Nonparametric analysis of unbalanced paired-comparison or ranked data

Douglas Martin Andrews

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation


https://lib.dr.iastate.edu/rtd/9266

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book. These are also available as one exposure on a standard 35mm slide or as a 17” x 23” black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Nonparametric analysis of unbalanced paired-comparison or ranked data

Andrews, Douglas Martin, Ph.D.
Iowa State University, 1989
Nonparametric analysis of unbalanced paired-comparison or ranked data

by

Douglas Martin Andrews

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:
Signature was redacted for privacy.

In Charge of Major Work
Signature was redacted for privacy.

For the Major Department
Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1989

Copyright © Douglas Martin Andrews, 1989. All rights reserved.
TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION ........................................... 1
   Literature Review ............................................ 1
   Paired comparisons ........................................... 1
   Ranking methods ............................................. 5
   Scope of this Dissertation ................................... 11

CHAPTER 2. RELATED PAIRED-COMPARISON PROCEDURES
   FOR BALANCED OR UNBALANCED DATA .......................... 14
   Methods Using Powers of the Preference Matrix ............ 14
      Kendall-Wei approach ..................................... 14
      Cowden's approach ....................................... 15
   Least-Squares ................................................ 18
   Method ......................................................... 18
   Analysis of variance ....................................... 20
   Special case .................................................. 21
   Dual Scaling .................................................. 22
   Generalized Row-Sums ....................................... 26

CHAPTER 3. PROPOSED PROCEDURE AND ITS PROPERTIES 30
CHAPTER 4. EXTENSION OF PRESENT APPROACH TO UNBALANCED RANKED DATA

Comparison of Rank and Paired-Comparison Approaches

Special case: completely balanced designs

Special case: balanced incomplete block designs

Special case: partially balanced incomplete block designs

General Case

Numerical Examples

Sire data

CHAPTER 5. SUMMARY AND CONCLUSIONS

BIBLIOGRAPHY

ACKNOWLEDGEMENTS

APPENDIX A. COVARIANCE OF TWO SCORES: GENERAL CASE

APPENDIX B. METHOD FOR FINDING $\hat{\Sigma}^{-\frac{1}{2}}$

APPENDIX C. EXPRESSION FOR $Q$ FROM A GROUP DIVISIBLE DESIGN

APPENDIX D. RELATIONSHIP BETWEEN $Q$ AND $C$ FOR PBIB RANK DESIGNS
CHAPTER 1. INTRODUCTION

Suppose we have \( t \) objects \( C_1, \ldots, C_t \), and that objects \( C_i \) and \( C_j \) are judged pairwise in \( n_{ij} \) independent comparisons, for \( i, j = 1, \ldots, t; \ i \neq j \). Considerable attention has been given to the design and analysis of such 'paired-comparison' experiments. The simplest design would require that all pairs of objects be compared exactly once (i.e., all \( n_{ij} = 1 \)) or, more generally, an equal number of times (i.e., all \( n_{ij} = n \)). (Such 'completely balanced' designs correspond to Round Robins, in sporting tournament parlance.) Yet it is often inconvenient or impractical to judge all \( \binom{t}{2} \) possible pairs of the \( t \) objects. 'Incomplete' designs, in which \( n_{ij} = 0 \) for some pairs \( (i, j) \), can yet be highly structured; for example, we might require that each object be judged in an equal number of comparisons, thus giving us a 'partially balanced' design. But the design need not have any apparent structure. It is these most general experiments — possibly incomplete or otherwise unbalanced — on which this paper will focus.
Paired comparisons

In the analysis of paired-comparison experiments it is often of interest to rank the objects or perhaps rate them on some linear scale. If the objects are compared by distinct judges with possibly different points of view, it may also be of interest to scale the judges in some way, or at least identify those judges whose preferences are markedly different from the other judges' preferences. Scheffé (1952) presents an analysis of variance for paired-comparison data which accounts for such a 'judge effect', and Nishisato (1980) develops a tool to scale the objects and the judges simultaneously. In this paper, however, we will assume that the judges are interchangeable, so that all of the comparisons may as well have been made by a single judge. In other words, we will assume that, in each of the \( n_{ij} \) comparisons between \( C_i \) and \( C_j \), \( C_i \) will be preferred with probability \( \pi_{ij} \in [0, 1] \). To begin with, we will not permit ties so that \( C_j \) will be preferred over \( C_i \) with probability \( \pi_{ji} = 1 - \pi_{ij} \).

We thus have \( \binom{N}{2} \) 'free' parameters \( \{\pi_{ij} : i < j\} \) to describe the probability with which each object will be preferred to each other. When our goal is to rank the objects, it becomes helpful to assume that each object has a certain 'intrinsic worth' \( \theta_i \) which we can estimate. (See, e.g., Brunk (1960).) These 'worth' or 'merit' parameters \( \{\theta_i\} \) are sometimes used to generate the preference probabilities \( \{\pi_{ij}\} \) in the following way:

\[
\pi_{ij} = \Pr\{C_i \text{ is preferred to } C_j\} = H(\theta_i - \theta_j),
\]

(1.1)
where $H(\cdot)$ is the cdf of a random variable distributed symmetrically about zero. If the preference probabilities $\{\pi_{ij}\}$ can be so described, we say they (or the characteristic under study) satisfy a **linear model**. One well-known special case of the linear model is the Bradley-Terry model, originally explored by Zermelo (1929), for which

$$H(\theta_i - \theta_j) = \frac{1}{4} \int_{-(\theta_i-\theta_j)}^{\infty} \text{sech}^2\left(\frac{1}{2} x\right) dx,$$

and from which it follows that $\pi_{ij} = \frac{\theta_i}{\theta_i + \theta_j}$. Many authors (e.g., Wilkinson (1957)) have proposed estimation procedures for the $\{\theta_i\}$ and tests based on the joint likelihood function, and some (e.g., Dykstra (1960)) have done this even for unbalanced experiments. Another notable special linear model is due to Thurstone, which takes for $H(\cdot)$ the cdf of a standard normal variate. Again, many have investigated estimation procedures for the $\{\theta_i\}$ under this model as well (e.g., Thompson (1975) for unbalanced experiments).

Notice that we could permit the judge(s) to state 'no preference', i.e., a tie. This would give the judge(s) an (ordered) scale of three 'points' from which to choose on each comparison:

- preference for $C_i$
- no preference
- preference for $C_j$.

A common practice in paired-comparison analysis is to treat a tie as 'half' a preference in favor of each object; thus if the 'winner' in a clear preference receives a score of 1 and the 'loser' 0, each would receive $\frac{1}{2}$ in the event of a tie. (Another approach gives
scores equal in magnitude but opposite in sign: the winner would be given, say, 1 and
the loser -1. An unfortunate consequence of such a system is that a tie is often treated
as if no comparison were made at all, since each object would logically be given 0.)
An extension of this idea of admitting ties is to let the judges declare a magnitude of
preference as well: Scheffé (1952) suggests using a 7- or 9-point scale, ranging from
'strong', 'moderate', and 'weak' preference for $C_j$, through 'no preference' and on
up to 'strong' preference for $C_i$. The points on Scheffé's 7-point scale are then given
values \{-3, -2, -1, 0, 1, 2, 3\} accordingly.

Yet the scale need not even be discrete; a judge might have a single 'point' to
split between the two objects for any given comparison. Or the 'encounter' between
two objects might be a measurement on some continuous scale. Most of the methods
which follow can take data in any such form without much loss of generality. So for
clarity and simplicity we will restrict our attention to the 2-point scale, for which we
may simply assign a value of 1 for $C_i$ and 0 for $C_j$ if $C_i$ is preferred to $C_j$ (and vice
versa if $C_j$ is preferred to $C_i$) for each of the $n_{ij}$ comparisons of the two objects.

We can then record the results of a paired-comparison experiment in a 'prefer­
ence' matrix $A = ((a_{ij}))$, where $a_{ij}$ is the number of preferences in favor of $C_i$ over
$C_j$. (Assume that no object is compared with itself; for convenience, take $a_{ii} = 0$.)
In a balanced experiment it is well-established that the row-sums of $A$ (i.e., the total
number of preferences in favor of each object) give an appropriate and meaningful
ranking. (Huber (1963) notes some optimal decision-theoretical properties of the
row-sum procedure.) But for incomplete or otherwise unbalanced experiments, the
row-sum ranking is no longer appropriate because of
• the (possibly) different numbers of comparisons involving each object, and

• the (possibly) varied strength of each object's opposition.

Sensible analysis of unbalanced data must address these concerns. In Chapter 2 we will highlight those paired-comparison procedures which account for these and other aspects of unbalanced data.

**Ranking methods**

Suppose instead that the objects are not compared pairwise, but instead are ranked in 'blocks', where \( k_j \) of the \( t \) objects are ranked in block \( B_j \), for \( j = 1(1)b \). The \( b \) rankings could be made by the same judge or by as many as \( b \) different judges; but we assume, as in the previous subsection, that the judges are interchangeable, so that distinct points of view are impossible.

There are several links between paired comparisons and ranks worth noting. The most obvious is that a paired comparison can be thought of as a ranking among two objects. In this light, any paired-comparison design can be analyzed as an incomplete block design with as many blocks as there are comparisons, and with all \( k_j = 2 \).

What might not be quite so obvious is that, conversely, ranked data may be analyzed as constrained paired-comparison data, since the ranking \( (C_1, C_2, C_3) \), say, can be 'decomposed' into comparisons \( \{C_1 \rightarrow C_2, C_2 \rightarrow C_3, C_1 \rightarrow C_3\} \), where '\( \rightarrow \) is read 'is preferred to'. It is crucial to note that, within any triad of ranked objects, the result of one of the constituent paired comparisons will be determined by the results of the other two. That \( C_1 \) is ranked ahead of \( C_2 \) and \( C_2 \) ahead of \( C_3 \) assures us that \( C_1 \) is ranked ahead of \( C_3 \); hence \( C_1 \rightarrow C_3 \). This forced transitivity is the main
distinction between rankings and paired comparisons. For in a paired-comparison experiment, we could easily have observed the intransitive 'circular triad' \( \{C_1 \rightarrow C_2, C_2 \rightarrow C_3, C_3 \rightarrow C_1\} \). Or, if the design is incomplete, one of the three pairs might not have been compared.

Bearing these differences in mind, we would like to know how reliably we may use paired-comparison methods to analyze ranked data. And since we will focus on paired-comparison methods developed for unbalanced data, we seek, for purposes of comparison, methods for analyzing unbalanced ranked data as well.

Denote by \( r_{ij} \) the rank for object \( C_i \) in block \( B_j \). If each of the \( b \) judges ranks all \( t \) objects, we can surely look to the rank sums \( \{r_i\} \) or rank means \( \{\bar{r}_i\} \) as appropriate scores for the objects (just as we could use the row sums of the preference matrix for balanced paired-comparison designs). This is the approach taken by Friedman (1937), who uses intrablock ranks in place of the actual observations in the analysis of variance in order to free the analysis from the usual normality assumptions. Noting that the sampling distribution of the rank means \( \{\bar{r}_i\} \) will be approximately normal "so long as the number of [judges] is not too small", he uses the sum of squares of these rank means about their mean, suitably standardized,

\[
X^2_F = \frac{12b}{t(t+1)} \sum_{i=1}^{t} \left( \bar{r}_i - \frac{1}{2}(t+1) \right)^2,
\]

as a test statistic for the equality of the objects. He points out that, as \( b \to \infty \), \( X^2_F \) has limiting \( \chi^2 \) distribution with \( t - 1 \) degrees of freedom, and then compares the efficiency of the resulting test with the efficiency of the usual normal-theory test from the analysis of variance. He also tabulates the exact null distribution of \( X^2_F \) for a few values of \( t \) and \( b \), both as a reference for small-sample testing and as an illustration.
of how quickly this distribution approaches its limit.

Much of the literature concerning multiple rankings focuses on measures of agreement among the blocks, or judges. Kendall's τ, Spearman’s ρs, and the usual product-moment correlation r are among the more familiar statistics for assessing the agreement between b = 2 complete rankings. (Incidentally, Kendall (1962) points out that each of these measures is a special case of a 'generalized correlation coefficient' Γ.) It should be clear that testing for agreement among judges is not entirely unrelated to testing for differences among the objects, since agreement among the judges typically reflects some perceived differences among the objects (though spurious correlations among the judges’ rankings could certainly arise, just as treatment differences could arise ‘by chance’). Indeed, for b = 2 Friedman shows that $X^2_p = (t - 1)(1 + ρ_s)$, thus establishing a direct link between these two statistics.

Noting then how agreement among judges relates to dispersion of the objects’ rank-sums, Kendall generalizes the notion of agreement for $b \geq 2$ by defining a ‘coefficient of concordance’

$$W = \frac{\text{observed value of } S}{\text{maximum possible value of } S},$$

where $S$ is the sum of squared rank-sums about their mean. Note that $W \in [0,1]$, with greater values of $W$ indicating greater agreement among the judges. When the $b$ rankings are complete (i.e., all the $k_j = t$), Kendall shows that (1.3) becomes

$$WBal = \frac{S_{Bal}}{\frac{1}{12} b^2 t(t^2 - 1)},$$

where

$$S_{Bal} = \sum_{i=1}^{t} [r_i - b \frac{t + 1}{2}]^2.$$

(1.5)
From this and (1.2) it follows that

$$b(t - 1)W_{Bal} = \frac{S_{Bal}}{\frac{1}{12}b(t + 1)}$$

is equal to Friedman's $\chi^2_n$ and hence has an asymptotic $\chi^2_{t-1}$ distribution.

Kendall also suggests using $W$ for balanced incomplete block (BIB) designs, in which each object appears in only $m$ of the $b$ blocks, and in which each object appears with each other object in $n$ blocks. For such designs, (1.3) becomes

$$W_{BIB} = \frac{S_{BIB}}{\frac{1}{12}nt(t^2 - 1)},$$

where now

$$S_{BIB} = \sum_{i=1}^{t} \left[ r_i - m \frac{t + 1}{2} \right]^2.$$

Durbin (1951) examines $W$ more closely in this context, finding its exact mean and variance, and approximating the distribution of

$$\frac{n(t^2 - 1)}{k + 1}W_{BIB} = \frac{S_{BIB}}{\frac{1}{12}nt(k + 1)}$$

with Beta and $\chi^2$ distributions. Schucany and Beckett (1976) work within the BIB framework as well; they develop a statistic for testing the agreement between two groups of judges.

Gulliksen and Tucker (1961) note that, although the paired comparisons derived from a single ranking form a transitive set, the constituent paired comparisons pooled from two or more rankings might contain considerable intransitivity. The authors count the number $d$ of circular triads in the collection of paired comparisons derived from a BIB experiment, and base a test on Kendall’s ‘coefficient of consistence’,
a linear function of \( d \). But BIB ranking designs 'decompose' into balanced paired-comparison designs, for which Kendall has shown \( d \) to be itself a linear function of the sum of squared (paired comparison) row-sums \( \{a_i\} \). And since, as will be shown in (4.1), these row sums are equivalent to the corresponding rank sums \( \{r_i\} \) for balanced designs, it is clear that the resulting test is equivalent to a test based on the sum of squared rank sums and hence to a test based on Kendall's \( W_{BIB} \) as well.

Like Gulliksen and Tucker, Straton (1975) works with paired-comparison data obtained from ranking experiments, though he broadens his scope to include partially balanced incomplete block (PBIB) designs. Because each object still appears in an equal number of blocks, Straton sees no need to look beyond the simple rank sums as scores for the objects. We will argue in Chapter 4 that such a method fails to take into account the (possibly) differing caliber of each object's opposition, much as the simple row sums of the preference matrix are deemed inadequate for unbalanced paired-comparison data.

A number of authors have taken a parametric approach to the general situation with differing block sizes \( \{k_j\} \) or with some objects appearing in more blocks than other objects. (For example, Levitt (1975) and Pettitt (1983) apply Luce's multiple comparison generalization of the Bradley-Terry model, and Rai (1971) applies a similar model to a certain class of unbalanced designs involving triad comparisons.) The chosen linear model then specifies a likelihood function, maximization of which (through any standard function-optimizing technique) leads to scores for the objects.

Benard and van Elteren (1953) tackle the general problem of \( m \) unbalanced rankings nonparametrically. (In fact, their method covers not just multiple rank orders
but more general unbalanced block designs as well, in which several observations on
an object may appear in a block.) They recognize that, in an unbalanced experiment,
objects might appear in widely differing numbers of blocks. And as even the most
mediocre object can boost its paired-comparison row-sum simply by being included
in a relatively large number of comparisons, so can such an object boost its rank-sum
simply by appearing in a relatively large number of blocks. To mitigate this effect,
the authors 'center' each rank \( r_{ij} \) by subtracting \( \frac{1}{2}(k_j + 1) \), the mean of the \( k_j \) ranks
in the \( j^{th} \) block, to get a 'reduced' rank \( u_{ij} \). They then use
\[
u_i = \sum_j u_{ij} = \sum_j [r_{ij} - \frac{1}{2}(k_j + 1)]
\]
as the score for \( C_i \) (for \( i = 1(1)t \)), and find the null covariance matrix \( \Sigma_u \) of the score
vector \( u \) in terms of the block sizes \( \{k_j\} \). Since \( u'1_t = 0 \) and hence \( \Sigma_u1_t = 0 \), they
eliminate one of the \( \{u_i\} \) to avoid singularity, and use
\[
B = \bar{u}'\tilde{\Sigma}_u^{-1}\bar{u}
\]
for testing equality of the objects, where \( \bar{u} \) is the vector of the \( t - 1 \) chosen scores, and
\( \tilde{\Sigma}_u \) is the covariance matrix of \( \bar{u} \). They show that \( B \to \chi^2_{t-1} \) as \( b \to \infty \) when the
objects form a connected set, and mention the special cases of \( b \) complete rankings and
the BIB designs discussed earlier. For the completely balanced case, \( B \) is equivalent
to Kendall's \( W_{Bal} \) from (1.4) and hence to Friedman's \( X^2_b \) from (1.2); for the BIB
designs, \( B \) is equivalent to \( W_{BIB} \) from (1.7).

Prentice (1979) builds on Benard and van Elteren's work, giving an expression
for the asymptotic non-centrality parameter of \( B \) under a certain local alternative
hypothesis. He furthermore suggests that the centered ranks \( \{u_{ij}\} \) are unsatisfactory
if the block sizes are much different, in that even mediocre performance in large blocks can be rewarded as much as superior performance in smaller blocks. (As an example, note that \( u_{ij} = 5 \) both for an object ranked 1st of 11 and for an object ranked 46th of 101.) He then scales each centered rank \( u_{ij} \) to get a standardized rank

\[
y_{ij} = \frac{u_{ij}}{k_j + 1} = \left( \frac{r_{ij}}{k_j + 1} - \frac{1}{2} \right),
\]

and sums these over blocks to get the objects' scores \( \{y_i\} \). Note that \( y_{ij} \in (-\frac{1}{2}, \frac{1}{2}) \), regardless of the block size \( k_j \). Prentice then proposes the statistic

\[
C = \bar{y}' \Sigma y \bar{y},
\]

shows that it has an asymptotic \( \chi^2_{l-1} \) distribution, finds an expression for its asymptotic non-centrality parameter, and uses this to show that the resulting test is more efficient than the test based on \( B \), for a certain class of balanced designs.

Prentice cites algebraic convenience to justify his choice of \( (k_j + 1)^{-1} \) in standardizing the ranks in (1.11), and admits that other standardizations are possible. Rai (1987) uses \( k_j^{-1} \), though he standardizes the uncentered ranks. Skillings and Mack (1981) use \( (k_j + 1)^{-\frac{1}{2}} \), which simplifies the covariance matrix of the resulting scores; they go on to examine multiple-comparison procedures and to tabulate the small-sample distribution of a statistic analogous to Prentice's \( C \) for a few specific unbalanced designs. Provided that the block sizes are not radically different, their scores will be quite comparable to Prentice's \( \{y_i\} \).
Scope of this Dissertation

Much of the paired-comparison literature focuses on the design of paired-comparison experiments. Yet even within the tidy framework of complete balance it is not trivial to construct a layout which is optimal in some sense, e.g., with respect to the order of presentation to the judge(s). John, Wolock, and David (1972) consider a rather broad class of partially balanced designs, and have listed those which are most 'efficient' in a certain sense. This paper will not deal with such design considerations, and in fact will deal with balanced and partially balanced experiments only as special cases of the more general paired-comparison experiments.

As discussed earlier, we regard the judges as indistinguishable or interchangeable, and we generally restrict these judges to record their preferences on a two-point scale. Furthermore, we will consider only those ranking or scoring procedures which are not tied to any particular linear model. In other words, we will consider methods which are applicable even when the \( \{ \pi_{ij} \} \) are not confined by a relation such as (1.1). In this sense these procedures can be considered 'nonparametric'.

Some existing methods for analysis of unbalanced paired-comparison data will be examined in closer detail in Chapter 2. Two of these methods involve powers of the preference matrix \( A \). A third approach appeals to the familiar least-squares criterion, and another is related to principal components analysis. Finally we will review a recent method for which the scores are the solution to a system of linear equations derived from the row-sums and from the number of comparisons on each pair of objects.

In Chapter 3 we will discuss our proposed method and will explore some of its
properties. In particular, we will derive expressions for the objects' scores under the general unbalanced case and under the special cases of complete and partial balance. Expectations, variances, and covariances of the scores will be calculated in each of these contexts; the resulting asymptotic distribution of the scores will lead to several tests of hypotheses. A modification of this scoring procedure will be introduced which alleviates some problems encountered when there are one or more pairs of objects which have not been compared. The proposed method is then compared to the existing methods in a few small numerical examples.

The method will be extended in Chapter 4 to analyze unbalanced ranked data; such analysis will be compared with existing methods of rank analysis tailored to unbalanced data. It will be seen that, although a few existing methods can account for differing block sizes and a few other basic aspects of incomplete data, they do not address all the problems of unbalanced data of concern in this paper. More numerical examples follow to illustrate the shortcomings of the existing rank approaches and the success of the proposed method in dealing with these problems.
CHAPTER 2. RELATED PAIRED-COMPARISON PROCEDURES
FOR BALANCED OR UNBALANCED DATA

Methods Using Powers of the Preference Matrix

Kendall-Wei approach

Wei (reported by Kendall (1955)) investigated a 're-allocation' of the simple row-sum scores, for the special case of a balanced experiment with all $n_{ij} = 1$. His score for an object $C_i$, say, is the sum of the row-sums of the objects which were 'defeated' by $C_i$, to use the language of sporting tournaments. (In fact, this method of scoring has long been used in chess tournaments to break ties between players with equal row-sums.) This amounts to taking as scores the row-sums of $A^2$. Clearly this method rewards wins against strong opponents. Kendall suggests that the resulting scores are helpful in breaking ties in the original row-sum scores. And although he feels that this “is as far as one would wish to go on practical grounds,” he nonetheless notes that this re-allocation can be continued *ad infinitum*, reasoning that the “continual re-allocation of scores is equivalent to taking successive powers of the [preference] matrix.” Kendall then shows that the corresponding repeated powering of the preference matrix does in fact converge to a limiting ranking, provided that $A$ is *indivisible*, i.e., if, in every possible partition of the objects into two non-empty sets,
some object in each set has won at least once from some object in the other set. He
extends these ideas to the more general balanced experiment with all $n_{ij} = n > 1$.
But once we get away from this balance, it is possible for an object to boost its
ranking simply by being judged in many comparisons. Moreover, the rankings may
not be faithfully reversed if 'wins' and 'losses' are interchanged.

Cowden's approach

Method  Cowden (1974) modifies the Kendall-Wei approach in proposing
the following iterative procedure for arriving at a set of scores:

- Start with a vector of 'win scores' $u^{(0)}$ and a vector of 'loss scores' $v^{(0)}$, where
  
  $$u_i^{(0)} \in (0, 1), \quad \text{and} \quad v_i^{(0)} = 1 - u_i^{(0)} \quad \text{for all } i.$$  

  (The final (stable) win scores will be larger or smaller according as $\sum_{i=1}^{t} u_i^{(0)}$
  is larger or smaller; the final ranks, however, will be little affected by the choice
  of the initial vector. Cowden suggests taking $u_i^{(0)} = v_i^{(0)} = \frac{1}{2}$ for convenience.)

- Define $A$ as before, inserting zeros for the diagonal elements $a_{ii}$ and for any
  $a_{ij}$ for which $C_i$ and $C_j$ have not been compared. Define $B = A'$ as the matrix
  of 'losses', since $b_{ij} = a_{ji}$ is the number of $C_i$'s losses to $C_j$.

- Iterate for $k = 1, 2, \ldots$ by setting
  
  $$p^{(k)} = Au^{(k-1)} \quad (2.1)$$
  
  $$q^{(k)} = Bv^{(k-1)}, \quad (2.2)$$
and taking

\[ u_i^{(k)} = \frac{p_i^{(k)}}{p_i^{(k)} + q_i^{(k)}} \]  \hspace{1cm} (2.3)

\[ v_i^{(k)} = 1 - u_i^{(k)} \]  \hspace{1cm} (2.4)

for \( i = 1, \ldots, t \).

The rankings usually stabilize within a few iterations, and the scores themselves stabilize shortly thereafter. Cowden claims that "a meaningful convergence of the ranks and scores will result" if \( A \) is indivisible. The relationship between \( u_i^{(\cdot)} \) and \( v_i^{(\cdot)} \) in (2.4) assures us of the desirable result that interchanging wins and losses reverses the ranking of the objects. Cowden's method usually gives fairly sensible results, though even in the balanced case it often gives different scores to objects with the same row-sum score, nonsensically breaking these row-sum ties in favor of the object(s) whose wins were over *weaker* opposition. Consider, for example, the following data with 10 comparisons per pair:

<table>
<thead>
<tr>
<th>row-sum</th>
<th>Cowden's score</th>
</tr>
</thead>
<tbody>
<tr>
<td>([ \ast \ 6 \ 6 \ 6 ])</td>
<td>([18 \ .576])</td>
</tr>
<tr>
<td>([4 \ \ast \ 5 \ 9])</td>
<td>([18 \ .584])</td>
</tr>
<tr>
<td>([4 \ 5 \ \ast \ 9])</td>
<td>([18 \ .584])</td>
</tr>
<tr>
<td>([4 \ 1 \ 1 \ \ast])</td>
<td>([6 \ .257])</td>
</tr>
</tbody>
</table>

Note that half of \( C_2 \) and \( C_3 \)'s wins came at the expense of the inferior object \( C_4 \), whereas \( C_1 \) won 6 of 10 from each of the others; yet Cowden's scores rank \( C_1 \) below \( C_2 \) and \( C_3 \). (Incidentally, Kendall's method breaks these row-sum ties in favor of \( C_1 \), as the row-sums of \( A^2 \) are \((252, 216, 216, 108)\).)
Simplification for early iterations

Suppose we begin with \( u^{(0)} = v^{(0)} = \frac{1}{2} 1_t \), as Cowden suggests. At the first iteration we then have \( p^{(1)} = \frac{1}{2} A 1_t \) and \( q^{(1)} = \frac{1}{2} B 1_t \), so that \( p_i^{(1)} \) and \( q_i^{(1)} \) are, respectively, half the number of comparisons won and lost by \( C_i \). Then from (2.3) and (2.4) we see that the first win scores \( u^{(1)} \) and loss scores \( v^{(1)} \) are simply the overall proportion of comparisons won and lost by each object.

Further simplifications, however, are impossible, since the denominator in (2.3) does not have a simplifiable form beyond the first iteration. Even under complete balance (i.e., all \( n_{ij} = n \), so that each object has \( r = n(t - 1) \) comparisons) we have

\[
p^{(1)} + q^{(1)} = \frac{1}{2} A 1_t + \frac{1}{2} B 1_t
\]

\[
= \frac{1}{2} (A + B) 1_t
\]

\[
= \frac{r}{2} 1_t,
\]

so that \( u^{(1)} = \frac{1}{r} A 1_t \) and \( v^{(1)} = \frac{1}{r} B 1_t \) are equivalent, respectively, to the row sums of \( A \) and \( B \); but even by the second iteration we then get

\[
p^{(2)} + q^{(2)} = A u^{(1)} + B v^{(1)}
\]

\[
= \frac{1}{r} A^2 1_t + \frac{1}{r} B^2 1_t
\]

\[
= \frac{1}{r} \left[ A^2 + (A')^2 \right] 1_t
\]

\[
= \frac{1}{r} \left[ (A + A')^2 - 2AA' \right] 1_t.
\]

Since \( (A + A') 1_t = r 1_t \) we have \( (A + A')^2 1_t = r^2 1_t \), and hence

\[
p^{(2)} + q^{(2)} = r 1_t - \frac{2}{r} AA' 1_t.
\]

Even under complete balance \( AA' \) will not, in general, simplify.
Least-Squares Method

Gulliksen (1956) took a least-squares approach to the problem of incomplete paired-comparison data. In an approach similar to the linear models described in Chapter 1, he presumes that there are some 'scale values' \( \{ \theta_i \} \) which represent the objects' worths in some sense, and considers the discrepancy between the predicted difference \( \theta_i - \theta_j \) and an observed difference \( d_{ij} \) for each pair \((i, j)\) which has been compared. (Gulliksen suggests for the observed difference a normal score based on \( p_{ij} \), the proportion of judgments in favor of \( C_i \) over \( C_j \). Kaiser and Serlin (1978) note that we only need require \( d_{ji} = -d_{ij} \).) The problem is then to determine the values \( \{ \hat{\theta}_i \} \) which minimize the sum of the squares of these discrepancies, namely,

\[
\sum^* (d_{ij} - (\hat{\theta}_i - \hat{\theta}_j))^2,
\]

where the sum \( \sum^* \) ranges over all pairs \((i, j)\) that have been compared. To this end, Gulliksen defines

\[
m_i = \text{number of objects with which } C_i \text{ has been compared},
\]

\[
N = \text{diagonal matrix with } \frac{1}{m_i + 1} \text{ as the } i^{th} \text{ diagonal entry},
\]

\[
M = ((m_{ij})), \text{ where } m_{ij} = \begin{cases} m_i & i = j \\ -1 & \text{if } C_i \text{ and } C_j \text{ have met} \\ 0 & \text{otherwise} \end{cases}
\]

\[
z = D1, \text{ the vector of row-sums of the matrix of observed differences } \{d_{ij}\}.
\]

His procedure is as follows:
• Choose an initial vector \( \hat{\theta}^{(0)} \). (Starting with a null vector assures that \( \hat{\theta}^{(1)} = \overline{N}z \), the average of the observed differences within each row.)

• Iterate, for \( k = 1, 2, \ldots \), by forming the predicted values \( \hat{M}\hat{\theta}^{(k-1)} \) and the discrepancies \( z - \hat{M}\hat{\theta}^{(k-1)} \), then updating the scores by adding the average of the discrepancies for each score:

\[
\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} + N(z - \hat{M}\hat{\theta}^{(k-1)}).
\] (2.5)

Note that

\[
\hat{\theta}_i^{(k)} = \hat{\theta}_i^{(k-1)} + \frac{1}{m_i + 1} \left[ z_i - \left( m_i\hat{\theta}_i^{(k-1)} - \sum_j^{(i)} \hat{\theta}_j^{(k-1)} \right) \right] \\
= \hat{\theta}_i^{(k-1)} + \frac{1}{m_i + 1} \left[ z_i - \sum_j^{(i)} \left( \hat{\theta}_i^{(k-1)} - \hat{\theta}_j^{(k-1)} \right) \right] \\
= \hat{\theta}_i^{(k-1)} + \frac{1}{m_i + 1} \sum_j^{(i)} \left[ d_{ij} - \left( \hat{\theta}_i^{(k-1)} - \hat{\theta}_j^{(k-1)} \right) \right],
\] (2.6)

where \( \sum_j^{(i)} \) denotes the sum over all \( C_j \) which have 'met' \( C_i \). From this form it is much clearer how the "average of the discrepancies for each scale value is then used to correct that value."
Analysis of variance

We can represent the corresponding partitioning of the total sum of squared differences in the standard analysis of variance format:

<table>
<thead>
<tr>
<th>source</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted</td>
<td>$t - 1$</td>
<td>$\sum^*(\hat{\theta}_i - \hat{\theta}_j)^2$</td>
</tr>
<tr>
<td>discrepancies</td>
<td>$M - t + 1$</td>
<td>$\sum^*[d_{ij} - (\hat{\theta}_i - \hat{\theta}_j)]^2$</td>
</tr>
<tr>
<td>total</td>
<td>$M$</td>
<td>$\sum^* d_{ij}^2$</td>
</tr>
</tbody>
</table>

\[ (2.7) \]

where $M = \frac{1}{2} \sum^i \sum^j m_{ij}$ is the total number of pairs compared at least once. For the special case of a single complete replication (i.e., all $n_{ij} = 1$), this is identical to Scheffé's (1952) analysis of variance for paired comparisons, if we ignore Scheffé's 'order effects'. Scheffé's 'deviations from subtractivity' would then be confounded with the residuals, and would become part of Gulliksen's 'discrepancies'.

Kaiser and Serlin continue along these lines, showing that a necessary and sufficient condition for a solution is that the objects form a connected set. They furthermore explore

\[ r^2 = \frac{SS(\text{predicted})}{SS(\text{total})} = \frac{\sum^*(s_i - s_j)^2}{\sum^* d_{ij}^2} \]

as a measure of the 'internal consistency' of the data. This $r^2$ will, as expected from the analysis of variance, lie between 0 and 1, with greater values indicating higher consistency. But the authors dutifully point out that this statistic should not be interpreted as a sign of 'stability', since it might be artificially close to 1 if there
were just barely enough data to connect the objects. Indeed, if there are just \( t - 1 \) preferences connecting the \( t \) objects, \( r^2 \) would always be 1.

**Special case**

The least-squares procedure simplifies greatly if all pairs have been compared (at least once). In that case we have

\[
m_i = t - 1 \quad \forall i,
\]

\[
N = \frac{1}{t} I,
\]

\[
M = \begin{bmatrix}
t - 1 & -1 & \cdots & -1 \\
-1 & t - 1 & \cdots & -1 \\
\vdots & \vdots & \ddots & \vdots \\
-1 & -1 & \cdots & t - 1
\end{bmatrix} = tI - J.
\]

So now

\[
I - NM = I - \frac{1}{t} I (tI - J) = I - I + \frac{1}{t} J = \frac{1}{t} J.
\]

From this and (2.5) we get

\[
\hat{\theta}^{(k)} = N z + (I - NM)\hat{\theta}^{(k-1)} = \frac{1}{t} (z + J\hat{\theta}^{(k-1)}).
\]

The scores may certainly be scaled so that their sum is zero. If at any iteration \( k - 1 \), say, the scores do indeed sum to zero, we get

\[
J\hat{\theta}^{(k-1)} = 0, \text{ leaving}
\]
\[ \hat{y}(k) = \frac{1}{t} z, \]

the average row-sum score. If, in particular, the initial vector is so chosen, the process will terminate after one iteration with these averaged row-sums. If the elements of the initial vector sum to \( c \), say, then \( J\hat{y}(0) = c1 \), and the process stops immediately with scores equivalent to the row-sum scores, merely shifted over by \( c/t \).

Note that, for the above simplifications, we have specified only that all pairs be compared. The objects need not have been involved in the same number of comparisons. So although we could think of this as a 'complete' experiment, it need not be 'balanced' in the traditional sense.

**Dual Scaling**

Nishisato (1980) has developed a powerful tool for analysis of a fairly wide range of data, including data from paired-comparison experiments. His 'dual scaling' procedure extends easily to handle missing data, incomplete designs, and preference scales with three or more points. Not only can the method rate the objects on a single linear scale, but it can also rate them on higher-dimensional scales as well (as in principal components analysis). Its application to all paired comparison experiments may, however, be quite dangerous, since fundamental to the dual scaling approach is the assumption that the comparisons are made by a heterogeneous group of judges. Data from a given judge will be weighted by the concordance of the judge's scores with those of the other judges. This is highly undesirable in those cases when we assume that the judges are identical or interchangeable, and their sets of comparisons are viewed as replicates. Since this assumption is made quite frequently in paired-
comparison situations, care must be taken not to misapply Nishisato's procedures.

Nishisato points out that “The dual scaling approach to ranking and paired-comparison data was pioneered by Guttman,” whose guiding “principle is that of internal consistency.” It is in this spirit that Nishisato characterizes dual scaling as “a procedure to find sets of paired-comparison judgments that produce transitive relations.” In Guttman’s words, the idea is to minimize “the variation within [judges], compared with that within the group as a whole.” The desired end is “the quantification ... best able to reproduce the judgment of each person in the population on each comparison.” It is important to note that the ‘quantification’ to which he refers is the scaling of both objects and judges. The scores for objects alone will not necessarily give the most meaningful ranking of the objects themselves; they are merely the most discriminative scores. Taken with the scores for the judges, one can then predict the response of a given judge on a given comparison.

But when the aim is not the reproduction or prediction of individual judgments but rather a (linear) assessment of the worth of the objects, the dual scaling approach is not at all appropriate. Consider Nishisato’s example with 4 objects and 10 judges, data from which are given below in preference-matrix format:

\[
\begin{array}{ccc}
-8 & 7 & 3 \\
2 & -9 & 2 \\
3 & 1 & -3 \\
7 & 8 & 7 \\
\end{array}
\]  

\begin{tabular}{c|c|c}
\textbf{scores} & \textbf{row sums} \\
0.121 & 18 \\
-0.036 & 13 \\
-0.276 & 7 \\
0.095 & 22 \\
\end{tabular}

(2.9)

Nishisato's procedure ranks the objects in the order \( C_1 > C_4 > C_2 > C_3 \), despite
the fact that $C_4$ was preferred in at least 70% of its comparisons with each of the other objects. He justifies the ranking in the following way:

The key lies in the internal consistency .... [The] judgment $[C_1 > C_4]$ comes from [judges] 6, 9, and 2, whose responses are highly consistent as we can see from their optimal scores .... [These] three [judges] had a greater impact on the maximization of the internal consistency than those of the remaining seven [judges]. This answer reflects the basic principle of our procedure, that is, duality. As the optimal scores for the subjects are proportional to the mean responses weighted by the scale values for the [objects], the scale values are proportional to the mean responses weighted by the [judges'] optimal scores.

Data from judges 3 and 8 in this example received negative weights because of their disagreement with the other judges. That some comparisons are considered more valuable than others is clearly undesirable if we have assumed that the judges are interchangeable and if our sole aim is to scale the objects.

Nishisato compares his solution to the column means of his subject-by-stimulus (i.e., judge-by-object) matrix, which are equivalent to the row-sums of the preference matrix. He mentions that $C_4$ will be scaled higher than $C_1$ whenever, as was the case in the example in (2.9), 70% of the judges preferred $C_4$ to $C_1$. This claim is certainly true for his data, since $C_4$ and $C_1$ had identical records against the remaining two objects. But in general this may not be the case, as illustrated by the following
completely balanced example:

\[
\begin{pmatrix}
\text{preference matrix} & \text{row-sums} \\
- & 9 & 9 & 3 & 21 \\
1 & - & 5 & 9 & 16 \\
1 & 5 & - & 9 & 16 \\
7 & 1 & 1 & - & 9
\end{pmatrix}
\]

Note that \( C_4 \) was preferred to \( C_1 \) in 70% of their comparisons, but that \( C_4 \) would be scaled last based on row sums, which are known to give a meaningful ranking of the objects for completely balanced experiments.

Nishisato states that a system that weights the judges equally is a “step backward” and “does not fully utilize the information contained in the data.” It is true that, once we accumulate all the judges’ preferences in a single matrix, we have lost the performance record of each judge. But, ironically, Nishisato’s method loses valuable information in that the function he maximizes involves the data only through the judge-by-object matrix, in which the \((i,j)\) element is the number of wins minus the number of losses for object \( C_j \) from judge \( i \). A judge could state a different set of preferences yet still be treated the same under such a method. Consider the comparisons by four judges, data from which are given below in preference-matrix
Data from each judge would be represented the same in the judge-by-object matrix, since the row sums are \{2, 2, 1, 1\} for each of the preference matrices. For Nishisato's method uses only a net score for each object from each judge, and has thus reduced the data so that individual comparisons are lost (unless, of course, each judge produces a perfectly transitive set of comparisons, i.e., a ranking).

The dual scaling method has some performance problems, especially when the number of objects is small. If, in a 3-object experiment, all the judges give circular triads, the procedure fails, since the judge-by-object matrix is null. Or if the row-sum scores for the objects are all equal, the method can give nearly any desired scaling through clever choice of the initial vector, even for larger experiments. Nishisato points out that problems will arise if the initial vector chosen for the iterative process is orthogonal to the solution vector, or if the initial vector is a scalar multiple of the unit vector.

### Generalized Row-Sums

In a recent technical report Chebotarev (1988) notes that any paired-comparison experiment can be thought of as an aggregation of \(m\), say, possibly incomplete sub-
experiments or 'rounds,' in which any pair of objects is compared at most once. If a comparison has been made between \( C_i \) and \( C_j \) in the \( k^{th} \) round, he denotes the result of this comparison by \( \alpha_{ij}^{(k)} \), and requires only that \( \alpha_{ij}^{(k)} = -\alpha_{ji}^{(k)} \). The simple row-sum for \( C_i \) is then

\[
u_i = \sum_{k=1}^{m} \sum_{j}^{(i;k)} \alpha_{ij}^{(k)}, \tag{2.10}
\]

where the sum \( \sum_{j}^{(i;k)} \) ranges over all objects \( C_j \) for which a comparison with \( C_i \) has been made in the \( k^{th} \) round. (For a two-point scale, it would be natural to take \( \alpha_{ij}^{(k)} \in \{-1, 1\} \), in which case \( u_i \) is the total number of comparisons in favor of \( C_i \) minus the total number against.) Chebotarev seeks a score for \( C_i \) which reduces to \( u_i \) under complete balance, and which takes the following form in the general unbalanced case:

\[
x_i = \sum_{k=1}^{m} \sum_{j}^{(i;k)} f^{(k)}_{ij}, \tag{2.11}
\]

where \( f_{ij}^{(k)} \) is a 'reward' function for \( C_i \) from its comparison with \( C_j \) in round \( k \). Chebotarev postulates that \( f_{ij}^{(k)} \) should depend not only on the direct result \( \alpha_{ij}^{(k)} \) but on the 'strengths' of the objects involved. He then shows that any function \( f \) which satisfies this desire, along with a number of other fairly natural conditions, must be of the form

\[
f_{ij}^{(k)} = \alpha_{ij}^{(k)} + \epsilon(x_j - x_i + mt\alpha_{ij}^{(k)}), \tag{2.12}
\]

where the constant \( \epsilon > 0 \) determines the extent to which the scores depend on the objects' relative strengths. When \( \epsilon = 0 \), we see at once from (2.12) and (2.11) that \( x_i \) reduces to \( u_i \); greater values of \( \epsilon \) lead to greater dependence on the objects' relative
standing. In case the $\alpha^{(k)}_{ij}$ are bounded, Chebotarev finds an upper bound on sensible values of $\epsilon$. Even so, choice of $\epsilon$ seems somewhat arbitrary, and little help is offered in guiding this choice.

Scores are calculated by solving the linear system of $t$ equations (2.11) in $t$ unknowns $\{x_i\}$. We can use (2.12) to express $x_i$ as

$$x_i = \sum_{k=1}^{m} \sum_j^{(i:k)} [(1 + \epsilon \mu t)\alpha^{(k)}_{ij} + \epsilon (x_j - x_i)]$$

$$= (1 + \epsilon \mu t)u_i + \epsilon \sum_j^{(i)} n_{ij} (x_j - x_i),$$

(2.13)

where $\sum_j^{(i)}$ denotes the sum over all objects $C_j$ which have been compared (at least once) with $C_i$. With $N_i = \sum_j^{(i)} n_{ij}$ denoting the total number of comparisons involving $C_i$, (2.13) becomes

$$(1 + \epsilon N_i)x_i - \epsilon \sum_j^{(i)} n_{ij} x_j = (1 + \epsilon \mu t)u_i.$$  

(2.14)

The set of $t$ equations (2.14) can be represented

$$(\mathbf{I} + \epsilon \mathbf{C})\mathbf{x} = \mathbf{c}\mathbf{u},$$

(2.15)

where $c = 1 + \epsilon \mu t$, and the elements of $\mathbf{C}$ are given by

$$c_{ij} = \begin{cases} N_i & i = j \\ -n_{ij} & i \neq j \end{cases}.$$  

When $\epsilon = 0$ we see at once that $c = 1$ and hence $\mathbf{x} = \mathbf{u}$, as desired.

Since $m$ appears only in the scalar constant $c$, we see from (2.15) that the relative standing of the scores does not depend on the number of rounds or sub-experiments.
This is not at all counterintuitive, since it should make no difference whether we organize the comparisons among a group of objects in a single round or in as many rounds as there are comparisons.

Chebotarev guarantees the existence and uniqueness of a solution

$$x = c(I + cC)^{-1}u,$$  \hspace{1cm} (2.16)

and lists several mathematical properties of these scores. Regarding the statistical properties of the scores, he supposes that $E(\alpha_{ij}^{(k)})$ is proportional to $X_i - X_j$, the true difference between $C_i$ and $C_j$. Under this assumption he shows that

$$x_i = u_i - \frac{c}{c} \sum^* (x_j - x_i),$$ \hspace{1cm} (2.17)

where the sum $\sum^*$ ranges over all objects $C_j$ and all rounds $k$ for which a comparison with $C_i$ has not been made. And since $x_i$ reduces to $u_i$ under complete balance, we see from this expression that $x_i$ can be viewed as the $i^{th}$ row sum from a complete experiment, less a correction for comparisons which were not observed.

Chebotarev's statistical model is reminiscent of the linear models described in the introduction (e.g., the Bradley-Terry model), under which the expected result of a comparison of two objects is a function of the difference in the overall worth of the objects, defined on a suitable scale. Hence this aspect of his work is beyond our scope. But Chebotarev's scores do indeed take into account the varied caliber of the objects; his method thus addresses some of our concerns about unbalanced data.
CHAPTER 3. PROPOSED PROCEDURE AND ITS PROPERTIES

David’s Approach

David (1987) points out the inadequacy of row-sum rankings $w = A1_t$ for unbalanced experiments, and adopts Kendall and Wei’s idea that the ‘indirect’ or ‘iterated’ wins

$$w^{(2)} = Aw = A^21_t$$

are relevant in estimating the worths of the objects. He notes that such an approach gives “more credit to a player for defeating a high-scoring than a low-scoring opponent, but this means, in effect, that a loss to the latter is punished less than a loss to the former.” Hence he decides to use the ‘direct’ losses $l = (A')1_t$ and the indirect losses $l^{(2)} = (A')l = (A')^21_t$ in conjunction with the direct and indirect wins. His proposed score is then

$$s = w^{(2)} - l^{(2)} + w - l$$

(3.1)

David goes on to examine some of the mathematical properties of this scoring system, including the following:

1. Let $s^*$ be the vector of scores when wins and losses are interchanged. Then

$$s^* = -s.$$
2. The scores sum to zero.

3. If the experiment is balanced, then \( s = t[w - \frac{1}{2}(t - 1)1_t] \).

Property 1 assures us that wins and losses are treated 'symmetrically', so that the rankings will be faithfully reversed, as desired, when wins and losses are interchanged. (Earlier we noted that this holds for Cowden's procedure. Yet this is not the case with Kendall and Wei's method, since their scores involved the row sums of \( A \) but not of \( A' \). Nor does Property 1 hold for some of the parametric models, including the Bradley-Terry model.) And Property 3 asserts that use of \( s \) is equivalent to the well-established row-sum method under complete balance.

Although this system certainly applies to experiments consisting of a single incomplete replication (i.e., where all the \( n_{ij} = 0 \) or \( 1 \)), use of the \( \{\alpha_{ij}\} \) themselves in the preference matrix \( A \) is often inappropriate in the more general case, since too much weight is then given to objects involved in (relatively) large numbers of comparisons. Indeed, when some \( n_{ij} \) are much greater than \( 1 \), the effects of the indirect wins and losses \( w^{(2)} \) and \( l^{(2)} \) swamp the effects of their direct counterparts:

<table>
<thead>
<tr>
<th>Preference matrix</th>
<th>( w )</th>
<th>( l )</th>
<th>( w^{(2)} )</th>
<th>( l^{(2)} )</th>
<th>( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( * 3 5 2 )</td>
<td>10</td>
<td>9</td>
<td>104</td>
<td>70</td>
<td>35</td>
</tr>
<tr>
<td>( 1 \ast 2 0 )</td>
<td>3</td>
<td>12</td>
<td>44</td>
<td>97</td>
<td>-62</td>
</tr>
<tr>
<td>( 5 8 \ast 4 )</td>
<td>17</td>
<td>8</td>
<td>94</td>
<td>75</td>
<td>28</td>
</tr>
<tr>
<td>( 3 1 1 \ast )</td>
<td>5</td>
<td>6</td>
<td>50</td>
<td>50</td>
<td>-1</td>
</tr>
</tbody>
</table>

Here \( C_3 \) has the most 'wins' \( w_3 = 17 \) and a dominant record over each of the other three objects; we might expect that it should receive the highest score. But note that the elements in \( w^{(2)} \) and \( l^{(2)} \) are several times bigger than those in \( w \) and \( l \). So
although $C_1$ won only half of the encounters with $C_3$ and had a losing record against $C_4$, the fact that $C_1$ had many encounters with the strong object $C_3$ boosted $w^{(2)}_1$ to such a height that $C_1$ received the highest score $s_1 = 35$.

David mentions that the more general situation can be handled by using in $A$ the proportion of times that each object was preferred to each other. If we use these $\{p_{ij}\}$ instead of the $\{\alpha_{ij}\}$, the scores for the objects from the experiment in (3.2) are $s' = [0.60, -4.20, 2.40, 1.20]$; we then see that $C_3$ is indeed top-rated, as expected. Henceforth we will use these proportions instead of the raw counts of preferences.

**Expression for David's Score $s_i$**

**General case**

Recall that we have $t$ objects and $n_{ij}$ comparisons between objects $C_i$ and $C_j$ ($n_{ij} = 0, 1, 2, \ldots$). Let $A = ((p_{ij}))$, where

$$p_{ij} = \begin{cases} \frac{\alpha_{ij}}{n_{ij}} & \text{for } n_{ij} > 0 \\ \frac{n_{ij}}{n_{ij}} & \text{for } n_{ij} = 0. \end{cases}$$

Thus $p_{ji} = 1 - p_{ij}$ for $n_{ij} > 0$, and $p_{ji} = p_{ij} = 0$ for $n_{ij} = 0$. Since $w = A1_t$ and $1 = A'1_t$, we see that $C_i$'s direct wins and losses are, respectively,

$$w_i = \sum_j (i) p_{ij} \quad \text{and} \quad \ell_i = \sum_j (i) p_{ji}, \quad (3.3)$$

where $\sum_j (i)$ denotes the sum over all objects $C_j$ that have met $C_i$. Define $m_i$ as the number of objects with which $C_i$ was compared, and $m_{ij}$ as the number of objects
with which both $C_i$ and $C_j$ were compared. Since $w_i + \ell_i = m_i$ we have

\[ w_i - \ell_i = 2w_i - m_i. \]  

(3.4)

Furthermore, since $w^{(2)} = \mathbf{A}w$ and $l^{(2)} = \mathbf{A}'l$, we see that $C_i$'s indirect wins and losses are, respectively,

\[
\begin{align*}
w_i^{(2)} &= \sum_j (i) p_{ij}w_j \\
\ell_i^{(2)} &= \sum_j (i) p_{ij}\ell_j \\
&= \sum_j (1 - p_{ij})(m_j - w_j) \\
&= \sum_j (1 - p_{ij})m_j - \sum_j (1 - p_{ij})w_j \\
&= \sum_j (1 - p_{ij})m_j - \sum_j w_j + w_i^{(2)}.
\end{align*}
\]

So

\[ w_i^{(2)} - \ell_i^{(2)} = \sum_j (i) \left[w_j - (1 - p_{ij})m_j\right]. \]  

(3.5)

Now we can use (3.4) and (3.5) in (3.1) to define $C_i$'s score:

\[
\begin{align*}
s_i &= w_i - \ell_i + w_i^{(2)} - \ell_i^{(2)} \\
&= 2w_i - m_i + \sum_j (i) \left[w_j - (1 - p_{ij})m_j\right] \\
&= \sum_j (i) \left[2p_{ij} + w_j - (1 - p_{ij})m_j\right] - m_i.
\end{align*}
\]  

(3.6)
Notice that, whenever \( C_i \) and \( C_j \) have met, 

\[
w_j = (1 - p_{ij}) + \sum_{k \neq i}^{(j)} p_{jk},
\]

where \( \sum_{k \neq i}^{(j)} \) denotes the sum over all objects \( C_k \) (excluding \( C_i \)) which have met \( C_j \). Putting this back into (3.6) and collecting the \( p_{ij} \) terms inside the \( \sum_{j}^{(i)} \) summation, we have

\[
s_i = \sum_{j}^{(i)} \left[ (m_j + 1)p_{ij} + 1 + \sum_{k \neq i}^{(j)} p_{jk} - m_j \right] - m_i
\]

\[
= \sum_{j}^{(i)} \left[ (m_j + 1)p_{ij} - m_j + \sum_{k \neq i}^{(j)} p_{jk} \right],
\]

(3.7)

since \( \sum_{j}^{(i)} 1 = m_i \). Notice that both \( p_{gh} \) and \( p_{hg} \) will appear in this expression whenever \( C_i \) has met both \( C_g \) and \( C_h \). Since \( p_{gh} + p_{hg} = 1 \), we would like to consolidate these 'complementary' proportions whenever possible. To this end, write

\[
\sum_{j}^{(i)} \sum_{k \neq i}^{(j)} P_{jk} = \sum_{j}^{(i)} \sum_{k \neq i}^{(i,j)} P_{jk} + \sum_{j}^{(i)} \sum_{k \neq i}^{(-i,j)} P_{jk},
\]

(3.8)

where \( \sum_{k \neq i}^{(i,j)} \) denotes the sum over the \( m_{ij} \) objects \( C_k \) which have met both \( C_i \) and \( C_j \), and \( \sum_{k \neq i}^{(-i,j)} \) denotes the sum over the \( m_j - m_{ij} - 1 \) objects \( C_k \) (excluding \( C_i \)) which have met \( C_j \) but not \( C_i \). Careful consideration reveals that the first double-sum of the RHS of (3.8) is equal to \( \frac{1}{2} \sum_{j}^{(i)} m_{ij} \), so that (3.7) becomes

\[
s_i = \sum_{j}^{(i)} \left[ (m_j + 1)p_{ij} - m_j + \frac{1}{2}m_{ij} + \sum_{k \neq i}^{(-i,j)} p_{jk} \right]
\]

\[
= \sum_{j}^{(i)} \left[ (m_j + 1)(p_{ij} - \frac{1}{2}) + \sum_{k \neq i}^{(-i,j)} (p_{jk} - \frac{1}{2}) \right].
\]

(3.9)
In this expression, no \( p_{gh} \) appears twice, and no \( p_{gh} \) appears with its complementary \( p_{hg} \). The only stochastic elements in the above are the independent proportions representing comparisons which involve \( C_i \) directly, and comparisons which involve two other objects, of which exactly one has met \( C_i \).

**Special case: no empty cells**

Suppose that there are no 'empty cells', i.e., all \( n_{ij} > 0 \). Then \( m_j = t - 1 \forall j \), and the inner sum \( \sum_{k \neq i} (-i,j) \) of (3.9) has no terms, since all objects have met \( C_i \). The score then reduces to

\[
s_i = t \sum_{j \neq i} (p_{ij} - \frac{1}{2}),
\]

which is, as David claims in Property 3, equivalent to the corresponding row-sum \( w_i \) of the preference matrix. Indeed, using (3.3) we see that

\[
s_i = tw_i - \left( \frac{t}{2} \right).
\]

Note that this simplification does not require that the design be balanced, but only that each pair of objects be compared at least once. If the design is indeed completely balanced, the score becomes

\[
s_i = \frac{t}{n} a_i - \left( \frac{t}{2} \right),
\]

where \( a_i \) is the total number of preferences in favor of \( C_i \).

**Special case: group divisible designs**

Between the two extremes presented in the previous two subsections lies a sequence of partially balanced designs. In one common and important class of such
designs, the objects are separated into \( m \) distinct groups of \( a = t/m \) objects each, where each object is compared \( n_1 \) times with each member of its own group and \( n_2 \) times with each member of every other group. If \( n_1 \) and \( n_2 \) are both greater than zero, we find ourselves in the case with no empty cells again, discussed above; if \( n_2 = 0 \), the groups are disconnected. So without loss of generality, we can take \( n_1 = 0 \) and \( n_2 = 1 \). In other words,

\[
n_{ij} = \begin{cases} 
0 & \text{if } C_i \text{ and } C_j \text{ are in the same group} \\
1 & \text{if } C_i \text{ and } C_j \text{ are in different groups,}
\end{cases}
\]

so that

\[
m_j = t - a
\]

\[
m_{ij} = \begin{cases} 
 t - a & \text{if } C_i \text{ and } C_j \text{ are in the same group} \\
t - 2a & \text{if } C_i \text{ and } C_j \text{ are in different groups.}
\end{cases}
\]

From (3.9) the score for \( C_i \) is now

\[
s_i = (t - a + 1) \sum_j (p_{ij} - \frac{1}{2}) + \sum_j \sum_{k \neq i} (-i,j) (p_{jk} - \frac{1}{2})
\]

\[
= (t - a + 2) \sum_j (p_{ij} - \frac{1}{2}) + \sum_j \sum_k (-i,j) (p_{jk} - \frac{1}{2})
\]

\[
= (t - a + 2) \sum_j (p_{ij} - \frac{1}{2}) - \sum_{k \in G_i} \sum_j (p_{kj} - \frac{1}{2})
\]

\[
= (t - a + 2) a_i^* - \sum_{k \in G_i} a_k^*
\]
say, where $G_i$ is the group of objects containing $C_i$, and $a_k^* G_i$ is then the $k^{th}$ centered row sum of the preference matrix. Note from (3.15) that $s_i$ is simply a multiple of $C_i$'s own 'strength' $a_i^*$ minus a term for the 'strength' of $C_i$'s group. Thus $C_i$ would be penalized if its group was stronger than the other groups; this might seem undesirable at first. But because comparisons are not made within groups, $C_i$ would have been spared direct competition with its (presumed) strong intra-group rivals. The objects which it meets directly must be that much weaker; hence the reduced score.

**Distribution Theory**

**Expectations, variances, and covariances of scores**

**General case** Thus far we have not required that the judges give their preferences on a two-point scale: a clear preference for one object or the other. David's method (and some of the procedures from Chapter 2) gives sensible scores if the data are recorded on a multiple-point or even continuous scale. The following distributional results, however, do require use of this two-point scale.

From (3.9) we can easily derive the expectation and variance of a score:

$$E(s_i) = \sum_j^i \left[ (m_j + 1)(\pi_{ij} - \frac{1}{2}) + \sum_{k \neq i}^j (\pi_{jk} - \frac{1}{2}) \right] \quad (3.16)$$

$$V(s_i) = \sum_j^i \left[ (m_j + 1)^2 \frac{\pi_{ij} \pi_{ji}}{n_{ij}} + \sum_{k \neq i}^j \frac{\pi_{jk} \pi_{kj}}{n_{jk}} \right] \quad (3.17)$$

The covariance of two scores $s_i$ and $s_j$ ($j \neq i$), say, is a bit more involved; derivation of the following expressions can be found in Appendix A:

$$\text{cov}(s_i, s_j) = -(m_i + 1)(m_j + 1)\frac{\pi_{ij} \pi_{ji}}{n_{ij}}$$
for $n_{ij} > 0$; for $n_{ij} = 0$ we have

$$\text{cov}(s_i, s_j) = -\sum_k (m_k + 1) \left( \frac{\pi_{ik} \pi_{ki}}{n_{ik}} + \frac{\pi_{jk} \pi_{kj}}{n_{jk}} \right) + \sum_k \sum_{l \neq i, j} \frac{\pi_{kl} \pi_{lk}}{n_{kl}} - \sum_{k \neq j} \sum_{l \neq i} \frac{\pi_{kl} \pi_{lk}}{n_{kl}}. \quad (3.19)$$

**Special case: no empty cells**  
From (3.10) the expectations, variances, and covariances are quite straightforward:

$$E(s_i) = \sum_{j \neq i} (\pi_{ij} - \frac{1}{2})$$

$$V(s_i) = \sum_{j \neq i} \frac{\pi_{ij} \pi_{ji}}{n_{ij}}. \quad (3.20)$$

$$\text{cov}(s_i, s_j) = \frac{-2 \pi_{ij} \pi_{ji}}{n_{ij}}. \quad (3.21)$$

For complete balance simply substitute $n$ for $n_{ij}$ in (3.20) and (3.21).

**Special case: null distribution**  
Under the 'hypothesis of randomness' $H_0 : \pi_{ij} = \frac{1}{2}$ for all $i \neq j$,  \quad (3.22)

either the objects are indistinguishable or their differences are irrelevant, so that no object is preferred to any other object. From (3.16) and (3.17), we see that the
expectations and variances then simplify to

\[ E(s_i) = 0 \]

\[ V(s_i) = \frac{1}{4} \sum_j^i \left[ \frac{(m_j + 1)^2}{n_{ij}} + \sum_{k \neq i}^j \frac{(-i,j)}{n_{jk}} \right], \]

respectively, though the covariances simplify little. If we furthermore insist that there

be no empty cells, however, we see from (3.20) and (3.21) that

\[ V(s_i) = \frac{i^2}{4} \sum_{j \neq i} \frac{1}{n_{ij}} \quad (3.23) \]

\[ \text{cov}(s_i, s_j) = -\frac{i^2}{4n_{ij}} \quad (3.24) \]

**Special case: group divisible designs, one outlier model**

For the
general group divisible design with \( m \) groups of \( a \) objects, we see from (3.13) that

\[ E(s_i) = (t - a + 1) \sum_j^i (\pi_{ij} - \frac{1}{2}) + \sum_j^i \sum_{k \neq i} (\pi_{jk} - \frac{1}{2}), \quad (3.25) \]

which simplifies little. But consider now the model

\[ \pi_{ai} = \pi > \frac{1}{2}, \text{ for } i \neq a \]

\[ \pi_{ij} = \frac{1}{2}, \text{ for } i, j \neq a, \text{ and } i \neq j. \quad (3.26) \]

In other words, the 'outlier' \( C_a \) is preferred to each other object with probability \( \pi \),

and the remaining objects are considered equals. For \( C_a \) we get

\[ E(s_a) = (t - a)(t - a + 1)(\pi - \frac{1}{2}). \quad (3.27) \]

For any other object \( C_i \),

\[ E(s_i) = \begin{cases} 
-(t - a)(\pi - \frac{1}{2}) & \text{, } i \text{ in } C_a's \text{ group}, \\
-(t - 2a + 2)(\pi - \frac{1}{2}) & \text{, } i \text{ not in } C_a's \text{ group}. 
\end{cases} \quad (3.28) \]
When the group size \( a = t/m \) is 1, we have a balanced design with all objects meeting all others; no object would be in \( C_a \)'s group, so that all the (lesser) objects would have the same expected score. For \( a = 2 \) we see from (3.28) that \( C_i \)'s score is the same, regardless of whether it is in \( C_a \)'s group. But for any larger group size, it will benefit an object \( C_i \) not to be in \( C_a \)'s group; in other words, it is to \( C_i \)'s advantage to be compared with the outlier.

**Extending designs by one comparison**

Consider again the one-outlier model (3.26), and suppose that we have data from an experiment \( T \) (though not necessarily a group divisible design, as above). If we extend \( T \) to an experiment \( T' \) by adding one or more comparisons between some object \( C_j \), say, and the outlier \( C_a \), how will the expectations of the various scores be affected?

If \( C_j \) and \( C_a \) had already met at least once in \( T \), the scores' expectations will not change. So assume that they had not met. The expected score for the outlier under \( T \) would then be

\[
E(s_a) = \sum_j \left[ (m_j + 1)(\pi_{aj} - \frac{1}{2}) + \sum_{k \neq a} (\pi_{jk} - \frac{1}{2}) \right]
\]

\[
= \sum_j (m_j + 1)(\pi - \frac{1}{2}),
\]

(3.29)

since \( \pi_{aj} = \pi \) for \( j \neq a \), and since \( \pi_{jk} = \frac{1}{2} \) for \( j, k \neq a \). Clearly none of the \( m_j \) will change for \( j \neq a, b \). So the outlier's expected score under \( T' \) will differ from that under \( T \) only through the addition of a term for \( j = b \) in (3.29), namely

\[
E(s'_a) - E(s_a) = (m'_b + 1)(\pi - \frac{1}{2})
\]

\[
= (m_b + 2)(\pi - \frac{1}{2}),
\]
since \( m'_b = m_b + 1 \). As \( \pi > \frac{1}{2} \), we see that \( C_a \)'s score is expected to increase.

To \( C_b \)'s expected score we must likewise add a term for \( j = a \):

\[
(m'_a + 1)(\pi_{ba} - \frac{1}{2}) + \sum_{k \neq b} (\pi_{ak} - \frac{1}{2})
\]

\[
= -(m'_a + 1)(\pi - \frac{1}{2}) + (m'_a - m'_{ab} - 1)(\pi - \frac{1}{2})
\]

\[
= -(m_{ab} + 2)(\pi - \frac{1}{2}),
\]

(3.30)

using \( m'_{ab} = m_{ab} \). (Note that, although \( m_a \) has increased by one, this difference does not figure into the difference between \( E(s_b') \) and \( E(s_b') \).) The supplemental comparisons will affect \( C_b \)'s score in less obvious ways as well. Consider the summand for any \( j \) other than \( a \):

\[
(m_j + 1)(\pi_{bj} - \frac{1}{2}) + \sum_{k \neq b} (\pi_{jk} - \frac{1}{2}).
\]

(3.31)

(Bear in mind that \( C_j \) must have met \( C_b \) for this term to appear in the expression for \( E(s_b') \).) Since \( C_j \) was not (directly) involved in the extra comparison(s), \( m_j \) will not change; the first term in (3.31) will hence be the same under \( T' \) as under \( T \). If \( C_j \) has not met \( C_a \), then the second term will remain unchanged as well. But if \( C_j \) has indeed met the outlier, this inner sum will no longer have a term for \( k = a \). This difference of \(- (\pi_{ja} - \frac{1}{2}) = \pi - \frac{1}{2} \) will arise for each of the \( m'_{ab} = m_{ab} \) objects \( C_j \) which have met both \( C_a \) and \( C_b \). Taking these differences with (3.30) we get

\[
E(s'_b) - E(s_b) = -(m_{ab} + 2)(\pi - \frac{1}{2}) + m_{ab}(\pi - \frac{1}{2})
\]

\[
= -2(\pi - \frac{1}{2}).
\]

Thus \( C_b \)'s score is expected to decrease when we include its comparisons with the outlier.
Consider finally the expected score of an object $C_i$, say, $i \neq a, b$:

$$E(s_i) = \sum_{j}^{(i)} \left[ (m_j + 1)(\pi_{ij} - \frac{1}{2}) + \sum_{k \neq i}^{(-i,j)} (\pi_{jk} - \frac{1}{2}) \right].$$

Terms for $j \neq a, b$ will not differ in $T$ and $T'$. If $C_i$ has met $C_a$, the corresponding $j = a$ term will have the following two changes:

- the inner sum must add the $k = b$ term $\pi_{ab} - \frac{1}{2} = \pi - \frac{1}{2}$, provided $C_i$ and $C_b$ have not met, and

- $m_a' = m_a + 1$, giving a difference of $\pi_{ia} - \frac{1}{2} = -(\pi - \frac{1}{2})$.

(Note that if $n_{ia} > 0$ and $n_{ib} > 0$ we have only the second of these two changes; if $n_{ia} > 0$ and $n_{ib} = 0$ the two changes will cancel, giving a net change of zero.) If $C_i$ has not met $C_a$, the inner sum of the corresponding $j = b$ term (if there is such a term, i.e., if $C_i$ and $C_b$ have met) will add the $k = a$ term $\pi_{ba} - \frac{1}{2} = -(\pi - \frac{1}{2})$. Thus the change in $C_i$'s expected score is the same for $n_{ia} = 0$ as for $n_{ia} > 0$:

$$E(s_i') - E(s_i) = \begin{cases} - (\pi - \frac{1}{2}), & \text{if } C_i \text{ has met } C_b \\ 0, & \text{otherwise.} \end{cases}$$

In other words, any object which had met $C_b$ in $T$ is punished in $T'$ only through the additional indirect comparison(s) with the outlier.

Asymptotics

It seems that little can be said about the asymptotic distribution of the scores if the $\{n_{ij}\}$ are allowed to grow in an uncontrolled or unrelated manner. Consider the
following restrictions on their growth. Suppose

\[
\frac{n_{ij}}{\sum_k n_{ik}} \to c_{ij} \quad \text{as the } \{n_{ij}\} \to \infty, \quad (3.32)
\]

where \(c_{ij}\) is some constant in the interval \((0, 1)\). Writing \(N_i = \sum_k n_{ik}\) we have

\[
\frac{1/n_{ij}}{\sum_k 1/n_{ik}} \to \frac{1/N_i c_{ij}}{\sum_k 1/N_i c_{ik}} = \frac{1/c_{ij}}{\sum_k 1/c_{ik}} = u_{ij}^2, \quad \text{say.} \quad (3.33)
\]

Consider at first the case with no empty cells, i.e. \(n_{ij} > 0\) for all pairs \((i, j)\). As noted in (3.10), the scores are then equivalent to row-sums of the appropriate proportions \(\{p_{ij}\}\):

\[
s_i = t \sum_{j \neq i} (p_{ij} - \frac{1}{2}).
\]

Under the hypothesis of randomness (3.22) we get

\[
q_{ij} = \frac{p_{ij} - \frac{1}{2}}{\frac{1}{2} \sqrt{1/n_{ij}}} \to N(0, 1) \quad \text{as } n_{ij} \to \infty,
\]

\[
E_o(s_i) = 0, \quad \text{and}
\]

\[
V_o(s_i) = \frac{i^2}{4} \sum_{j \neq i} \frac{1}{n_{ij}}.
\]

Since the proportions comprising \(s_i\) are independent, we see that the asymptotic
distribution of a standardized score

\[ d_i = \frac{t \sum_{j \neq i} (p_{ij} - \frac{1}{2})}{\sqrt{\frac{t^2}{4} \sum_{k \neq i} \frac{1}{n_{ik}}}} \]

\[ = \sum_{j \neq i} \frac{p_{ij} - \frac{1}{2}}{\frac{1}{2} \sqrt{\sum_{k \neq i} \frac{1}{n_{ik}}}} \]

\[ = \sum_{j \neq i} q_{ij} \sqrt{\frac{1/n_{ij}}{\sum_{k \neq i} 1/n_{ik}}} \]

is that of \( \sum_{j \neq i} Z_{ij} u_{ij} \), where the \( \{Z_{ij}\} \) are independent N(0,1) variates, and the \( \{u_{ij}\} \) are defined by equation (3.33). So since

\[ \sum_{j \neq i} u_{ij}^2 = \sum_{j \neq i} \frac{1/c_{ij}}{\sum_{k \neq i} 1/c_{ik}} = 1, \]

we see that \( d_i \) has asymptotic N(0,1) distribution.

But what of the joint distribution of the \( \{d_i\} \)? Following the previous argument, \( d_i \) will have the same limiting distribution as

\[ d_i' = \sum_{j \neq i} q_{ij} u_{ij}. \]

An arbitrary linear function \( L = \sum_{i=1}^{t} h_i d_i \) of the scores will then have the same limiting distribution as

\[ L' = \sum_{i=1}^{t} h_i d_i' \]

\[ = \sum_{i=1}^{t} h_i \sum_{j \neq i} q_{ij} u_{ij} \]
But \( n_{ji} = n_{ij} \) and \( (p_{ji} - \frac{1}{2}) = -(p_{ij} - \frac{1}{2}) \), so that \( q_{ji} = -q_{ij} \). The second double sum in (3.34) then becomes, upon interchanging the order and indices of summation,

\[
- \sum_{i=1}^{t} \sum_{j>i}^{t} h_{ij} q_{ij} u_{ij}.
\]

Now (3.34) gives us

\[
L' = \sum_{i=1}^{t} \sum_{j>i}^{t} (h_{ij} u_{ij} - h_{ji} u_{ji}) q_{ij}.
\]

Since the \( \{p_{ij}\} \) (and hence the \( \{q_{ij}\} \)) are independent, we see that \( L' \) (and hence \( L \)) has limiting normal distribution, as the \( \{n_{ij}\} \to \infty \) under the restrictions described above in (3.32). Since any linear function of the standardized scores has a (limiting) normal distribution, we know that the scores themselves have a (limiting) multivariate normal distribution.

In the general (unbalanced) case, the scores are no longer equivalent to the row-sums of the preference matrix. Each \( s_i \) is, however, nothing but a linear function of independent binomial proportions, as seen from (3.9). We can still express \( s_i \) as a linear function of the standardized proportions, then follow the argument above.

**Tests**

**Hypothesis of randomness**

For incomplete experiments we will not be able to test the full 'hypothesis of randomness' (3.22), since we have no information on any \( \pi_{ij} \) for which \( C_i \) and \( C_j \).
have not met. We could, however, test

\[ H_0^* : \pi_{ij} = \frac{1}{2} \quad \forall (i,j), \, n_{ij} > 0 \]  \hspace{1cm} (3.35)

with a Pearson \( \chi^2 \)-type statistic. From the \( n_{ij} \) comparisons between \( C_i \) and \( C_j \), we observed \( \alpha_{ij} \) preferences in favor of \( C_i \) and had expected \( n_{ij}/2 \). Hence we can use

\[
\chi^2 = \sum_{i=1}^{t} \sum_{j<i}^{(i)} \frac{(\alpha_{ij} - n_{ij}/2)^2}{n_{ij}/2}
\]

\[
= 2 \sum_{i=1}^{t} \sum_{j<i}^{(i)} n_{ij}(p_{ij} - \frac{1}{2})^2
\]

\[
= 4 \sum_{i=1}^{t} \sum_{j<i}^{(i)} n_{ij}(p_{ij} - \frac{1}{2})^2.
\]

to test this 'incomplete' hypothesis of randomness (3.35). The degrees of freedom associated with this statistic will be the number of pairs that have been compared (at least once). Note that, under complete balance, this is equivalent to Bradley's (1955) multi-binomial test.

**Equality of treatments**

**Test statistic; null distribution** It is also of interest to test whether the treatments are of equal strength or value. In David (1963,1988) it is shown that, under complete balance,

\[
D_n = \frac{4}{nt} \sum_{i=1}^{t} [a_i - \frac{1}{2}n(t - 1)]^2
\]

has limiting \( \chi^2_{t-1} \) distribution as \( n \to \infty \), where \( a_i = \sum \alpha_{ij} \) is the \( i^{th} \) row-sum of the preference matrix \( A \). We will now derive a similar quadratic form of the scores
\{s_i\} to handle unbalanced data. This quadratic form will also have limiting \(\chi^2_{t-1}\) distribution, and will reduce to \(D_n\) under complete balance.

First note that the covariance matrix \(\Sigma\) of the scores \(\{s_i\}\) is singular, since \(s'1 = 0\) implies that

\[
\Sigma 1 = E[(s - E(s))(s - E(s))']1 = E[(s - E(s))s']1 = 0.
\]

Clearly there is dependence among the rows (and columns) of \(\Sigma\). In fact, the rank of \(\Sigma\) is \(t - 1\) if and only if the objects form a connected set. (See Benard and van Elteren (1953) for a proof of the more general result that a similar type of covariance matrix has rank \(t - s\) if and only if there are \(s\) disconnected sets of objects.) We therefore focus on the covariance matrix \(\bar{\Sigma}\) of any set of \(t - 1\) scores; for convenience we take the first \(t - 1\) scores, and denote by \(\bar{s}\) the vector of these \(t - 1\) scores.

Consider that the linear transformation \(\bar{b} = \bar{\Sigma}^{-\frac{1}{2}}\bar{s}\) has covariance matrix \(\Gamma_{t-1}\). (Appendix B details a method for finding \(\bar{\Sigma}^{-\frac{1}{2}}\).) Then, in light of the joint asymptotic multivariate normality of \(s\), we can say that \(\bar{b} \to N(0, \Gamma_{t-1})\) and hence that the test statistic

\[ Q = \bar{b}'\bar{b} = \bar{s}'\bar{\Sigma}^{-1}\bar{s} \tag{3.37} \]

has an asymptotic \(\chi^2_{t-1}\) distribution.

**Non-centrality parameter for \(Q\)** We are also concerned with the non-null distribution of \(Q\). Suppose that \(s\) has some non-zero mean vector \(\mu\). (Denote by
\( \bar{\mu} \) the mean of \( \bar{s} \). From, for example, Corollary 2.1 of Searle (1971), the quadratic form \( \bar{s}' P \bar{s} \) has an asymptotic \( \chi^2 \left[ \text{tr}(P \bar{\Sigma}), \bar{\mu}' P \bar{\mu} \right] \) distribution if and only if

1. \( \bar{\Sigma} P \bar{\Sigma} = \bar{\Sigma} \)

2. \( \bar{\mu} P \bar{\Sigma} = \bar{\mu} \bar{\Sigma} \)

3. \( \bar{\mu} P \bar{\Sigma} \bar{\mu} = \bar{\mu} P \bar{\mu} \).

Since \( P = \bar{\Sigma}^{-1} \) for our case, the three conditions follow immediately, and our non-centrality parameter for \( Q \) is \( \lambda = \bar{\mu}' \bar{\Sigma}^{-1} \bar{\mu} \).

**Special case: complete balance**

Putting \( n_{ij} = n \forall i \neq j \) in (3.2.3) and (3.2.4) gives us

\[
\begin{align*}
\text{var}(s_i) &= \frac{t^2(t-1)}{4n} \\
\text{cov}(s_i, s_j) &= -\frac{t^2}{4n}, \quad i \neq j
\end{align*}
\]

for complete balance under the hypothesis of randomness. Then

\[
\bar{\Sigma} = \frac{t^2}{4n} \begin{bmatrix} t-1 & -1 & \cdots & -1 \\ -1 & t-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & t-1 \end{bmatrix}
= \frac{t^2}{4n} \left( I_t - J_{t-1} \right).
\] (3.38)

Since

\[
(aI_n + bJ_{n})^{-1} = \frac{1}{a} \left( I_n - \frac{b}{a + nb} J_{n} \right)
\] (3.39)
(see, e.g., Searle (1982), p. 132), we can easily invert (3.38):
\[
\hat{\Sigma}^{-1} = \frac{4n}{t^2} \cdot \frac{1}{t} \left( I_{t-1} + \frac{1}{t-(t-1)} J_{t-1} \right) \\
= \frac{4n}{t^3} \left( I_{t-1} + J_{t-1} \right) .
\]

Using this in (3.37) we have
\[
Q_{Bal} = \hat{s}' \hat{\Sigma}^{-1} \hat{s} \\
= \frac{4n}{t^3} \left( \hat{s}' I_{t-1} \hat{s} + \hat{s}' J_{t-1} \hat{s} \right) 
\]

Noting that \( s' 1_t = 0 \) implies that \( \hat{s}' I_{t-1} = -s_t \), we see that \( J_{t-1} \hat{s} = -s_t 1_{t-1} \) and hence \( \hat{s}' J_{t-1} \hat{s} = s_t^2 \). And since \( \hat{s}' I_{t-1} \hat{s} = \sum_{i=1}^{t-1} s_i^2 \), (3.42) gives us
\[
Q_{Bal} = \frac{4n}{t^3} \sum_{i=1}^{t} s_i^2 .
\]

Using (3.10), note that under complete balance
\[
s_i = t \sum_{j \neq i} (p_{ij} - \frac{1}{2}) \\
= \frac{t}{n} \sum_{j \neq i} (\alpha_{ij} - \frac{n}{2}) \\
= \frac{t}{n} \left( a_i - \frac{n(t-1)}{2} \right) ,
\]
where \( a_i = \sum_{j \neq i} \alpha_{ij} \). Putting this back into (3.43) we get
\[
Q_{Bal} = \frac{4}{nt} \sum_{i=1}^{t} \left[ a_i - \frac{1}{2} n(t-1) \right] ,
\]
which agrees with David's statistic \( D_n \) from (3.36) for this special case of complete balance.
In a derivation similar to (3.43), the non-centrality parameter under this special case is

$$\lambda_{Bal} = \mu'\tilde{\Sigma}^{-1}\mu$$

$$= \frac{4n}{t^3} \sum_{i=1}^{t} \mu_i^2.$$  \hspace{1cm} (3.45)

As an accessible special case, consider the one-outlier model (3.26). Taking $C_t$ to be the outlier, we see from (3.27) and (3.28) that

$$\mu_t = t(t-1)(\pi - \frac{1}{2})$$

$$\mu_i = -t(\pi - \frac{1}{2}), \hspace{1cm} i \neq t.$$  

Putting these back into (3.46) we get

$$\lambda_{Bal} = 4n(t-1)(\pi - \frac{1}{2})^2.$$  \hspace{1cm} (3.47)

**Special case: group divisible designs**  
Suppose we have a group divisible design with $m$ groups of size $a = t/m$. Denote by $s_{gi}^t$ the $i^{th}$ score within the $g^{th}$ group, and let $S_g = \sum_{i=1}^{a} s_{gi}^t$ be the sum of scores for the $g^{th}$ group. Adopting some obvious notation from the analysis of variance, let

$$\text{SST} = \sum_{g=1}^{m} \sum_{i=1}^{a} s_{gi}^2$$

$$\text{SSB} = \frac{1}{a} \sum_{g=1}^{m} S_g^2$$

$$\text{SSW} = \sum_{g=1}^{m} \sum_{i=1}^{a} (s_{gi}^t - \frac{1}{a} S_g)^2$$

be the Total sum of squared scores, the sum of squared scores Between groups, and the sum of squared scores Within groups, respectively. Appendix C details the derivation
of the following expressions for $Q$

$$Q_{GD} = k_0\text{SST} + k_1\text{SSB} = k_0\text{SSW} + k_2\text{SSB},$$

where

$$k_0 = \frac{4}{(t-a)(t-a+2)^2},$$

$$k_1 = \frac{4[t(t-a)-(a-2)]}{t(t-a)(t-a+2)(t-2a+2)^2},$$

$$k_2 = \frac{4}{t(t-2a+2)^2}.$$  

Similarly, we can show that the non-centrality parameter is then

$$\lambda_{GD} = k_0 \sum_{g=1}^{m} \sum_{i=1}^{a} \mu_{gi}^2 + k_1 \sum_{g=1}^{m} M_g^2 = k_0 \sum_{g=1}^{m} \sum_{i=1}^{a} (\mu_{gi} - \frac{1}{a} M_g)^2 + k_2 \frac{1}{a} \sum_{g=1}^{m} M_g^2,$$

where $\mu_{gi} = E(s_{gi})$ and $M_g = \sum_{i=1}^{a} \mu_{gi}$.

We can apply (3.50) to the one-outlier situation (3.26), under which we saw from (3.27) and (3.28) that

$$\mu_{gi} = \begin{cases} 
(t-a+1)(t-a)(\pi - \frac{1}{2}) & g = m, i = a \\
-(t-a)(\pi - \frac{1}{2}) & g = m, i \neq a \\
-(t-2a+2)(\pi - \frac{1}{2}) & g \neq m, 
\end{cases}$$

where, for convenience, we have taken the outlier to be the last object in the last group. From (3.51) we see that the expected group scores are

$$M_g = \begin{cases} 
-a(t-2a+2)(\pi - \frac{1}{2}) & g \neq m \\
(t-a)(t-2a+2)(\pi - \frac{1}{2}) & g = m, 
\end{cases}$$
so that the sum of squared mean scores between groups is then
\[ \frac{1}{a} \sum_{g=1}^{m} M_g^2 = \frac{t}{a}(t - a)(t - 2a + 2)(\pi - \frac{1}{2})^2. \]  

(3.53)

Since the scores are constant for every group but the outlier’s, the sum of mean scores within groups is just
\[ \sum_{g=1}^{m} \sum_{i=1}^{a} (\mu_{gi} - \frac{1}{a} M_g)^2 = \sum_{i=1}^{a} (\mu_{mi} - \frac{1}{a} M_m)^2 = \frac{a - 1}{a} (t - a + 2)^2(t - a)(\pi - \frac{1}{2})^2, \]

after some algebra. Putting this and (3.53), along with the constants from (3.49), back into (3.50) gives
\[ \lambda_{GD} = \frac{4(a - 1)}{a} (t - a)(\pi - \frac{1}{2})^2 + \frac{4}{a} (t - a)(\pi - \frac{1}{2})^2 \]
\[ = 4(t - a)(\pi - \frac{1}{2})^2. \]  

(3.54)

**Top score**

Let \( E_i \) be the event that \( s_i > M \), for some \( M > 0 \). Then
\[ \Pr\{s_{(t)} > M\} = \Pr\left\{ \bigcup_{i=1}^{t} E_i \right\} \leq \sum_{i=1}^{t} \Pr\{E_i\}, \]

(3.55)

where \( s_{(t)} \) is the top score. If the \( \{n_{ij}\} \) are large enough for the normal approximation of the \( \{s_i\} \) to be reasonable, we see that, under the hypothesis of randomness,\n\[ \Pr\{E_i\} = \Pr\{s_i > M\} = \Pr\left\{ \frac{s_i}{\sigma_i} > \frac{M}{\sigma_i} \right\} = 1 - \Phi\left( \frac{M}{\sigma_i} \right), \]

where \( \Phi \) is the cdf of the standard normal distribution, and \( \sigma_i^2 \) is the variance of \( s_i \). We can now solve (iteratively) for \( M \) so that the RHS of (3.55) is equal to any
prespecified $\alpha$. If $s_{(t)}$ is greater than this $M$, conclude that the corresponding object is better than the average. Alternatively, we could use

$$
\sum_{i=1}^{t} \Pr \left\{ \frac{s_i}{\sigma_i} > \frac{s_{(t)}}{\sigma_i} \right\} = t - \sum_{i=1}^{t} \Phi \left( \frac{s_{(t)}}{\sigma_i} \right),
$$

as an upper bound for the $p$-value of this test.

If the $\{\sigma_i\}$ are not much different, another approach is plausible. Let $s_{i^*}$ be the score with the largest variance, $\sigma_{i^*}^2$, so that, under the hypothesis of randomness,

$$
\Pr\{E_{i^*}\} = \max_i \Pr\{E_i\}, \quad \text{and hence}
$$

$$
\Pr\{s_{(t)} > M\} \leq t \Pr\{E_{i^*}\}. \quad (3.56)
$$

Having set a significance level $\alpha$, we seek the $M$ such that

$$
\alpha = t \Pr\{E_{i^*}\}
$$

$$
= t \Pr\{s_{i^*} > M\}
$$

$$
= t \Pr \left\{ \frac{s_{i^*}}{\sigma_{i^*}} > \frac{M}{\sigma_{i^*}} \right\}.
$$

We can now solve for

$$
M = \sigma_{i^*} \Phi^{-1} \left( 1 - \frac{\alpha}{t} \right),
$$

and proceed as before. Note that if the $\{\sigma_i\}$ are much different, the inequality in (3.56) is quite harsh. Yet the suggestion that these variances be of comparable size is not entirely unreasonable: even for the highly unbalanced data in section 3, the variances of the four scores were (.477, .444, .453, .462).
Difference of two scores

Consider the difference between two prespecified objects \( C_i \) and \( C_j \), say. Under our restrictions (3.32) on the \( \{n_{ij}\} \), the statistic

\[
\frac{s_i - s_j}{\sqrt{V(s_i - s_j)}}
\]

has an asymptotic normal distribution. Having set \( \alpha \), we seek a \( M \) such that

\[
\alpha = \Pr \{ |s_i - s_j| > M \}
\]

\[
= 2 \Pr \left\{ \frac{s_i - s_j}{\sqrt{V(s_i - s_j)}} > \frac{M}{\sqrt{V(s_i - s_j)}} \right\}.
\]

This leads us to

\[
M = \sqrt{V(s_i - s_j)} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right)
\]

\[
= \sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),
\]

where \( \sigma_{ij} \) is the covariance between \( s_i \) and \( s_j \). So we conclude that \( C_i \) and \( C_j \) are of differing strength if \( |s_i - s_j| \) is greater than \( M \) from (3.57).

Multiple Comparisons

Scheffé's procedure

Consider the standardized scores \( d_i = s_i/\sigma_i \) for \( i = 1, \ldots, t \). We are interested in contrasts of the form

\[
\hat{\theta}_{ij} = d_i - d_j
\]
\[ E(\hat{\theta}_{ij}) = 0 \quad \text{and} \]
\[ V(\hat{\theta}_{ij}) = V(d_i) + V(d_j) - 2\text{cov}(d_i, d_j) \]
\[ = 1 + 1 - 2 \frac{\sigma_{ij}}{\sigma_i \sigma_j} \]
\[ = 2 \left( 1 - \frac{\sigma_{ij}}{\sigma_i \sigma_j} \right), \]
where \( \sigma_{ij} = \text{cov}(s_i, s_j) \). Since the variances and covariances involved are known, the denominator degrees of freedom \( \nu \) in Scheffé’s
\[ S^2 = (t - 1)F_\alpha(t - 1, \nu) \]
is, effectively, \( \infty \). Hence \( S^2 \rightarrow \chi^2_{\alpha, t - 1} \). We then have, asymptotically under (3.32),
\[ \Pr \left\{ \forall i \neq j : \theta_{ij} \in (d_i - d_j) \pm \sqrt{\chi^2_{\alpha, t - 1} \cdot 2 \left( 1 - \frac{\sigma_{ij}}{\sigma_i \sigma_j} \right)} \right\} = 1 - \alpha. \quad (3.58) \]

Tukey’s procedure

Tukey’s (1951) procedure requires that the scores have a common variance and a common pairwise correlation \( \rho \). The former requirement is satisfied, since the standardized scores \( \{d_i\} \) all have variance 1. Supposing we could meet the latter requirement, the half-width of Tukey’s confidence interval for the difference between two
scores would be

\[ T = \frac{1}{2} \sum_{i=1}^{t} |c_i| w_{\alpha,t} \sqrt{1 - \rho} \]

where \( w_{\alpha,t} \) is the upper \( \alpha \) point of the range of \( t \) independent standard normal variates. We can put an upper bound on \( T \) by setting

\[ \rho^* = \min_{(i,j)} \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \]

giving us

\[ \Pr \left\{ \forall i \neq j : \theta_{ij} \in (d_i - d_j) \pm w_{\alpha,t} \sqrt{1 - \rho^*} \right\} \leq 1 - \alpha. \]

**Modified Scores**

**Method**

Instead of using the raw proportions \( p_{ij} = \frac{\alpha_{ij}}{n_{ij}} \) in the preference matrix we could have used the 'modified' proportions

\[ \tilde{p}_{ij} = \frac{\alpha_{ij} + 1}{n_{ij} + 2}. \quad (3.60) \]

Use of such modified proportions tends to dampen the scores, in that \( \tilde{p}_{ij} \) will always be closer to \( \frac{1}{2} \) than will \( p_{ij} \). But this modification will help us distinguish between the situation where \( C_i \) is preferred to \( C_j \) in a single comparison and the situation where \( C_i \) is preferred to \( C_j \) in, say, 10 of 10 comparisons. The unmodified proportion \( p_{ij} \) would be 1 in either case, whereas \( \tilde{p}_{ij} \) would be \( \frac{2}{3} \) in the first instance and \( \frac{10}{11} \) in the second. Object \( C_i \)'s dominance is surely more apparent in the latter case than
in the former, and this difference is reflected in the modified proportions. Certain outcomes will still be indistinguishable with the modified proportions (e.g., 1 out of 1, 3 out of 4, 5 out of 7, 7 out of 10, etc.), but \( \tilde{p}_{ij} \) nonetheless takes into account the volume of data which compare \( C_i \) and \( C_j \). (We can think of \( \tilde{p}_{ij} \) as a Bayesian estimate of \( \pi_{ij} \) from a \( \beta(1,1) \) prior updated by \( n_{ij} \) comparisons.)

Distribution theory

The expectations, variances, and covariances for these modified proportions are, respectively,

\[
E(\tilde{p}_{ij}) = \frac{n_{ij} \pi_{ij} + 1}{n_{ij} + 2}
\]

\[
V(\tilde{p}_{ij}) = \frac{n_{ij} \pi_{ij} \pi_{ji}}{(n_{ij} + 2)^2}
\]

\[
\text{cov}(\tilde{p}_{ij}, \tilde{p}_{kl}) = \begin{cases} 
V(\tilde{p}_{ij}) & (i, j) = (k, l) \\
-V(\tilde{p}_{ij}) & (i, j) = (l, k) \\
0 & \text{otherwise.}
\end{cases}
\]

Since \( n_{ij} = 0 \) gives \( \tilde{p}_{ij} = \frac{1}{2} \), we can proceed as if there were no empty cells, even when some pairs have not been compared. Following (3.10) we see that the corresponding modified scores \( \{\tilde{s}_i\} \) will be equivalent to row-sums of the (modified) proportions:

\[
\tilde{s}_i = t \sum_{j \neq i} (\tilde{p}_{ij} - \frac{1}{2}).
\]

In matrix notation this is equivalent to

\[
\tilde{s} = t \tilde{W} - \left( \frac{t}{2} \right) \mathbf{1}, \quad (3.61)
\]
where \( \mathbf{\bar{w}} \) is the vector of row-sums of the (modified) preference matrix. We then have

\[
E(\tilde{s}_i) = \frac{1}{t} \sum_{j \neq i} \left( \frac{n_{ij}\pi_{ij} + 1}{n_{ij} + 2} - \frac{1}{2} \right)
\]

\[
V(\tilde{s}_i) = \frac{t^2}{(n_{ij} + 2)^2} \sum_{j \neq i} \frac{n_{ij}\pi_{ij} \pi_{ji}}{n_{ij} + 2}
\]

\[
\text{cov}(\tilde{s}_i, \tilde{s}_j) = -\frac{t^2 n_{ij}\pi_{ij} \pi_{ji}}{(n_{ij} + 2)^2}
\]

Comparison with unmodified scores

Suppose that, for each pair \((i, j)\), say, one of the objects is favored in all \(n_{ij}\) comparisons. (This must, in fact, be the case if all \(n_{ij} = 0\) or \(1\).) Then our unmodified (maximum likelihood) estimate \(p_{ij}\) of \(\pi_{ij}\) would be 0 or 1, and its estimated variance \(\frac{p_{ij}p_{ji}}{n_{ij}}\) would be zero. As each score is a linear combination of these proportions, our estimates of the scores' variances will all be zero. Use of the modified proportions alleviates this problem, since we get non-zero estimates even in these situations.

Speaking more generally and dropping the subscripts for the time being, let us consider a single random variable \(\alpha \sim \text{binomial}(n, \pi)\). Our unmodified estimator of \(\pi\) is \(\hat{p} = \alpha/n\). We see that

\[
E(\hat{p}) = \pi \quad \text{and hence}
\]

\[
bias(\hat{p}) = 0, \quad \text{and}
\]

\[
V(\hat{p}) = \text{MSE}(\hat{p}) = \frac{\pi(1 - \pi)}{n}.
\]

Our modified estimator of \(\pi\) is \((\alpha + 1)/(n + 2)\), for which we have

\[
E(\tilde{\hat{p}}) = \frac{n\pi + 1}{n + 2}
\]
Comparing these two estimators, we see that

\[
\frac{\text{MSE}(\hat{p})}{\text{MSE}(p)} = \frac{\frac{1}{(n+2)^2} \left[ n\pi(1 - \pi) + (1 - 2\pi)^2 \right]}{\frac{\pi(1 - \pi)}{n}}
= \frac{n}{(n+2)^2} \left[ (n-4)\pi(1 - \pi) + 1 \right].
\]

(3.62)

For this ratio to equal unity we must have

\[
\pi = \frac{1}{2} \left( 1 \pm \sqrt{1 - \frac{n}{2n+1}} \right).
\]

For \(\pi\) within these bounds, the ratio in (3.62) is less than 1; so only for the extreme values of \(\pi\) close to 0 or 1 will \(p\) have a smaller \(\text{MSE}\) than \(\hat{p}\). As \(n \to \infty\) these bounds close in on

\[
\pi \approx \{0.146, 0.854\}.
\]

For small \(n\) these bounds are even more favorable to the unmodified proportions.

Note, finally, that for \(\pi = \frac{1}{2}\)

\[
\frac{\text{MSE}(\hat{p})}{\text{MSE}(p)} = \left( \frac{n}{n+2} \right)^2,
\]
since \( \tilde{p} \) is then also unbiased.

This suggests that, for moderate (i.e., central) values of the \( \{ \pi_{ij} \} \), the modified score \( \tilde{s}_i \) is more efficient than the unmodified score \( s_i \). This is difficult to check, however, since the ratio of their MSEs is unmanageable in the general case. For the case with no empty cells and under the hypothesis of randomness, the difference

\[
\text{MSE}(\tilde{s}_i) - \text{MSE}(s_i) = -t^2 \sum_{j \neq i} \frac{n_{ij} + 1}{n_{ij}(n_{ij} + 2)^2}
\]

is indeed negative. But even if \( \text{MSE}(\tilde{p}) < \text{MSE}(p) \) for all pairings involving \( C_i \), it does not necessarily follow that \( \tilde{s}_i \) is more efficient than \( s_i \), since

\[
\frac{\text{MSE}(\tilde{s}_i)}{\text{MSE}(s_i)} = \frac{\sum_{j \neq i} \text{MSE}(\tilde{p}_{ij}) + \sum_{j \neq i} \sum_{k \neq i,j} \text{bias}(\tilde{p}_{ij}) \cdot \text{bias}(\tilde{p}_{ik})}{\sum_{j \neq i} \text{MSE}(p_{ij})}.
\]

From this we see that if all the \( \{ \pi_{ij} \} \) involving a given object \( C_i \) tend to lie on the same side of \( \frac{1}{2} \), the second term in the numerator of the above RHS will be positive, so that the ratio might be greater than one.

**Numerical Examples**

**Cowden’s data**

Consider the following set of data, reported by Cowden (1974), showing the records against each other of the four top women’s tennis professionals from 1971:

\[
\begin{bmatrix}
* & 1 & 2 & . \\
0 & * & 1 & 11 \\
6 & 1 & * & 1 \\
. & 1 & 0 & *
\end{bmatrix}
\]
It is immediately apparent that these data are highly unbalanced, and are incomplete in that players 1 and 4 did not meet that year. Note that the row-sum vector $(3,12,8,1)$ is clearly not an appropriate criterion for ranking because of the volume of matches between players 2 and 4. Player 2 in this case would enjoy an inflated score simply by virtue of her dominance over player 4 and the good fortune to have 12 matches scheduled with this opponent.

**Cowden's method**

Let us start Cowden's procedure with $u^{(0)} = v^{(0)} = \frac{1}{2} 1_t$, as recommended. The intermediate calculations from (2.1) and (2.2) are straightforward for the first iteration:

\[
p^{(1)} = Au^{(0)} = \begin{bmatrix} 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 11 \\ 6 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} .5 \\ .5 \\ .5 \\ .5 \end{bmatrix} = \begin{bmatrix} 1.50 \\ 6.00 \\ 4.00 \\ 0.50 \end{bmatrix}
\]

\[
q^{(1)} = Bv^{(0)} = \begin{bmatrix} 0 & 0 & 6 & 0 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} .5 \\ .5 \\ .5 \\ .5 \end{bmatrix} = \begin{bmatrix} 3.00 \\ 1.50 \\ 1.50 \\ 6.00 \end{bmatrix}
\]
The resulting score vectors from (2.3) and (2.4) are then

\[
\begin{bmatrix}
4.50 & .33 & .67 \\
7.50 & .80 & .20 \\
5.50 & .73 & .27 \\
6.50 & .08 & .92
\end{bmatrix}
\]

Note that the score vectors are, respectively, the proportion of matches won and lost for each player. For the second iteration we get

\[
p^{(2)} = Av^{(1)} = \begin{bmatrix} 0 & 1 & 2 & 0 & .33 & 2.26 \\ 0 & 0 & 1 & 1 & .80 & 1.57 \\ 6 & 1 & 0 & 1 & .73 & 2.88 \\ 0 & 1 & 0 & 0 & .08 & 0.80 \end{bmatrix}
\]

\[
q^{(2)} = Bv^{(1)} = \begin{bmatrix} 0 & 0 & 6 & 0 & .67 & 1.64 \\ 1 & 0 & 1 & 1 & .20 & 1.86 \\ 2 & 1 & 0 & 0 & .27 & 1.53 \\ 0 & 1 & 1 & 0 & .92 & 2.47 \end{bmatrix}
\]

and score vectors

\[
\begin{bmatrix}
3.90 & .58 & .42 \\
3.43 & .46 & .54 \\
4.41 & .65 & .35 \\
3.27 & .24 & .76
\end{bmatrix}
\]
As the procedure continues, we focus on the vector of 'win scores' on which the ranks are based. These scores converge quite slowly, stabilizing to three decimal places only after 33 iterations:

\[
\begin{align*}
\mathbf{u}(3) & = \begin{bmatrix} .458 \\ .687 \\ .751 \\ .068 \end{bmatrix} \\
\mathbf{u}(4) & = \begin{bmatrix} .595 \\ .465 \\ .715 \\ .157 \end{bmatrix} \\
\mathbf{u}(5) & = \begin{bmatrix} .525 \\ .614 \\ .757 \\ .070 \end{bmatrix} \\
\mathbf{u}(6) & = \begin{bmatrix} .593 \\ .481 \\ .742 \\ .120 \end{bmatrix} \\
\mathbf{u}(7) & = \begin{bmatrix} .559 \\ .572 \\ .757 \\ .075 \end{bmatrix} \\
\mathbf{u}(8) & = \begin{bmatrix} .589 \\ .495 \\ .753 \\ .104 \end{bmatrix} \\
& \vdots \\
\mathbf{u}(33) & = \begin{bmatrix} .585 \\ .522 \\ .759 \end{bmatrix}
\end{align*}
\]

Note that each score jumps back and forth in cycles of decreasing amplitude across its (eventual) limit. To hasten convergence, Cowden suggests taking as \(u_i^{(k+1)}\) the mean of \(u_i^{(k)}\) and \(u_i^{(k-1)}\) "whenever it seems useful". But even if this is done at every stage, the scores converge to three places only after 30 iterations, with slightly different final values (.593, .529, .764, .089).

**Least-squares method** For the least-squares approach we need data in the form of observed 'differences' for all the pairs which have been compared. For simplicity, take as the observed difference \(d_{ij}\) between \(C_i\) and \(C_j\) the difference \(p_{ij} - p_{ji}\). (Gulliksen favors a transformation of these proportions; for clarity of illustration, however, we will use the raw proportions.) If we begin the procedure with a null initial vector \(s(0) = 0\), our initial predicted differences \(s_i^{(0)} - s_j^{(0)}\) are all zero, and the discrepancies (between the predicted and observed differences) are then simply the observed differences. Following (2.6) the next set of scores \(s^{(1)}\) is just the vector
of row 'averages', obtained by dividing each row sum by $m_i + 1$:

$$\begin{array}{ccc}
\text{discrepancies} & \text{row-sums} & \text{averages} \\
\begin{bmatrix}
* & 1.00 & -.50 \\
-1.00 & * & 0.00 & .83 \\
.50 & 0.00 & * & 1.00 \\
* & -.83 & -1.00 & *
\end{bmatrix}
& \begin{bmatrix}
.50 \\
-1.7 \\
1.50 \\
-1.83
\end{bmatrix}
& \begin{bmatrix}
.17 \\
-.04 \\
.38 \\
-.61
\end{bmatrix}
\end{array}$$

The sum of squared discrepancies was initially equal to the total sum of squares (namely, 2.94), since the initial predicted differences were all zero. By the next iteration, this sum of squares drops to 0.96 as the 'fit' between the predicted and observed differences gets better:

$$\begin{array}{ccc}
\text{observed} & \text{predicted} = s_{i}^{(1)} - s_{j}^{(1)} \\
\begin{bmatrix}
* & 1.00 & -.50 \\
-1.00 & * & 0.00 & .83 \\
.50 & 0.00 & * & 1.00 \\
* & -.83 & -1.00 & *
\end{bmatrix}
& \begin{bmatrix}
* & .21 & -.21 \\
-.21 & * & -.42 & .57 \\
.21 & .42 & * & .99 \\
* & -.57 & -.99 & *
\end{bmatrix}
\end{array}$$

$$\begin{array}{ccc}
\text{discrepancies} & \text{row-sums} & \text{averages} \\
\begin{bmatrix}
* & .79 & -.29 \\
-.79 & * & .42 & .26 \\
.29 & -.42 & * & .01 \\
* & -.26 & -.01 & *
\end{bmatrix}
& \begin{bmatrix}
.50 \\
-.11 \\
-.12 \\
-.27
\end{bmatrix}
& \begin{bmatrix}
.17 \\
-.03 \\
-.03 \\
-.09
\end{bmatrix}
\end{array}$$

The sum of squared discrepancies is now only 0.81, and will bottom out at .79 as the scores approach the least-squares solution. Convergence is fairly rapid, as the scores
stabilize to three decimal places after 7 iterations:

\[
\begin{bmatrix}
  .370 \\
  -.065 \\
  .352 \\
  -.753 \\
\end{bmatrix}
\begin{bmatrix}
  .386 \\
  -.066 \\
  .351 \\
  -.766 \\
\end{bmatrix}
\begin{bmatrix}
  .390 \\
  -.065 \\
  .351 \\
  -.771 \\
\end{bmatrix}
\begin{bmatrix}
  .392 \\
  -.065 \\
  .351 \\
  -.773 \\
\end{bmatrix}
\begin{bmatrix}
  .393 \\
  -.065 \\
  .351 \\
  -.774 \\
\end{bmatrix}
\]

**Chebotarev's method** If we take the \(a_{ij}^{(k)}\) to be \(\pm 1\), Chebotarev's row sums are just the net number of preferences in favor of each object:

\[
u = A\mathbf{1}_t - A'\mathbf{1}_t = \begin{bmatrix}
3 \\
12 \\
8 \\
1 \\
\end{bmatrix} - \begin{bmatrix}
6 \\
3 \\
3 \\
12 \\
\end{bmatrix} = \begin{bmatrix}
-3 \\
9 \\
5 \\
-11 \\
\end{bmatrix}.
\]

This would, of course, be equal to Chebotarev's score vector \(x\) if we choose \(\epsilon = 0\).

It is evident from Chebotarev's calculations that he takes the number of rounds to be \(m = \max n_{ij}^{(k)}\), which would be \(n_{24} = 12\) for our data. Equation (2.16) then becomes

\[
x = (1 + 48\epsilon) \left( \begin{bmatrix}
9 & -1 & -8 & 0 \\
-1 & 15 & -2 & -12 \\
-8 & -2 & 11 & -1 \\
0 & -12 & -1 & 13 \\
\end{bmatrix} \right)^{-1} \begin{bmatrix}
-3 \\
9 \\
5 \\
-11 \\
\end{bmatrix}.
\]
Chebotarev's scores do indeed change considerably as $\epsilon$ varies:

$$
\begin{align*}
\epsilon : & \quad 0.00 \quad 0.01 \quad 0.05 \\
& \begin{bmatrix}
-3.00 \\
9.00 \\
5.00 \\
-11.00 \\
\end{bmatrix} \\
& \begin{bmatrix}
-3.50 \\
10.28 \\
6.48 \\
-13.26 \\
\end{bmatrix} \\
& \begin{bmatrix}
-3.83 \\
11.76 \\
10.16 \\
-18.08 \\
\end{bmatrix} \\
& \begin{bmatrix}
0.00 \\
0.25 \\
0.50 \\
1.00 \\
\end{bmatrix} \\
& \begin{bmatrix}
-3.20 \\
-1.13 \\
0.92 \\
2.71 \\
\end{bmatrix} \\
& \begin{bmatrix}
11.63 \\
10.21 \\
8.64 \\
7.23 \\
\end{bmatrix} \\
& \begin{bmatrix}
12.69 \\
16.40 \\
18.93 \\
20.86 \\
\end{bmatrix} \\
& \begin{bmatrix}
-21.12 \\
-25.48 \\
-28.49 \\
-30.81 \\
\end{bmatrix}
\end{align*}
$$

We see that $C_2$'s top row-sum is overshadowed as $\epsilon$ gets larger, i.e., as we put more weight on the strengths of the objects.

**David's method**  
David's method uses in $A$ the proportion of matches won within each pairing:

$$
A = \begin{bmatrix}
* & 1.00 & 0.25 & \\
0.00 & * & 0.50 & 0.92 \\
0.75 & 0.50 & * & 1.00 \\
& 0.08 & 0.00 & *
\end{bmatrix}
\quad w = \begin{bmatrix}
1.25 \\
1.42 \\
2.25 \\
0.08
\end{bmatrix}
\quad w^{(2)} = \begin{bmatrix}
1.98 \\
1.20 \\
1.73 \\
0.12
\end{bmatrix}
\quad s = \begin{bmatrix}
1.92 \\
-0.25 \\
2.25 \\
-3.92
\end{bmatrix}
$$

$l' = \begin{bmatrix}
0.75 & 1.58 & 0.75 & 1.92
\end{bmatrix}$

$l^{(2)}' = \begin{bmatrix}
0.56 & 1.28 & 0.98 & 2.20
\end{bmatrix}$
The various components $w, l, w^{(2)},$ and $l^{(2)}$ of his score vector $s$ from (3.1) are, respectively, the row-sums of $A, A', A^2,$ and $(A')^2$.

In contrast to the previous two iterative methods, David's scores are calculated in a few short steps. Perhaps even simpler would be to calculate the $\{s_i\}$ via (3.9):

\[
\begin{align*}
    s_1 &= (m_2 + 1)(p_{12} - \frac{1}{2}) + (p_{24} - \frac{1}{2}) + (m_3 + 1)(p_{13} - \frac{1}{2}) + (p_{34} - \frac{1}{2}) \\
    &= 4(.50) + (.42) + 4(-.25) + (.50) \\
    &= 1.92 \\
    s_2 &= (m_1 + 1)(p_{21} - \frac{1}{2}) + (m_3 + 1)(p_{23} - \frac{1}{2}) + (m_4 + 1)(p_{24} - \frac{1}{2}) \\
    &= 3(-.50) + 4(.00) + 3(.42) \\
    &= -.25 \\
    s_3 &= (m_1 + 1)(p_{31} - \frac{1}{2}) + (m_2 + 1)(p_{32} - \frac{1}{2}) + (m_4 + 1)(p_{34} - \frac{1}{2}) \\
    &= 3(.25) + 4(.00) + 3(.50) \\
    &= 2.25 \\
    s_4 &= (m_2 + 1)(p_{42} - \frac{1}{2}) + (p_{21} - \frac{1}{2}) + (m_3 + 1)(p_{43} - \frac{1}{2}) + (p_{31} - \frac{1}{2}) \\
    &= 4(-.42) + (-.50) + 4(-.50) + (.25) \\
    &= -3.92
\end{align*}
\]
Modified scores  The modified scores \( \{ \tilde{s}_i \} \) are easily calculated by (3.61) with \( t = 4 \) from the matrix of modified proportions \( \{ \tilde{p}_{ij} \} \) (rather than the \( \{ p_{ij} \} \)):

\[
\begin{bmatrix}
  * & 2/3 & 3/10 & 1/2 \\
  1/3 & * & 2/4 & 12/14 \\
  7/10 & 2/4 & * & 2/3 \\
  1/2 & 2/14 & 1/3 & * \\
\end{bmatrix}
\begin{bmatrix}
  \tilde{w} \\
  s = 4\tilde{w} - 6 \cdot 1
\end{bmatrix}
\begin{bmatrix}
  1.47 \\
  1.69 \\
  1.87 \\
  0.98 \\
\end{bmatrix}
\begin{bmatrix}
  -0.13 \\
  .76 \\
  1.47 \\
  -2.09 \\
\end{bmatrix}
\]

Comparison of the methods  Since direct comparison of the various methods is rather awkward because of the different scales used, it is helpful to standardize each set of scores by subtracting off the mean and dividing by the standard deviation of each set:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>David's s</td>
<td>0.68</td>
<td>-0.09</td>
<td>0.79</td>
<td>-1.38</td>
</tr>
<tr>
<td>modified ( \tilde{s} )</td>
<td>-0.09</td>
<td>0.49</td>
<td>0.95</td>
<td>-1.36</td>
</tr>
<tr>
<td>least-squares</td>
<td>0.77</td>
<td>-0.08</td>
<td>0.69</td>
<td>-1.39</td>
</tr>
<tr>
<td>Cowden's scores</td>
<td>0.34</td>
<td>0.12</td>
<td>0.94</td>
<td>-1.41</td>
</tr>
<tr>
<td>Chebotarev's scores</td>
<td>-0.06</td>
<td>0.55</td>
<td>0.89</td>
<td>-1.38</td>
</tr>
</tbody>
</table>

(For Chebotarev's scores we have here used \( \epsilon = 0.25 \), a value which he favors in his examples.) All five methods agree that player 4 is by far the weakest, but there are considerable differences in the scores of the other three. The least-squares approach favors player 1; the other four methods agree that player 3 is best. Cowden's method agrees entirely with David's on all the ranks; the ranking from the modified scores \( \{ \tilde{s}_i \} \) differs from these two in that player 2 is ranked ahead of player 1, due mostly
to the 'dampening' effect of $\hat{p}_{12} = 2/3$ in contrast to $p_{12} = 1$. That Chebotarev's method also favors $C_2$ to $C_1$ is not surprising, since his method does not account well for drastically different $n_{ij}$.

The value of the multibinomial test statistic for these data is 12.33 on 5 degrees of freedom. As $\chi^2_{5,0.95} = 11.1$, this test bears strong evidence against the hypothesis of randomness $H_0: \pi_{ij} = \frac{1}{2}$ for $(i,j)$.

Indeed, with the exception of the 1-1 record between players 2 and 3, the pairwise records are all fairly lopsided.

The value of our quadratic form for testing the equality of the players, however, is only $Q = 3.34$. The significance of this value depends of course on the null distribution of the statistic. We can generate the exact null distribution of $Q$ by exhaustively listing all possible outcomes of the experiment (given the $\{n_{ij}\}$), calculating the value of $Q$ from each outcome, and tabulating the probabilities with which $Q$ takes on each of these values. Selected values of this distribution are given below, from
which we see that our observed value of $Q$ falls about the $62^{nd}$ percentile.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$\Pr{Q \leq q}$</th>
<th>$\Pr{Q &gt; q}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.0312</td>
<td>0.9759</td>
</tr>
<tr>
<td>2</td>
<td>0.3138</td>
<td>0.7174</td>
</tr>
<tr>
<td>3</td>
<td>0.5713</td>
<td>0.4541</td>
</tr>
<tr>
<td>4</td>
<td>0.7735</td>
<td>0.2613</td>
</tr>
<tr>
<td>5</td>
<td>0.9238</td>
<td>0.0844</td>
</tr>
<tr>
<td>6</td>
<td>0.9725</td>
<td>0.0319</td>
</tr>
<tr>
<td>7</td>
<td>0.9903</td>
<td>0.0122</td>
</tr>
<tr>
<td>8</td>
<td>0.9972</td>
<td>0.0035</td>
</tr>
<tr>
<td>9</td>
<td>0.9989</td>
<td>0.0012</td>
</tr>
<tr>
<td>10</td>
<td>0.9997</td>
<td>0.0003</td>
</tr>
<tr>
<td>11</td>
<td>0.9999</td>
<td>0.0001</td>
</tr>
<tr>
<td>12</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The exact upper .05 and .01 critical values for $Q$ are seen to be about 5.6 and 7.1, respectively. To illustrate how conservative the $\chi^2$ approximation is for such small experiments, consider that the upper .05 and .01 critical values for a $\chi^2$ variate with 3 degrees of freedom are, respectively, 7.81 and 11.84. From the exact distribution of $Q$ we see that use of these values leads to tests of size .0043 and .000055, respectively.

But from the approximation or from the exact distribution, the observed value of $Q$ is obviously nowhere close to significance, despite player 4's standing far below the other three. The lack of power here seems due primarily to the sparsity of the
data: aside from the 20 matches from the (1,3) and (2,4) pairings, there were only 4 other matches recorded. There is, furthermore, considerable intransitivity in these data, as borne out by the least-squares ANOVA from (2.7):

<table>
<thead>
<tr>
<th>source</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted</td>
<td>3</td>
<td>2.15</td>
</tr>
<tr>
<td>discrepancies</td>
<td>2</td>
<td>.79</td>
</tr>
<tr>
<td>total</td>
<td>5</td>
<td>2.94</td>
</tr>
</tbody>
</table>

Kaiser and Serlin's suggested measure of 'internal consistency' from (2.8) is here only $r^2 = 0.73$.

We can use the covariance expressions from section 3 to construct the covariance matrix of David's scores \( \{s_j\} \):

\[
\Sigma = \begin{bmatrix}
4.77 & -2.94 & 0.38 & -2.21 \\
-2.94 & 4.44 & -2.00 & 0.50 \\
0.38 & -2.00 & 4.53 & -2.91 \\
-2.21 & 0.50 & -2.91 & 4.62
\end{bmatrix}
\]

Compare now the observed differences of the standardized scores \( \{d_i\} \), where \( d_i = s_i / \sigma_i \), to the minimum significant differences (for \( \alpha = 0.05 \)) calculated by Scheffé's method:

\[
\begin{bmatrix}
* \\
-1.0 & * \\
0.2 & 1.2 & * \\
-2.7 & -1.7 & -2.9 & *
\end{bmatrix}
\begin{bmatrix}
* \\
5.1 & * \\
3.8 & 4.8 & * \\
4.8 & 3.7 & 5.1 & *
\end{bmatrix}
\]
The minimum significant difference between \( d_1 \) and \( d_2 \), say, is calculated as the half-width of the confidence interval for \( \theta_{ij} \) from (3.58), namely,

\[
\sqrt{X_{0.05,3}^2 \cdot \left(1 - \frac{\sigma_{12}}{\sigma_1 \sigma_2}\right)} = \sqrt{7.81 \cdot 2 \cdot \left(1 - \frac{-2.94}{\sqrt{4.77 \cdot 4.44}}\right)} = 5.1.
\]

We see that none of the pairwise differences among David's (standardized) scores is significant. Again, the sparsity of the data for most pairings drove up the variances of the scores.

Tukey's minimum significant difference from (3.59) is

\[
w_{0.05,4} \sqrt{1 - \rho^*} = 3.63 \sqrt{1 - (-.64)} = 4.65.
\]

Note that all of the observed differences are far less than (our modification of) Tukey's minimum significant difference. Recall that we used the smallest pairwise correlation \( \rho^* \) between scores as a lower bound on Tukey's common correlation \( \rho \). In our example this bound was achieved at considerable expense, as the pairwise correlations ranged from -0.64 to 0.11. For data that are not quite so unbalanced, Tukey's minimum significant difference is typically less than even the smallest of Scheffé's differences.

Lanctot's data

The following data pertain to the dominance behavior of a certain species of prairie voles. Specifically, the \((i,j)\) entry is the amount of time (in minutes) which
vole $i$ spent dominating a water resource while in a cell with vole $j$ for 3 minutes.

\[
\begin{bmatrix}
  * & 2.09 & 1.31 & .91 & 0.00 \\
  1.10 & * & 1.69 & .25 & .42 \\
  .04 & .36 & * & .78 & .74 \\
  .87 & 1.31 & .60 & * & .44 \\
  1.87 & 2.16 & & * & 2.24 \\
  & 1.42 & 1.15 & .04 & *
\end{bmatrix}
\]

Some of the time neither vole was dominant, so that $\alpha_{ij} + \alpha_{ji} < 3$ for most pairs. Occasionally both voles were considered to be in control of the water resource, so that, for example, $\alpha_{12} + \alpha_{21} > 3$.

As mentioned in the first chapter, some of the methods can still produce scores from these data, despite the fact that there have not been an integral number of 'comparisons' between each pair. To implement David's method, for example, we need only construct the matrix containing the proportion of time spent in control of the resource:

\[
\begin{bmatrix}
  * & .66 & .97 & .51 & .00 \\
  .34 & * & .82 & .16 & .16 \\
  .03 & .18 & * & .57 & .34 \\
  .49 & .84 & .43 & * & .28 \\
  1.00 & .84 & & * & .98 \\
  & & .66 & .72 & .02 & *
\end{bmatrix}
\]

Calculating scores either by (3.1) or by (3.9) we get

\[s' = [1.29, -2.10, -5.60, -0.90, 6.96, 0.34].\]
Since the distributional results for these scores required use of the two-point scale for judging comparisons, the many tests based on these results would be inappropriate.

Cowden's method can handle these data as well; the procedure produces scores 

\[.55, .42, .25, .44, .92, .47\].

The least-squares approach (as conceived by Gulliksen) and the method of modified scores could each generate scores, but such results would be inappropriate since these two methods were designed to operate on proportions of preferences won and lost (or a function of these proportions).
CHAPTER 4. EXTENSION OF PRESENT APPROACH TO UNBALANCED RANKED DATA

Comparison of Rank and Paired-Comparison Approaches

Of the existing approaches to analysis of ranked data, few methods – namely, Prentice's and Skillings and Mack's – use both the 'centering' and 'scaling' ideas that seem to be essential for sensible analysis of unbalanced data. Just as our paired-comparison method accounted for the differing numbers of comparisons involving each object and for the varied caliber of each object's opposition, so can such rank methods account for the differing numbers of blocks containing each object and for the differing block sizes. And since, as noted above, standardizing by $(k_j + 1)^{-1/2}$ as opposed to $(k_j + 1)^{-1}$ usually has little effect on the relative standing of the objects' scores, we will focus, for computational convenience, on Prentice's method exclusively for comparing existing rank analysis with our paired-comparison analysis of ranked data.

Special case: completely balanced designs

When each object is ranked by each of the $b$ judges, the resulting completely balanced rank design 'decomposes' into paired comparisons that form a completely
balanced paired-comparison design with \( b \) comparisons per pair. Since the number of 'preferences' in favor of an object \( C_i \) derived from any one of its ranks \( r_{ij} \) is just \( r_{ij} - 1 \), we see that each rank-sum is then simply \( r_i = a_i + b \), where \( a_i \) is the total number of preferences for \( C_i \), i.e., the corresponding row-sum of the preference matrix.

From (3.12) it then follows that

\[
s_i = \frac{t}{b} (r_i - b) - \left( \frac{t}{2} \right) = \frac{t}{b} r_i - \frac{t(t+1)}{2}.
\]

(4.1)

And by putting \( k_j = t \) into (1.11) and summing over blocks we see that the \( \{y_i\} \) are also just linear functions of the rank-sums and hence of the \( \{s_i\} \) as well:

\[
y_i = \frac{1}{t+1} r_i - \frac{b}{2} = \frac{b}{t(t+1)} s_i.
\]

Use of the paired-comparison scores is thus equivalent to use of Prentice's scores.

Prentice's \( C \) from (1.12) reduces to Friedman's \( X^2 \) for these balanced rank designs; and since \( Q_{Bal} \) is a function of the sum of squared row-sums for balanced paired-comparison designs, we can use (3.44) and (1.6) to express \( Q_{Bal} \) in terms of \( C_{Bal} \):

\[
Q_{Bal} = \frac{4}{bt} \sum_{i=1}^{t} \left[ a_i - b \left( \frac{t-1}{2} \right) \right]^2
= \frac{4}{bt} \sum_{i=1}^{t} \left[ r_i - b \left( \frac{t+1}{2} \right) \right]^2
= \frac{4}{bt} S_{Bal}
= \frac{t+1}{3} C_{Bal},
\]

(4.2)

where \( S_{Bal} \) is defined in (1.5) for these balanced rank designs. Hence the two statistics are equivalent for testing for the equality of the objects, and \( \frac{3}{t+1} Q_{Bal} \) clearly has the same asymptotic \( \chi^2_t \) distribution as \( C_{Bal} \).
Special case: balanced incomplete block designs

When objects are ranked in blocks of \( k < t \) and each object is ranked with each other object in \( n \) of the \( b \) blocks, the resulting BIB rank design reduces to a completely balanced paired-comparison design with \( n \) comparisons per pair. Again we have a simple relationship between each rank-sum and the corresponding row-sum: \( r_i = a_i + m \), where \( m \) is the number of blocks in which each object appears. So again the paired-comparison score \( s_i \) is a linear function of the simple rank-sum \( r_i \). And since the blocks are still of equal sizes, Prentice's score \( y_i \) is also simply a linear function of \( r_i \) and hence of \( s_i \).

Noting that the number of comparisons involving a given object can be expressed as either \( m(k - 1) \) or \( n(t - 1) \), we can use (3.44) and (1.9) to show that

\[
Q_{BIB} = \frac{4}{nt} \sum_{i=1}^{t} \left[ a_i - \frac{n(t - 1)}{2} \right]^2 \\
= \frac{4}{nt} \sum_{i=1}^{t} \left[ a_i - \frac{m(k - 1)}{2} \right]^2 \\
= \frac{4}{nt} \sum_{i=1}^{t} \left[ r_i - \frac{m(k + 1)}{2} \right]^2 \\
= \frac{4}{nt} S_{BIB} \\
= \frac{k + 1}{3} C_{BIB},
\]

where \( S_{BIB} \) is defined in (1.8) for these BIB rank designs. So again the two statistics are equivalent, and \( \frac{3}{k+1} Q_{BIB} \) has an asymptotic \( \chi^2_{t-1} \) distribution.
Special case: partially balanced incomplete block designs

Suppose we have a PBIB rank design that decomposes into a group divisible paired-comparison design. For example, suppose we have $t = 6$ objects in $b = 4$ blocks of $k = 3$, reducing to $m = 3$ groups of $a = 2$ objects:

<table>
<thead>
<tr>
<th>blocks:</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>groups:</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

(Note that each object is not ranked with the other object in its group, but is ranked exactly once with each remaining object.) In general there will be $b = a^2$ blocks of $k = m$ when there are $m$ groups of $a$. In Appendix D it is shown that

$$Q_{GD} = \frac{m+1}{3}C_{PBIB}$$

thus paralleling the results in (4.2) and (4.3). Similarly, $\frac{3}{m+1}Q_{PBIB}$ then has an asymptotic $\chi^2_{t-1}$ distribution.

Once again, the structure in the data assures that $Q$ and $C$ are equivalent. But now that the corresponding paired-comparison design is no longer completely balanced, our scores $s$ will no longer be equivalent to Prentice’s scores $y$. Consider the following ranked data from a group divisible design with $m = 3$ groups of $a = 3$
objects each:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

(Reading from left to right, the objects are listed from highest-ranked to lowest-ranked within each block.) Note below that, since the blocks are all of the same size, Prentice’s scores are equivalent to the rank sums:

<table>
<thead>
<tr>
<th>object</th>
<th>rank-sums</th>
<th>(y)</th>
<th>(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>7</td>
<td>.25</td>
<td>11</td>
</tr>
<tr>
<td>(C_2)</td>
<td>4</td>
<td>-.50</td>
<td>-13</td>
</tr>
<tr>
<td>(C_3)</td>
<td>4</td>
<td>-.50</td>
<td>-13</td>
</tr>
<tr>
<td>(C_4)</td>
<td>7</td>
<td>.25</td>
<td>6</td>
</tr>
<tr>
<td>(C_5)</td>
<td>6</td>
<td>.00</td>
<td>-2</td>
</tr>
<tr>
<td>(C_6)</td>
<td>7</td>
<td>.25</td>
<td>6</td>
</tr>
<tr>
<td>(C_7)</td>
<td>7</td>
<td>.25</td>
<td>7</td>
</tr>
<tr>
<td>(C_8)</td>
<td>7</td>
<td>.25</td>
<td>7</td>
</tr>
<tr>
<td>(C_9)</td>
<td>5</td>
<td>-.25</td>
<td>-9</td>
</tr>
</tbody>
</table>
Note also that 5 objects are tied for best with rank sum equal to 7. But although Prentice’s scores are locked into the rank sums, our method breaks this tie, since each object’s opposition may be of different caliber. In particular, $C_2$ and $C_3$ are clearly weaker than the rest; bearing in mind that intragroup comparisons are not made, we see that $C_1$ has not enjoyed the luxury of comparisons with these weaker objects. $C_1$ has had to earn its rank sum of 7 against the remaining (stronger) opposition. Hence our method rewards $C_1$ with the top score $s_1 = 11$. None of the rank approaches considered have any such mechanism for preferential treatment of objects facing stronger opposition.

**General Case**

For unpatterned rank data in which no structure is readily apparent, the rank scores may differ considerably from the paired-comparison scores. Consider the following data with $t = 4$ objects and $b = 5$ blocks:

$$
\begin{array}{ccc}
1 & 2 & \\
1 & 3 & \\
1 & 4 & \\
4 & 3 & 2 & 1 \\
2 & 3 & 4
\end{array}
$$

(4.7)

Each object is ranked with each other object twice – once ahead and once behind. Each object receives a paired-comparison score $s_i = 0$, as we expect. Yet Prentice’s score vector $y' = (.20, -.02, -.07, -.12)$ gives distinct scores to all four objects. Looking back to Prentice’s standardization of the centered ranks in (1.11), we see that
each of the constituent comparisons from a ranking of \( k_j \) objects is given 'weight' 
\[(k_j + 1)^{-1}.\] Hence a comparison in a smaller block is weighted more than a comparison in a larger block. It is through this that Prentice's method favors \( C_1 \), since all of its 'wins' occurred in blocks of size 2, and all of its 'losses' in the block containing all 4 objects. This is markedly different from the extension of our paired-comparison approach, in which the influence of a particular comparison depends on the total number of comparisons between the two objects, and not on the size of the block from which the comparison is derived.

A related point is illustrated by the following data with \( t = 5 \) objects and \( b = 5 \) blocks:

\[
\begin{array}{cccc}
1 & 2 \\
1 & 2 \\
1 & 2 \\
2 & 3 & 4 & 5 \\
5 & 4 & 3 & 1 \\
\end{array}
\]

Here \( C_1 \) is ranked ahead of \( C_2 \) only, and is ranked behind all other objects; conversely, \( C_2 \) is ranked behind \( C_1 \) only, and is ranked ahead of all others. As there is a certain parity among the remaining objects, our scores \( s' = (-5, 5, 0, 0, 0) \) seem quite reasonable. But Prentice's scores \( y' = (.2, -.2, 0, 0, 0) \) tell quite a different story, as the top and bottom scores are reversed. This drastic switch in the overall ranking is due primarily to the fact that \( C_2 \)'s lone weakness to \( C_1 \) was exploited in 3 of the 5 blocks. None of the considered rank methods can account for the differing numbers of blocks containing a given pair of objects.

The rank approaches can be particularly inadequate when the data is sparse, as
in the following example with \( t = 6 \) objects and only \( b = 3 \) blocks:

\[
\begin{array}{ccc}
1 & 2 & 3 \\
4 & 5 & 6 \\
1 & 3 & 4 & 6
\end{array}
\]

(4.9)

Note that the two subsets of objects \( \{C_1, C_2, C_3\} \) and \( \{C_4, C_5, C_6\} \) are connected only in the third block. Intra-subset rankings are given in the first two blocks, and an inter-subset ordering is established in the third block. The first subset is seen to be stronger than the second; it should be evident that the indices have been selected so that lower-indexed objects are stronger. This clear ordering among the objects is borne out by our scores \( s' = (9, 2, 1, -1, -2, -9) \). But Prentice's scores \( y' = (.55, .00, -.15, .15, .00, -.55) \) suggest, for example, that \( C_2 \) and \( C_5 \) deserve equal scores, since each was ranked second in a block of three. Considering that \( C_2 \) was ranked in the middle of the superior subset and \( C_5 \) in the middle of the inferior subset, we see how Prentice’s method ignores the strength of each object’s opposition. Specifically, the rank approaches fail to account for the connectedness between subsets of objects; such information is of crucial importance with sparse data.

For unpatterned data, there is no exact functional relationship between \( Q \) and \( C \). Generally speaking, we can say that

\[
Q \approx \frac{k^* + 1}{3} C,
\]

where \( k^* \) is an averaged block size in some sense. But there is no longer a one-to-one relationship between the values of \( Q \) and the values of \( C \).

The \( \chi^2 \) approximation to the distribution of \( C \) is rather crude for unpatterned data, especially in the upper tail, which is of primary concern for determining critical
values for significance tests. Specifically, the test based on $C$ tends to be quite conservative. (Oddly enough, the distribution of $Q$ itself, not $\frac{3}{k^2+1}Q$ or any such multiple of $Q$, is fairly close to a $\chi^2_{k-1}$ distribution in the upper tail.) But bearing in mind the discrete nature of these statistics (e.g., $C$ and $Q$ can assume only 18 distinct values for a balanced rank design with $b = 3$ blocks and $t = 4$ objects), it is recommended to generate the desired statistic's true distribution (either by simulation or exact computation) when an exact size-$\alpha$ test is required.

Numerical Examples

Quality of film emulsions

Examples of unbalanced ranked data are hard to find in the literature, presumably because there are so few methods for analysis of such data. The illustrative examples accompanying these few existing methods are typically fabricated. So to get some unbalanced ranked data, we will start with some completely balanced data, then randomly remove several objects from each block.

Johnson and Leone (1977) report six judges' (complete) rankings of the quality of several varieties of color film coated with different emulsions. Below on the left are the rankings remaining after 3 to 6 objects were removed from each block; on the
right are the corresponding ranks \( r_{ij} \) for each object within each block.

\[
\begin{array}{cccccccc}
6 & 3 & 1 & 4 & & & & \\
8 & 3 & 2 & 7 & & & & \\
8 & 7 & & & & & & \\
3 & 7 & 6 & 5 & 4 & & & \\
5 & 8 & 4 & 6 & & & & \\
5 & 1 & 2 & 4 & 7 & & & \\
\end{array}
\]

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
2 & 3 & 1 & 4 & & & & \\
2 & 3 & & & & & 1 & 4 \\
& & & & & 1 & 2 & \\
& & & & & 5 & 1 & 2 & 3 & 4 & \\
& & & & & 2 & 4 & 1 & 3 & & \\
4 & 3 & 2 & 5 & 1 & & & & \\
\end{array}
\]

From the raw ranks \( r_{ij} \) we can easily calculate the standardized ranks \( y_{ij} \) from (1.11); Prentice's scores are simply the sums of these standardized ranks:

\[
\begin{array}{cccccccc}
\text{object:} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\text{score:} & .07 & -.10 & .53 & -.90 & .47 & .00 & -.63 & .57 \\
\end{array}
\]

Despite the fact that, on the average, half of the objects were removed from each block, only two of the 28 pairs of objects failed to appear together in at least one block. Hence the resulting paired-comparison experiment has only two 'empty cells' and is thus nearly complete, though far from balanced. As may be checked from the equivalent preference matrix below, 2 of the pairs were 'compared' three times, 9 were
Our scores, calculated either by (3.1) or (3.9), are then

\[ s' = [-2.00, -6.67, 15.50, -19.67, 10.50, -1.67, -11.00, 15.00]. \]

Prentice’s and our scores are both centered about zero; to facilitate comparison of the two methods, we can scale the scores so that each set has unit variance:

<table>
<thead>
<tr>
<th>object:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prentice's method:</td>
<td>.13</td>
<td>-.18</td>
<td>.98</td>
<td>-1.66</td>
<td>.87</td>
<td>.00</td>
<td>-1.16</td>
<td>1.05</td>
</tr>
<tr>
<td>Proposed method:</td>
<td>-.16</td>
<td>-.52</td>
<td>1.22</td>
<td>-1.54</td>
<td>.82</td>
<td>-.13</td>
<td>-.86</td>
<td>1.18</td>
</tr>
</tbody>
</table>

The standardized scores for \( C_2, C_3, \) and \( C_7 \) are quite different in the two approaches, so much so that the two methods disagree on the top two rankings: Prentice’s method favors \( C_8 \), whereas our method favors \( C_3 \). The two methods disagree on the standings of \( C_6 \) and \( C_1 \) as well.
Sire data

In order to analyze the difference in milk production over various milking seasons, the volume of milk output was recorded for each of a certain herd’s cows whose first lactation occurred in one of six consecutive seasons from 1979 to 1981. Because the volume of milk delivered in first lactation depends on the cow’s age and on the number of days open (i.e., the time between pregnancies), the measurements were adjusted to account for these two factors. A natural blocking variable for this analysis would be the genetic stock of the cows. Thinking of the sires then as blocks, we can look at the average milk production for each of a sire’s daughters that had her first lactation in a given season. We might be hesitant to use the traditional normal-theory block analysis. We might then consider ranking, for each sire, the milk production for the various seasons in which at least one of that sire’s daughters had her first lactation. Such data for four sires is presented below for the six seasons of interest:

<table>
<thead>
<tr>
<th>season:</th>
<th>1 2 3 4 5 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 1</td>
</tr>
<tr>
<td></td>
<td>2 6 3 5 1</td>
</tr>
<tr>
<td></td>
<td>2 4 6 3</td>
</tr>
<tr>
<td></td>
<td>5 6</td>
</tr>
<tr>
<td></td>
<td>1 . . . 2</td>
</tr>
<tr>
<td></td>
<td>1 5 3 . 2 4</td>
</tr>
<tr>
<td></td>
<td>. 4 1 3 . 2</td>
</tr>
<tr>
<td></td>
<td>. . . 2 1</td>
</tr>
</tbody>
</table>

Again, the seasons’ rankings within blocks (i.e., sires) are given on the left, and the corresponding ranks \( r_{ij} \) are on the right. The standardized scores for the two
methods are

<table>
<thead>
<tr>
<th>season</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prentice's method:</td>
<td>-1.29</td>
<td>1.62</td>
<td>-.77</td>
<td>.26</td>
<td>.00</td>
<td>.18</td>
</tr>
<tr>
<td>Proposed method:</td>
<td>-1.40</td>
<td>1.51</td>
<td>-.35</td>
<td>.64</td>
<td>-.47</td>
<td>.06</td>
</tr>
</tbody>
</table>

As $C_2$ was ranked first in the only two blocks in which it appears, it should not be surprising that both methods give it the highest score. Similarly, since $C_1$ was ranked last in each of its two appearances, it is scored far and away the lowest. The middle four scores, however, differ a fair bit from one method to the other. In particular, the standings of $C_3$ and $C_5$ are reversed, with $C_5$ scored higher by Prentice's method by virtue of its performance in the scant fourth block.

Prentice's $C$ for testing equality of the treatments is 6.86 for these data, which falls at about the 79th percentile of the exact null distribution of $C$. The exact upper .05 and .01 critical values for $C$ are 8.3 and 8.9, respectively. Our observed value $Q = 10.62$ falls at about the 71st percentile of its null distribution; the corresponding critical values are 13.3 and 14.9 for $Q$. The critical values 11.07 and 15.09 from the $\chi^2$ approximation (with 5 degrees of freedom) are clearly much closer to the critical values of $Q$ than to those of $C$ for this experiment.
CHAPTER 5. SUMMARY AND CONCLUSIONS

Much of the paired-comparison work in the literature is parametric, in the sense that the analysis is based on a particular linear model. Of the nonparametric methods, there has been some work done concerning balanced paired-comparison experiments. But nonparametric analysis of unstructured or even partially balanced paired-comparison data has received relatively little attention. The aspects of unbalanced data of greatest concern in this paper have been

- the different numbers of comparisons involving each object, and
- the varied caliber of each object's opposition.

Each of the nonparametric approaches addressing these concerns has its strengths and weaknesses. Cowden's method, as an iterative method, is theoretically intractable and at times unpredictable. Gulliksen's method also has the drawback of an iterative solution; yet its basis on optimizing the least-squares criterion is surely appealing, as is its vehicle for checking goodness-of-fit. Chebotarev's method only requires solving a system of linear equations; it is unclear, however, how much weight should be placed on the strength of each object's opposition (i.e., choosing $\epsilon$). David's scores are by far the easiest to calculate. And since his method was developed with the above two concerns in mind, it is not surprising that his method accounts for these concerns
more consistently and thoroughly than the other methods considered.

There are even fewer approaches in the literature to the analysis of unbalanced ranked data, and our paired-comparison extension to ranked data compares quite favorably to these rank methods. None of the rank methods can adequately account for the differing numbers of blocks in which the pairs of objects appear, or for the connectedness of the data in highly incomplete experiments. The paired-comparison extension provides a natural solution to these problems.

The tests here proposed for David's scores should be helpful, especially for larger experiments. For small experiments, however, the suggested approximations are quite crude and the tests quite conservative. For such small or sparse designs, it is recommended to calculate (via computer or by hand) the exact distribution of the statistic considered for hypothesis testing.
BIBLIOGRAPHY


ACKNOWLEDGEMENTS

Thanks to Dr. H.A. David, who has taught me more about scholarship than he can possibly know, and certainly far more than is apparent in this paper. His diplomatic criticism of my work and his tolerance of my outside interests have also been greatly appreciated.

Thanks to committee members Ken Heimes, Wolfgang Kliemann, Glen Meeheden, Bill Meeker, Shashikala Sukhatme, and especially to Richard Groeneveld for a particularly thorough reading and some helpful comments.

Thanks also to my parents, for obvious reasons.

Special thanks to Dr. Nuwan Nanayakkara, whose companionship has made my stay in Ames a delight.

This work was supported in part by the U.S. Army Research Office, fund number DAAL 03-89-K-0010.
APPENDIX A. COVARIANCE OF TWO SCORES: GENERAL CASE

Recall from (3.9) that

\[ s_i = \sum_j^{(i)} \left[ (m_j + 1)(p_{ij} - \frac{1}{2}) + \sum_{k \neq i}^{(-i,j)} (p_{jk} - \frac{1}{2}) \right]. \]

Then, for \( i \neq j \), we can see that the covariance of the scores for two objects \( C_i \) and \( C_j \) is

\[
\text{cov}(s_i, s_j) = \text{cov} \left( \sum_i^{(i)} (m_k + 1)p_{ik}, \sum_l^{(j)} (m_l + 1)p_{jl} \right) \quad (A.1)
\]

\[
+ \text{cov} \left( \sum_k^{(i)} (m_k + 1)p_{ik}, \sum_l^{(j)} \sum_b^{(-j,l)} p_{lb} \right) \quad (A.2)
\]

\[
+ \text{cov} \left( \sum_k^{(i)} (-i,k) p_{ka}, \sum_l^{(j)} (m_l + 1)p_{jl} \right) \quad (A.3)
\]

\[
+ \text{cov} \left( \sum_k^{(i)} (-i,k) p_{ka}, \sum_l^{(j)} \sum_b^{(-j,l)} p_{lb} \right). \quad (A.4)
\]

The first term (A.1) is simply

\[
\sum_k^{(i)} \sum_l^{(j)} (m_k + 1)(m_l + 1) \text{cov}(p_{ik}, p_{jl}) = \begin{cases} 
-(m_j + 1)(m_i + 1) \frac{\pi_{ij} \pi_{ji}}{n_{ij}}, & n_{ij} > 0 \\
0 & n_{ij} = 0,
\end{cases}
\]
since $p_{ik}$ and $p_{jl}$ are uncorrelated unless $k = j$ and $l = i$. Similarly, (A.2) is seen to be

$$
(i)(j)(-j,l) \sum_k \sum_l \sum_{b \neq j} (m_k + 1) \text{cov}(p_{ik}, p_{lb}) =
$$

$$
= \left\{
\begin{array}{ll}
(i) & \sum (m_k + 1) \frac{\pi_{ik}\pi_{ki}}{n_{ik}} n_{ij} > 0, n_{jk} = 0 \\
(i) & -\sum_k (m_k + 1) \frac{\pi_{ik}\pi_{ki}}{n_{ik}} n_{ij} = 0, n_{jk} > 0 \\
0 & \text{otherwise}
\end{array}
\right.
$$

$$
(i,-j) \sum (m_k + 1) \frac{\pi_{ik}\pi_{ki}}{n_{ik}} n_{ij} > 0
$$

$$
= \left\{
\begin{array}{ll}
(i,-j) & \sum (m_k + 1) \frac{\pi_{ik}\pi_{ki}}{n_{ik}} n_{ij} > 0 \\
(i,j) & -\sum_k (m_k + 1) \frac{\pi_{ik}\pi_{ki}}{n_{ik}} n_{ij} = 0
\end{array}
\right.
$$

The third term (A.3) follows by symmetry from (A.2), and the fourth term (A.4) is

$$
(i)(-i,k)(j)(-j,l) \sum_k \sum_{a \neq i} \sum_l \sum_{b \neq j} \text{cov}(p_{ka}, p_{lb}) =
$$

$$
= \left\{
\begin{array}{ll}
(i)(-i,k) & \sum_k \sum_{a \neq i,j} \frac{\pi_{ka}\pi_{ak}}{n_{ka}} n_{ja} = 0, n_{ka} > 0, n_{jk} > 0 \\
(i) & \sum (m_k + 1) \frac{\pi_{kl}\pi_{lk}}{n_{kl}} n_{il} = 0, n_{kl} > 0, n_{jk} = 0 \\
0 & \text{otherwise}
\end{array}
\right.
$$

$$
= \sum_k (m_k + 1) \frac{\pi_{kl}\pi_{lk}}{n_{kl}} - \sum_{k \neq j} \sum_{l \neq i,k} \frac{\pi_{kl}\pi_{lk}}{n_{kl}}.
$$

Putting these expressions together, we get, for $n_{ij} > 0$,

$$
\text{cov}(s_i, s_j) = -(m_i + 1)(m_j + 1) \frac{\pi_{ij}\pi_{ji}}{n_{ij}}
$$
and for $n_{ij} = 0$, 

$$
\text{cov}(s_i, s_j) = - \sum_k (m_k + 1) \left( \frac{\pi_{ik} \pi_{ki}}{n_{ik}} + \frac{\pi_{jk} \pi_{kj}}{n_{jk}} \right) \\
+ \sum_k \sum_{l \neq i, j} \frac{\pi_{kl} \pi_{lk}}{n_{kl}} - \sum_{k \neq j} \sum_{l \neq i} \frac{\pi_{kl} \pi_{lk}}{n_{kl}},
$$

as reported in (3.18) and (3.19).
APPENDIX B. METHOD FOR FINDING $\tilde{\Sigma}^{-\frac{1}{2}}$

One way to find $\tilde{\Sigma}^{-\frac{1}{2}}$ is through the LR decomposition of $\tilde{\Sigma}$ (see, e.g., Anderson (1984)). Starting with

$$\Sigma^{(1)} = \tilde{\Sigma} = ((\tilde{\sigma}_{ij})),$$

define recursively

$$\Sigma^{(g)} = ((\tilde{\sigma}_{ij}^{(g)})) = F^{(g-1)}\Sigma^{(g-1)}, \quad g = 2, 3, \ldots, t - 1,$$

where $F^{(g-1)} = (f_{ij}^{(g-1)})$ is chosen to annihilate the entries in the $(g - 1)^{th}$ column of $\Sigma^{(g)}$ below the diagonal, while leaving all other sub-diagonal elements of the first $g - 2$ columns unchanged. One such $F^{(g-1)}$ has elements

$$
\begin{cases}
  f_{ii}^{(g-1)} = 1, & i = 1(1)t - 1 \\
  f_{i,g-1}^{(g-1)} = \frac{-\tilde{\sigma}_{i,g-1}^{(g-1)}}{\tilde{\sigma}_{g-1,g-1}^{(g-1)}}, & i = g(1)t - 1 \\
  f_{ij}^{(g-1)} = 0, & \text{otherwise}.
\end{cases}
$$

Note that $F^{(g-1)}$ is lower triangular, and would be equal to $I_{t-1}$ but for the sub-diagonal elements of the $(g - 1)^{th}$ column. So $F = F^{(t-2)}F^{(t-3)} \ldots F^{(1)}$ is also lower triangular. Moreover, $\Sigma^{(g)} = F\tilde{\Sigma}$ is now upper triangular and $D = F\tilde{\Sigma}F'$ diagonal. Now simply take $\tilde{\Sigma}^{-\frac{1}{2}} = D^{-\frac{1}{2}}F$. 
APPENDIX C. EXPRESSION FOR $Q$ FROM A GROUP DIVISIBLE DESIGN

Suppose we have a group divisible design with $m$ groups of size $a = t/m$. Let

\begin{align*}
  v_0 &= \text{var}(s_i) \\
  c_0 &= \text{cov}(s_i, s_j), \quad n_{ij} = 0 \\
  c_1 &= \text{cov}(s_i, s_j), \quad n_{ij} = 1.
\end{align*} \tag{C.1}

Note that $c_0$ is the covariance for the scores of two objects in the same group, and $c_1$ is the covariance for the scores of two objects in different groups. So within each group the scores have covariance matrix

\[
\begin{bmatrix}
  v_0 & c_0 & \cdots & c_0 \\
  c_0 & v_0 & \cdots & c_0 \\
  \vdots & \vdots & \ddots & \vdots \\
  c_0 & c_0 & \cdots & v_0
\end{bmatrix} = (v_0 - c_0)I_a + c_0J_a, \tag{C.2}
\]

and between groups we have covariance matrix $c_1J_a$. Setting $v = v_0/c_1$ and $c = c_0/c_1$, we can write the matrix in (C.2) as $c_1\Sigma_1$, where $\Sigma_1 = (v - c)I_a + cJ_a$. Then
the covariance matrix of all the scores can be written as \( \Sigma = c_1 \Sigma_{1:m} \), where

\[
\Sigma_{1:m} = \begin{bmatrix}
\Sigma_1 & J_a & \cdots & J_a \\
J_a & \Sigma_1 & \cdots & J_a \\
\vdots & \vdots & \ddots & \vdots \\
J_a & J_a & \cdots & \Sigma_1
\end{bmatrix}
\]

\[= \mathbf{I}_m \otimes (\Sigma_1 - J_a) + \mathbf{J}_m \otimes J_a,
\]

where \( \otimes \) denotes the direct (Kronecker) product. Since \( \Sigma \) is singular, we again focus on \( \hat{\Sigma} \), the covariance of the first \( t - 1 \) scores, which can be written as

\[
\hat{\Sigma} = c_1 \begin{bmatrix}
\Sigma_{1:m-1} & J_{a(m-1),a-1} \\
J_{a-1,a(m-1)} & \hat{\Sigma}_m
\end{bmatrix},
\]

where \( c_1 \Sigma_{1:m-1} \) is the covariance matrix of the first \( m - 1 \) groups of \( a \) scores, and \( c_1 \hat{\Sigma}_m \) is the covariance matrix of the \( m^{th} \) group without the last object.

We ultimately seek \( \hat{\Sigma}^{-1} \). Noting from (C.3) that \( \hat{\Sigma} \) is of the form

\[
c_1 \begin{bmatrix}
E & F \\
F' & G
\end{bmatrix},
\]

we can express its inverse as

\[
\hat{\Sigma}^{-1} = \frac{1}{c_1} \left\{ \begin{bmatrix}
E^{-1} & 0 \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
-E^{-1}F \\
0
\end{bmatrix} \begin{bmatrix}
G - F'E^{-1}F \\
I
\end{bmatrix}^{-1} \begin{bmatrix}
-F'E^{-1} \\
I
\end{bmatrix} \right\}.
\]

(C.4)

So we first attack

\[
E^{-1} = \hat{\Sigma}_{1:m-1}^{-1} = (\mathbf{I}_{m-1} \otimes (\Sigma_1 - J_a) + \mathbf{J}_{m-1} \otimes J_a)^{-1}.
\]

(C.5)
We already know from (3.39) how to invert linear combinations of *scalar* products of $I$ and $J$; is there an analogous expression for inverting linear combinations of *direct* products of $I$ and $J$? In other words, we want to find matrices $A$ and $B$ such that

$$I_{m-1} \otimes I_a = \sum_{1:m-1} \sum_{1:m-1}^{-1} \left(I_{m-1} \otimes (\Sigma_1 - J_a) + J_{m-1} \otimes J_a\right) \left(I_{m-1} \otimes A + J_{m-1} \otimes B\right)$$

$$= I_{m-1} \otimes (\Sigma_1 - J_a)A + J_{m-1} \otimes [(\Sigma_1 + (m - 2)J_a)B + J_aA].$$

For this to hold, we obviously require

$$(\Sigma_1 - J_a)A = I_a \quad \text{(C.7)}$$

$$(\Sigma_1 + (m - 2)J_a)B + J_aA = 0_a. \quad \text{(C.8)}$$

Using (3.39) we see at once from (C.7) that

$$A = (\Sigma_1 - J_a)^{-1}$$

$$= \left[(v - c)I_a + (c - 1)J_a\right]^{-1}$$

$$= \frac{1}{v - c} \left(I_a - \frac{c - 1}{(v - c) + a(c - 1)} J_a\right). \quad \text{(C.9)}$$

As for the second requirement (C.8), note that

$$J_aA = \frac{a}{v - c} \left(J_a - \frac{a(c - 1)}{(v - c) + a(c - 1)} J_a\right)$$

$$= \frac{1}{(v - c) + a(c - 1)} J_a$$

$$= b_1 J_a, \quad \text{(C.10)}$$

say, and

$$(\Sigma_1 + (m - 2)J_a)^{-1} = [(v - c)I_a + (c + m - 2)J_a]^{-1}$$

$$= \frac{1}{v - c} \left(I_a - \frac{c + m - 2}{(v - c) + a(c + m - 2)} J_a\right). \quad \text{(C.11)}$$
applying (3.39) again. Putting (C.10) and (C.11) back into (C.8) then gives us

\[
B = -(\Sigma_1 + (m-2)J_a)^{-1}J_aA
\]

\[
= \frac{-b_1}{v-c} \left( J_a - \frac{a(c+m-2)}{(v-c) + a(c+m-2)}J_a \right)
\]

\[
= \frac{-b_1}{(v-c) + a(c+m-2)}J_a
\]

\[
= b_2J_a,
\]
say. Substituting this for B in (C.6) then gives us

\[
\Sigma_{1:m-1}^{-1} = I_{m-1} \otimes A + J_{m-1} \otimes b_2J_a,
\]

where A is defined in (C.9).

Looking back to (C.4), we see that our next step should be an expression for

\[
E^{-1}F = \Sigma_{1:m-1}^{-1}J_{a(m-1),a-1}.
\]

Writing \(J_{a(m-1),a-1} = 1_{m-1} \otimes J_{a,a-1}\) and noting, as in the derivation in (C.10), that \(AJ_{a,a-1} = b_1J_{a,a-1}\), we have

\[
E^{-1}F = \Sigma_{1:m-1}^{-1}J_{a(m-1),a-1}
\]

\[
= (I_{m-1} \otimes A + J_{m-1} \otimes b_2J_a)(1_{m-1} \otimes J_{a,a-1})
\]

\[
= 1_{m-1} \otimes b_1J_{a,a-1} + (m-1)1_{m-1} \otimes ab_2J_{a,a-1}
\]

\[
= 1_{m-1} \otimes (b_1 + (m-1)ab_2)J_{a,a-1}
\]

\[
= 1_{m-1} \otimes b_3J_{a,a-1},
\]

say. Similarly,

\[
F'F^{-1} = 1_{m-1}' \otimes b_3J_{a-1,a}.
\]

Then

\[
F'E^{-1}F = (1_{m-1}' \otimes b_3J_{a-1,a})(1_{m-1} \otimes J_{a,a-1})
\]
\[= (m - 1) \otimes b_3 J_{a-1}, a J_{a,a-1}\]

\[= ab_3 (m - 1) J_{a-1},\]

giving us

\[
(G - F' F^{-1} F')^{-1} = (\Sigma_m - ab_3 (m - 1) J_{a-1})^{-1}
\]

\[= \left( (v - c) I_{a-1} + (c - ab_3 (m - 1)) J_{a-1}\right)^{-1}\]

\[= \frac{1}{v - c} \left( I_{a-1} - \frac{c - ab_3 (m - 1)}{(v - c) + (a - 1)(c - ab_3 (m - 1))} J_{a-1}\right)\]

\[= \frac{1}{v - c} (I_{a-1} - b_4 J_{a-1}),\]

say. Observe from this and (C.13) that

\[
E^{-1} F (G - F' F^{-1} F')^{-1} = (1_{m-1} \otimes b_3 J_{a,a-1}) \left( \frac{1}{v - c} \otimes (I_{a-1} - b_4 J_{a-1})\right)
\]

\[= 1_{m-1} \otimes \frac{b_3}{v - c} (J_{a,a-1} - (a - 1) b_4 J_{a,a-1})\]

\[= 1_{m-1} \otimes \frac{b_3}{(v - c) + (a - 1)(c - ab_3 (m - 1))} J_{a,a-1}\]

\[= 1_{m-1} \otimes b_5 J_{a,a-1},\]

say; from this and (C.14) we then have

\[
E^{-1} F (G - F' F^{-1} F')^{-1} F' F^{-1} = (1_{m-1} \otimes b_5 J_{a,a-1}) (1_{m-1} \otimes b_3 J_{a-1}, a)
\]

\[= J_{m-1} \otimes (a - 1) b_3 b_5 J_{a}.\]

Using these last three expressions, we can finally construct

\[
\begin{bmatrix}
-E^{-1} F \\
I
\end{bmatrix} \begin{bmatrix}
(G - F' F^{-1} F')^{-1} \\
-F' F^{-1}
\end{bmatrix} = \begin{bmatrix}
J_{m-1} \otimes (a - 1) b_3 b_5 J_{a} \\
1_{m-1} \otimes b_5 J_{a,a-1}
\end{bmatrix} \begin{bmatrix}
1_{m-1} \otimes b_5 J_{a,a-1} \\
1_{m-1} \otimes b_5 J_{a,a-1}
\end{bmatrix}(C.15)
We need only add this to
\[
\begin{bmatrix}
E^{-1} & 0 \\
0 & 0
\end{bmatrix}
\]
to get \(c_1\hat{\Sigma}^{-1}\) in (C.4). Recall from (C.9) and (C.12) that
\[
E^{-1} = I_{m-1} \otimes \frac{1}{v-c}(I_a - b_1(c-1)J_a) + J_{m-1} \otimes b_2J_a.
\]

Putting this and (C.15) back into (C.4) we have at last
\[
c_1\hat{\Sigma}^{-1} = \frac{1}{v-c}I_{t-1} + \begin{bmatrix}
I_{m-1} \otimes \frac{-b_1(c-1)}{v-c}J_a & 0 & a(m-1),a-1 \\
0 & 0 & 0
\end{bmatrix}
\]
\[
+ \begin{bmatrix}
J_{m-1} \otimes (b_2 + (a-1)b_3b_5)J_a & -b_51_{m-1} \otimes J_{a,a-1} \\
-b_51_{m-1} \otimes J_{a-1,a} & \frac{b_4}{v-c}J_{a-1}
\end{bmatrix}
\]
\[
= b_6I_{t-1} + \begin{bmatrix}
b_7I_{m-1} \otimes J_a & 0 \\
0 & 0
\end{bmatrix}
\]
\[
+ \begin{bmatrix}
b_8J_{m-1} \otimes J_a & -b_51_{m-1} \otimes J_{a,a-1} \\
-b_51_{m-1} \otimes J_{a-1,a} & b_9J_{a-1}
\end{bmatrix}
\]
\[
= A_1 + A_2 + A_3,
\]
(C.16)
say. Our test statistic \(Q_{GD}\) from (3.37) is then just
\[
Q_{GD} = \frac{1}{c_1} (\hat{s}'A_1\hat{s} + \hat{s}'A_2\hat{s} + \hat{s}'A_3\hat{s}).
\]
(C.17)

Clearly,
\[
\hat{s}'A_1\hat{s} = b_6 \sum_{i=1}^{t-1} s_i^2.
\]
(C.18)

To simplify the other two terms of (C.17), define \(S_g\) as the sum of scores in the \(g^{th}\) group (for \(g = 1(1)m\)), and let \(\hat{s}_m = S_m - s_t\) be the score for the \(m^{th}\) group without
the last object. Then, using \( A_2 \) and \( A_3 \) from (C.16), we have

\[
\begin{align*}
\hat{s}' A_2 \hat{s} &= b_7 \sum_{g=1}^{m-1} S_g^2, \\
\hat{s}' A_3 \hat{s} &= \hat{s}' \\
&= \hat{s}' \left[ \left( \frac{m-1}{b_8} \sum_{g=1}^{m-1} S_g - b_5 \hat{s}_m \right) \frac{1}{a(m-1)} \right] \\
&= \left( b_8 \sum_{g=1}^{m-1} S_g - b_5 \hat{s}_m \right) \sum_{g=1}^{m-1} S_g \\
&\quad + \left( -b_5 \sum_{g=1}^{m-1} S_g + b_9 \hat{s}_m \right) \hat{s}_m. \\
\end{align*}
\]

But \( \sum_{g=1}^{m} S_g = \sum_{i=1}^{t} s_i = 0 \) implies that \( \sum_{g=1}^{m-1} S_g = -S_m \), so that (C.20) becomes

\[
\begin{align*}
&\quad [-b_8 S_m - b_5 (S_m - s_t)] (-S_m) + [b_5 S_m + b_9 (S_m - s_t)] (S_m - s_t) \\
&= (2b_5 + b_8 + b_9) S_m^2 - 2a(-b_5 + b_9) S_m s_t + b_9 s_t^2 \\
&= b_6 s_t^2 + b_7 S_m^2 + R, \\
\end{align*}
\]

where the remainder R can be shown to be zero. Putting the three pieces (C.18), (C.19), and (C.22) back into (C.17), we get

\[
Q_{GD} = \frac{b_6}{c_1} \sum_{i=1}^{t} s_t^2 + \frac{b_7}{c_1} \sum_{g=1}^{m} S_g^2 \\
= k_0 \sum_{i=1}^{t} s_t^2 + k_1 \sum_{g=1}^{m} S_g^2, \\
\]

say.
Switching now to double subscripts, so that \( s_{gi} \) is the score for the \( i^{th} \) object of the \( g^{th} \) group, we can express \( Q_{GD} \) from (C.23) as

\[
Q_{GD} = k_0 \sum_{g=1}^{m} \sum_{i=1}^{a} s_{gi}^2 + k_1 \sum_{g=1}^{m} S_g^2, \tag{C.24}
\]

Noting that

\[
\sum_{i=1}^{a} (s_{gi} - \frac{1}{a} S_g)^2 = \sum_{i=1}^{a} s_{gi}^2 - \frac{1}{a} S_g^2,
\]

we could also write (C.24) as

\[
Q_{GD} = k_0 \sum_{g=1}^{m} \sum_{i=1}^{a} (s_{gi} - \frac{1}{a} S_g)^2 + k_2 \cdot \frac{1}{a} \sum_{g=1}^{m} S_g^2, \tag{C.25}
\]

where

\[
k_2 = k_0 + ak_1. \tag{C.26}
\]

The constants from (C.24) and (C.26) are seen to be

\[
k_0 = \frac{b_6}{c_1} = \frac{1}{c_1(v-c)} = \frac{1}{v_0 - c_0},
\]

\[
k_1 = \frac{b_7}{c_1} = \frac{-b_1(c-1)}{c_1(v-c)} = \frac{-(c_0 - c_1)}{(v_0 - c_0)((v_0 - c_0) + a(c_0 - c_1))}, \tag{C.27}
\]

\[
k_2 = \frac{1}{(v_0 - c_0) + a(c_0 - c_1)}.
\]

For these group divisible designs, we see from section (3.?) that the variances and covariances from (C.1) are

\[
v_0 = \frac{1}{4}(t-a)(t-a)^2 + 2t - a
\]

\[
c_0 = -\frac{1}{4}(t-a)(2t - 3a + 4)
\]

\[
c_1 = -\frac{1}{4}(t - 2a + 2)^2.
\]
Putting these expressions for \( v_0, c_0, \) and \( c_1 \) back into (C.27) we get

\[
k_0 = \frac{4}{(t - a)(t - a + 2)^2}
\]

\[
k_1 = \frac{4[t(t - a) - (a - 2)^2]}{t(t - a)(t - a + 2)^2(t - 2a + 2)^2}
\]

\[
k_2 = \frac{4}{t(t - 2a + 2)^2}.
\]

(C.28)

Adopting some obvious notation from the analysis of variance,

\[
SST = \sum_{g=1}^{m} \sum_{i=1}^{a} s_{gi}^2
\]

\[
SSB = \frac{1}{a} \sum_{g=1}^{m} S_g^2
\]

\[
SSW = \sum_{g=1}^{m} \sum_{i=1}^{a} (s_{gi} - \frac{1}{a} S_g)^2,
\]

(C.29)

our two expressions for \( Q_{GD} \) from (C.24) and (C.25) become

\[
Q_{GD} = k_0 SST + k_1 SSB
\]

\[
= k_0 SSW + k_2 SSB,
\]

(C.30)

where the constants and sums of squares are defined by (C.28) and (C.29), respectively.
APPENDIX D. RELATIONSHIP BETWEEN $Q$ AND $C$ FOR PBIB RANK DESIGNS

Suppose we have a PBIB rank design with $b = a^2$ blocks of $k = m$ objects, reducing to a group divisible paired-comparison design with $m$ groups of $a$ objects. We would like to compare Prentice’s $C = \tilde{y}'\tilde{\Sigma}_y^{-1}\tilde{y}$ with $Q = \tilde{s}'\tilde{\Sigma}^{-1}\tilde{s}$. Let us first invert $\tilde{\Sigma}_y$, the covariance matrix of the $t-1$ chosen scores $\{y_{gi}\}$, where $y_{gi}$ denotes the score for the $i^{th}$ object in the $g^{th}$ group (for $i = 1(1)a$ and $g = 1(1)m$). From the expressions given in §3 of Prentice (1979) we see that

$$\text{var}(y_{gi}) = \frac{t-a}{12(m+1)}$$

$$\text{cov}(y_{gi}, y_{hj}) = \begin{cases} 
-1/12(m+1) & g \neq h \\
0 & g = h.
\end{cases}$$

Then $\Sigma_y$, the covariance of all the scores, can be expressed as

$$\Sigma_y = \frac{1}{12(m+1)} \begin{bmatrix}
(t-a)I_a & -J_a & \cdots & -J_a \\
-J_a & (t-a)I_a & \cdots & -J_a \\
\vdots & \vdots & \ddots & \vdots \\
-J_a & -J_a & \cdots & (t-a)I_a
\end{bmatrix}
= \frac{1}{12(m+1)} \{I_m \otimes [(t-a)I_a + J_a] - J_m \otimes J_a \}.$$
Through a process similar to inversion of $\tilde{\Sigma}$ for group divisible designs, found in Appendix C, we see that

$$
\frac{1}{12(m+1)}\tilde{\Sigma}^{-1} = \frac{1}{t-a}I_{t-1} - \begin{bmatrix}
I_{m-1} \otimes \frac{1}{t(t-a)}Ja & 0_{a(m-1),a-1} \\
0_{a-1,a(m-1)} & 0_{a-1}
\end{bmatrix}
+ \begin{bmatrix}
J_{m-1} \otimes b_1Ja & 1_{m-1} \otimes b_2Ja,a-1 \\
1'_{m-1} \otimes b_2Ja_{-1,a} & b_3Ja_{-1}
\end{bmatrix}
= A_1 - A_2 + A_3,
$$

(D.1)
say, where the constants in $A_3$ depend only on $t$ and $a$. Clearly

$$
\tilde{y}'A_1\tilde{y} = \frac{1}{t-a} \left( \sum_{g=1}^{m} \sum_{i=1}^{a} y_{gi}^2 - 2yma \right)
$$

$$
\tilde{y}'A_2\tilde{y} = \frac{1}{t(t-a)} \sum_{g=1}^{m-1} Y_g^2,
$$

where $Y_g$ is the sum of scores for the $g^{th}$ group. It can be shown that

$$
\tilde{y}'A_3\tilde{y} = \frac{1}{t-a}y_{ma}^2 - \frac{1}{t(t-a)}Y_m^2,
$$

so that (D.1) gives us

$$
C_{PBJ} = \tilde{y}'\tilde{\Sigma}^{-1}\tilde{y} = \frac{12(m+1)}{t-a} \left( \sum_{g=1}^{m} \sum_{i=1}^{a} y_{gi}^2 - \frac{1}{t} \sum_{g=1}^{m} Y_g^2 \right)
$$

(D.2)

We can use the familiar equality

$$
\sum_{i=1}^{a} (y_{gi} - \frac{1}{a}Y_g)^2 = \sum_{i=1}^{a} y_{gi}^2 - \frac{1}{a}Y_g^2,
$$
in (D.2) to express $C_{PBIB}$ as

$$C_{PBIB} = \frac{12(m+1)}{t-a} \left[ \sum_{g=1}^{m} \sum_{i=1}^{a} (y_{gi} - \frac{1}{a} Y_g)^2 + \left( \frac{1}{a} - \frac{1}{t} \right) \sum_{g=1}^{m} Y_g^2 \right]$$

$$= \frac{12(m+1)}{t-a} \left[ SSW_y + \frac{t-a}{t} SSB_y \right], \quad (D.3)$$

where $SSW_y$ and $SSB_y$ denote, respectively, the sum of squares Within and Between groups for the scores $\{y_{gi}\}$.

Now we would like to express Prentice's $\{y_{gi}\}$, the sums of standardized ranks, in terms of the simple ranks $\{r_{j(gi)}\}$, where $r_{j(gi)}$ is the rank in block $j$ for the $i^{th}$ object of the $g^{th}$ group. From (1.11) we see that

$$y_{gi} = \sum_{j} \left( \frac{r_{j(gi)}}{k_j} - \frac{1}{2} \right)$$

$$= \frac{1}{m+1} \sum_{j} \left( r_{j(gi)} - \frac{m+1}{2} \right)$$

$$= \frac{1}{m+1} \left( r(gi) - \frac{a(m+1)}{2} \right)$$

$$= \frac{1}{m+1} \left( r(gi) - \frac{t+a}{2} \right), \quad (D.4)$$

where $r(gi)$ is the rank-sum for $C_{gi}$, the $i^{th}$ object of the $g^{th}$ group, and $\sum_{j}^{[gi]}$ denotes the sum over all blocks $j$ in which $C_{gi}$ is ranked. Writing $R_g = \sum_{i=1}^{a} r_{j(gi)}$ as the sum of these rank-sums for the $g^{th}$ group, we see from (D.4) that

$$Y_g = \frac{1}{m+1} \left( R_g - \frac{a(t+a)}{2} \right). \quad (D.5)$$

We can then use (D.4) and (D.5) to express the sums of squares from (D.3) as

$$SSB_y = \frac{1}{a} \sum_{g=1}^{m} Y_g^2.$$
where $SSW_r$ and $SSB_r$ are the analogous sums of squares for the rank-sums $\{r_{(gi)}\}$. Putting these back into (D.3) then gives

\[
CPBIB = \frac{12}{(m+1)(t-a)} \left( SSW_r + \frac{t-a}{t} SSB_r \right). \tag{D.6}
\]

We now aim toward an expression for $Q_{P\text{BIB}}$ in terms of $SSB_r$ and $SSW_r$. First note that, since $C_{(gi)}$ appears in $m$ of the $b$ blocks, we have the following simple relation between its row-sum and its rank-sum:

\[
a_{(gi)} = r_{(gi)} - m,
\]

where $a_{(gi)}$ is the total number of preferences in favor of $C_{gi}$. Applying this to (3.14) we can express $C_{gi}$'s paired-comparison score as

\[
s_{(gi)} = (t-a+2)[r_{(gi)} - m - \frac{1}{2}(t-a)] - \sum_{k=1}^{a} [r_{(gk)} - m - \frac{1}{2}(t-a)]
\]

\[
= (t-a+2)[r_{(gi)} - m - \frac{1}{2}(t-a)] - \left( R_g - t - \frac{a(t-a)}{2} \right), \tag{D.7}
\]

so that the $g^{th}$ group score is then

\[
S_g = (t-a+2) \left( R_g - t - \frac{a(t-a)}{2} \right) - a \left( R_g - t - \frac{a(t-a)}{2} \right) \tag{D.8}
\]
From (D.7) and (D.8) we see that the difference between a score and its group average is just
\[ s_{(gi)} - \frac{1}{a} S_g = (t - a + 2) \left( r_{(gi)} - \frac{1}{a} R_g \right). \]

From this and (D.9) the sums of squares for the paired-comparison scores are now seen to be
\[
SS_{B_s} = (t - 2a + 2)^2 SS_B_r \tag{D.10}
\]
\[
SS_{W_s} = (t - a + 2)^2 SS_W_r. \tag{D.11}
\]

Looking back to the constants in (3.49) we can express \( Q_{GD} \) from (3.48) as
\[
Q_{GD} = \frac{4}{t - a} SS_W_r + \frac{4}{t} SS_B_r
\]
\[
= \frac{4}{t - a} \left( SS_W_r + \frac{t - a}{t} SS_B_r \right). \]

Comparing this to (D.6) we can at last state that
\[
Q_{GD} = \frac{m + 1}{3} C_{PBIB}.
\]