Chapter 1

Introduction

1.1 Overview

Comparative judgement is a process employed for selection and elimination of items with an objective to make the best choice or learn about the rank-order of the choices. With contexts as varied as economic competitions, sport tournaments, psychological experiments, preference learning algorithms, and animal behaviour experiments, the interest of both academicians and practitioners in comparative experiments is ubiquitous. The items involved in these comparisons can be employees, players, teams, stimuli, consumer products and many more. The comparative experiments performed to realise any particular objective result in a ranked list of items and are amenable for statistical analysis.

Comparative ordinal data originate when objects or items are compared by a subject, judge or responder. The responder in these experiments can be an agent, a person or a consumer among many and we denote by subject the person or the group who makes the choice. There are also situations when the comparison can be performed without the involvement of a human-subject. A sports tournament or an animal behaviour experiment is an example of such a comparison. In such experiments, nature makes the choice. The models developed in this thesis are applicable to data with or without the involvement of the judge.

When items are compared by people as subjects, it becomes easier if the items are presented in pairs. In the literature, such comparisons are known as paired comparisons. Thurstonian and Bradley–Terry models are the most popular models for analysis of paired comparison data. Both the models postulate that the comparison between any two items is independent Bernoulli experiment with probability of success dependent on fixed latent characteristic of items involved in the comparison. The latent characteristic

of the items is referred to as strength or ability of the items being compared. The ease of estimation of model parameters for Bradley–Terry model and its generalization to incorporate ties, multiple comparison and item or subject specific covariates has made it very popular among statisticians.

In a paired comparative experiment, if all the items are compared to each other, the resulting structure of comparison is known as round-robin or full comparison. Such a design of comparison of items is relevant if the objective is to rank-order all the items being compared. However, for a situation when the objective is to select only few items, the experimenter would prefer elimination design of comparisons. For example, in a single-elimination structure, one item is eliminated after every comparison. A practitioner can also choose to compare the objects in a mixture of designs. These structures of comparison are also referred to as tournaments in this thesis. Here, we do not study the adequacy of a particular design of tournament to achieve the objective of the designer. Instead, our objective is to use the comparison data arising out of repetition of a specific tournament and come up with models to infer about the strength of all the items compared.

In many situations, an experimenter does not realise his or her entire objective through a comparative judgement experiment. The choice of the tournament structure reflects the confidence of the experimenter in his or her prior judgement about the rank and strength of the objects being compared. For example, a totally ignorant experimenter would like to conduct a round-robin tournament. On the other hand, if an experimenter is partially confident of his or her assessment of the rank order, a rank-dependent structure of the tournament such as standard knockout can be employed. In this regard, this thesis contributes to the development of probability models for the rank-consistent strength of the compared items.

The comparative judgement models are tools to infer about the probability of an item being preferred over others. Modern models dependent on the latent strength of the items characterises the probability as a function of the latent characteristic of the compared items. In practice, many of the market, industrial, sports and political tournaments are open to gambling markets in countries such as UK and USA and Australia. The market behaviour observed in terms of betting odds in gambling market hands us yet another tool to measure the aggregate preference of an item over the other. As a part of the secondary contribution in this thesis, we have tried to link the two estimates of the probability of the preference of the compared items and draw behavioural insights and also make better predictions of the outcomes in a tournament.

The models described in this thesis can also be generalised to comparative rank data when more than two items are compared at a time. Here we make a comparative study of the varying units of comparison and also illustrate the methodology to estimate the strength of the items from a multiple comparison data.

1.2 Paired Comparison Models

For the purpose of this thesis, we redefine a general class of paired comparison models as follows. A paired comparison experiment comprises K items that are to be compared in pairs. Whenever i-th item is compared with j-th item, the probability that an item is preferred over other (here, i over j) is denoted as p_{ij}. Noether [54] postulated the existence of a chance variable X_{ij} given as

$$X_{ij} = d_i - d_j + \epsilon_{ij}, \tag{1.1}$$

where d_i is a fixed parameter associated with i-th item, ϵ_{ij} is the error term associated with a comparison. The equation was also interpreted by authors as the amount of preference given to one item over the other. The distribution assumption made on the error component is that it is symmetric about zero. Noether [54] further divided the equation on preference comparison to a sensation from each item given as

$$X_i = d_i + \epsilon_i, \tag{1.2}$$

where ϵ_i are identically distributed and $X_{ij} = X_i - X_j$. Under these assumptions, Noether [54] studied the sensitivity of the parameter estimation from the observed data, to the parametric assumption on the distribution of ϵ_i. Based on stochastic assumption, the two classical models that were proposed are Thurstone model [65] with normal errors and (SBT) Bradley–Terry model [10] with double exponential errors. Probably because of the ease of calculation and the scope of generalisation, the Gumbel error model or the Bradley–Terry model has been more popular among academicians. We discuss in Chapter 5, how SBT models have been generalised to different kinds of paired comparisons and multiple comparisons. According to the SBT model, the probability of an item-i being preferred over item-j is given by

$$P_{ij} = P[X_{ij} > 0] = P[\epsilon_{ij} > d_j - d_i] = \frac{\exp(d_i)}{\exp(d_i) + \exp(d_j)} = \frac{s_i}{s_i + s_j}, \tag{1.3}$$

where $s_i = \exp(d_i)$. Hereafter, in this study, we refer to s_i as the strength of the i-th item.

1.3 Related Literature

Comparative judgement in form of paired comparisons (PC) has always been a topic of interest for statisticians and practitioners (psychologists and survey designers) alike. The literature compendium by Davidson and Farquhar [20], a monograph written by David [18] and the more recent study by Cattelan [13] provides a clear indication of the immense amount of interest that the paired comparison models have gathered from various domains of application. Development of the analysis has always been hinged on the various models that have been devised to infer from the data generated from paired comparison experiments. Although most of the analysis is based on the Bradley–Terry or Thurstone models, details of which can be found in Thurstone [65], Bradley and Terry [10], many alternatives such as Stern [62, 63], Abbas and Aslam [1], Davidson and Beaver [19] and extensions of models to situations such as ties, continuum, categorical responses to paired comparison are among many that have also been proposed [4, 9, 41]. Many of the recent papers focus on categorization of the situation or domain of study when one model can be preferred [14, 53].

The design of the experiment that results in the PC data has resulted in domain and structure specific inferences based on the model. Proper extension of the model to capture the design in the models has also been addressed. A section of the literature deals with experiments where judges/subjects are involved in creating preference data for the objects. Such data result in a rich structure of inconsistent comparisons. For applying Bradley–Terry models to such scenarios, the model is typically re-framed as log-linear models [22] or written as logit models to incorporate various covariates (subject/object specific) [21, 28, 67]. The covariates are modeled to influence a typical estimate of the Bradley–Terry model. The particularity of such model lies in the fact that the estimates need not be a surrogate for the true abilities [65]. Instead, they measure how various categories of the subjects perceive the abilities to be. Bradley–Terry models have also been used to analyse paired comparison matrix different from this thesis and similar to the PC matrix in analytical hierarchy process (AHP) [29] .

A typical section of literature particularly relevant to our study is paired comparisons with no involvement of subjects. A representative domain of such a section is sports. A tournament played between teams or individuals is an example of paired comparison design. Much of the initial literature is based on chess but the different structure

of the game such as racing where more than one team competes at a time or tennis where a typical game has more than one subgames to be won. The literature contains analysis of win–loss records in racquetball, tennis, football, basketball and racing and even innovation contests [6, 36, 38, 39, 41, 35, 64, 69]. Different questions targeted are ranking of various players, temporal evolution of strength as an attempt towards prediction, effectiveness of the structure of game, an optimal design of the tournament and whether the subjective rankings imparted are good indicators [8, 15, 16, 32, 31, 35].

Yet another section of the literature concentrates on design of the experiments and uses sports for the validation of the model in context of a particular design and related significant inferences that can be drawn from such designs. Win–loss records (Bradley–Terry and Thurstone) and point tally models (Poisson and Gaussian) have historically been used to model the wins of a typical object(team/individual) in a match.Annis and Craig [6], Annis [5] have generalised the two separate models as a special case of the hybrid model which is capable of differentiating between close wins and convincing wins. It provides a measurement for the degree of the victory. Annis and Davis [7] typically talks about the influence of a structure of the tournament in the estimates of the abilities of the objects. The fact that not all designs of the tournament are framed after proper statistical analysis and are decided subjectively by a panel of experts of the game provides us with problems of framing valid models to estimate true abilities of the objects from a particular design and structure of the tournament.

One of the most popular paired comparison models has been the Bradley–Terry model. A generalised class of algorithms referred to as MM algorithm allows a fast computation of the maximum likelihood (ML) estimate of the various extensions of the BT models using an iterative procedure. A tutorial on the same is provided by Hunter and Lange [44].

Recently, Bayesian estimation of the strength of the participants in a comparative judgement experiment has also become popular among statisticians. Adams [2] used WinBUGS to set up a Bayesian estimation framework for the BT models. Gormley et al. [34] proposed a Bayesian estimation of a mixture model based on the Bradley–Terry–Luce model. While Adams [2] used built-in MH algorithm to approximate the target distribution, Gormley et al. [34] designed a proposal distribution for the MH algorithm. Guiver and Snelson [37] have proposed an alternative approximation of the posterior distribution using expectation propagation (EP) algorithm. More recently, Caron and Doucet [12] introduced a set of latent variables to generalise the MM algorithm under a Bayesian setting and also framed data augmentation sampler based on the latent variables. The authors show that such a method can be easily

applied to all the existing generalisation of the Bradley–Terry models. In this context, we discuss methodology for addressing inference problems for Bradley–Terry model applied to the paired comparison data from repeated designs.

1.4 Arrangement of Thesis

This thesis is divided into five main chapters. In chapter 2, we develop rank consistent extensions of Bradley–Terry models. The model has been applied to the outcome from National Collegiate Athletic Association. The main contribution of this study is the application of rank-consistent Bradley–Terry models for comparative judgement outcomes from the repetition of a fixed tournament structure. Parameter estimation, goodness-of-fit using suitably framed test statistic and its null distribution, change point analysis in a nested model framework, as well as other estimation aspects are discussed in this chapter.

In Chapter 3, we elaborate on inference problems associated with the strength estimates of the paired comparison models for the data from repetition of a single structure of the tournament. We have developed a general mechanism to estimate confidence intervals and perform testing of hypothesis based on the parameters of the paired comparison models. The framework is general and applicable to small, moderate and large sample size. We have also developed novel bootstrap technique for resampling based inferences.

In Chapter 4, we develop Bayesian framework for rank consistent models for the paired comparison data. In this chapter, we discus how we can build a prior distribution over the space of strength parameter and make use of Bayesian procedures to update our belief. We also compare the forecasts under various models.

In Chapter 5, we develop a model to predict the outcome of a comparison in the tournament based on the rank of the items and third-party assessment of the preference probability. In our context, betting odds set by the bookmakers serve as third-party assessment and tournament is a sports tournament with paired comparison between players. The model is based on Bradley–Terry framework where the participating players are linked by a measure of their competitive ability. We illustrate the application of our model with a data set comprising records from an international tennis tournament for women and men. Bayesian approach has been adopted to make inferences about the parameters in the model. The estimates allow us to infer about the degree by which a bookmaker skews the 'true-odds'. Predictions based on the estimated model is compared with true observation and various strategies of selecting bets based on the

model have been discussed. We propose two very promising betting strategies that have yielded positive results albeit in the short run.

In Chapter 6, we demonstrate the application of strength parameter based probability models to partial-ranked data arising from multiple comparisons. The models discussed is a generalisation of the paired comparison models discussed at length in this thesis. We also discuss the the efficiency of the multiple comparison model estimates in the presence of size-dependent noise. We have made a comparative study of situations when the size of comparison is allowed to vary.