij} = w_{ij}v_{ij}^T M^+ L^T (LM^+L^T)^{-1} LM^+ v_{ij} \\ &\text{for the } A_L\text{-criterion,} &\quad d_{ij} = w_{ij}v_{ij}^T M^+ L^T LM^+ v_{ij}. \end{aligned}$$

Of note is that these are positive, as is the case with all standard design criteria. For the multiplicative algorithm a feasible choice is $f(d,\delta) = d^{\delta}$, the original form of this function when the algorithm was first conceived for determining optimal designs. The choices of δ we opt for here correspond to choices which have been shown to be monotonic for the standard D-criterion and A-criterion, namely $\delta = 1, 1/2$ respectively.

Coffee example

We choose to determine locally optimal designs at the current maximum likelihood estimates; i.e. at p^* =
$(0.190257, 0.122731, 0.155456, 0.106993, 0.091339, 0.149406, 0.080953, 0.102865)$.
We use the following choices of $f(d,\delta)$: for the D_L-criterion: $f(d,\delta) = d$; for the A_L-criterion: $f(d,\delta) = d^{1/2}$.
Iterations begin at $\lambda_{ij}^{(0)} = 1/(J(J-1))$.
We take L to be the matrix defining the 7 differences $p_1 - p_j$, $j = 2, 3, \ldots, 8$.

We sumarise the implications if a further experiment is to be run and parameter values are in the region of the maximum likelihood estimates: for D_L-optimality no comparisons would be made between coffee types 1 and 3 and between coffee types 1 and 6; under both designs maximum weight is put on the comparisons between coffee types 1 and 7, which have the largest and smallest estimated Bradley Terry parameters; the A_L-optimal weights of the 7 comparisons with the first coffee type

exceed 0.07 while the remainder are less than 0.03, which is in keeping with the focus of the choice of L on differences with this coffee type.
For comparison we note that uniform weights of 1/28 = 0.0357143.

7 Discussion

There are several extensions of this work in respect of both parameter estimation and local optimal designing (arguably new): for rankings; for "no preference" options; for factorially structured treatments.

References

Bradley, R. A. and Terry, M. E. (1952). The rank analysis of incomplete block designs I, The method of paired comparisons. *Biometrika 39*, 324-345.
Bradley, R. A. and El-Helbawy, A. T. (1976). Treatment contrasts in paired comparisons: basic procedures with applications to factorials. *Biometrika 63*, 255-262.
Kiefer, J. (1974) General equivalence theory for optimum designs (approximate theory) *Annals of Statistics 2*, 849-879.
Kuk, A.C. Y. (1995). Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *Statistician 44*, 523-528.
Mandal, S. and Torsney, B. (2000) Algorithms for the construction of optimising distributions. *Communications in Statistics (Theory and Methods) 29*, 1219-1231.
Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology 21*, 384-400.
Torsney, B. (1988) Computing Optimizing Distributions with Applications in Design, Estimation and Image Processing *Optimal Design and Analysis of Experiments, Editors Y. Dodge, V.V. Fedorov, H.P. Wynn. North Holland.*, 361-370.
Torsney, B. (2004) Fitting Bradley Terry models using a multiplicative algorithm. *Proceedings in Computational Statistics (COMPSTAT2004, August 2004, Prague, Czeck Republic), Editor - Jaromir Antoch, Physica Verlag*, 214-226.
Torsney, B. and Alahmadi, A. M. (1992) Further developments of algorithms for constructing optimizing distributions. *Model Oriented data Analysis (V. Fedorov, W.G. Muller, I.N. Vuchkov Eds). Proceedings of 2nd IIASA- Workshop, St. Kyrik, Bulgaria, 1990. Physica Verlag* , 121-129.
Whittle, P. (1973). Some general points in the theory of optimal experimental design *J. Roy. Statist, Soc. B 35*, 123-130.