University of Southampton

Faculty of Human, Social and Mathematical Sciences

Mathematical Sciences

# An Application of HodgeRank to Predicting the Outcome of Competitive Events: A Case Study of Horse Racing

by

Conrad D'Souza

Thesis for the degree of Doctor of Philosophy

January 2020

**Abstract**

Financial markets rely heavily on the informational efficiency of their participants, setting security prices by aggregating the decisions made by these participants. Errors in the decision making process can pervade financial markets, creating opportunities for savvy traders to exploit and achieve excess risk-adjusted returns.

Systematic errors arise when the complexity of a decision problem exceeds the capabilities of an individual who, in turn, employs heuristics to reduce the complexity of the task. These so called cognitive biases have been observed in human behaviour affecting attitudes to risk, reliance on certain information sources, and the veracity of judgements.

Inconsistent data, containing conflicting information, increases the complexity of decision problems and the likelihood that individuals will make sub-optimal decisions. This project studies the impact on decision making and market efficiency of inconsistent ranking data, data used for ranking purposes containing intransitive patterns of preferences.

HodgeRank, a topologically-inspired ranking algorithm, is used to understand and explore inconsistent ranking data (Jiang et al. 2011). This technique separates pairwise comparison matrices into consistent and inconsistent components, and derives a ranking solution from the consistent ranking data. This project extends the algorithm to account for (i) the reliability of the information contained in the data and (ii) information contained in the inconsistent ranking data.

A study of parimutuel horserace wagering markets is undertaken to establish whether inconsistent ranking data is fully accounted for in real world settings. The results of a statistical and economic evaluation demonstrate that the presence of inconsistent ranking data reduces the quality of decisions made by bettors and that market inefficiencies exist as a result. HodgeRank, in conjunction with conditional logit models and Kelly wagering strategies, is capable of exploiting this inefficiency and achieving abnormal returns.

# Contents

# List of Figures



# List of Tables

# Declaration of Authorship

I, Conrad D'Souza, declare that this thesis titled, **An Application of HodgeRank to Predicting the Outcome of Competitive Events: A Case Study of Horse Racing** and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- None of this work has been published.

Signed:

---

Date:

---

# Acknowledgements

Firstly, I would like to thank my supervisors Dr. Ruben Sanchez-Garcia, Dr. Tiejun Ma, Prof. Johnnie E. V. Johnson and Prof. Ming-Chien Sung, for their support over the years. Their expertise has pushed me in new directions and their guidance has ensured I was never lost.

In particular I would like to express my deepest gratitude to Ruben who has been a source of boundless enthusiasm, inspiration and patience throughout.

I would like to thank my family for their support, in particular my mother whose unwavering belief has been of great comfort.

I would like to thank Jonathan St Clair Terry who has been my sounding board over the years and kept me level-headed.

I would like to thank Jo Munson, Jess Ogden, Becki Nash, Charlie Carson, Briony Gray and Eva Walsh who have provided much needed advice, support, housing, and most importantly friendship.

I would like to thank Fabio Strazzeri and James Strudwick for their comments and ideas over the years.

I would like to thank the Web Science Centre for Doctoral Training for the opportunity to undertake a PhD and the diverse training I recieved.

Lastly, I would like to dedicate this thesis to Laura Hyrjak whose company is sorely missed.

# 1 Introduction

Financial markets are traditionally understood through models in which agents are 'rational'. According to Barberis & Thaler (2003), these rational agents immediately update their beliefs in line with a Bayesian interpretation of Bayes' theorem and make decisions which maximise their subjective expected utility (Savage 1972, Von Neumann & Morgenstern 1944).

These models of rational behaviour provide an appealingly intuitive understanding of financial markets. Market values, which are determined and adjusted by the collective behaviour of participants in the market, immediately and accurately incorporate available information. If a security is under- or over-priced, rational agents have an incentive to trade until the market value agrees with the intrinsic value of the security.

The Efficient Market Hypothesis is the theory that, in financial markets, "prices always 'fully reflect' available information" (Fama 1969)[p. 383]. This does not require that individual agents are rational but rather that the market as a whole behaves in a rational fashion, although clearly models assuming rational behaviour will also satisfy the Efficient Market Hypothesis. In an efficient market, no participant can consistently outperform the market and generate excess risk-adjusted returns (Fama 1976). That is not to say that abnormal returns cannot be achieved over shorted time frames but that they should only be possible over longer periods by accepting greater levels of risk or being privy to information that is not widely available.

Three forms of the Efficient Market Hypothesis are commonly employed, differing in their definitions of 'available information' (Jensen 1978):

(i) **Weak:** Prices fully reflect past price movements

(ii) **Semi-strong:** Prices fully reflect all publicly available information

(iii) **Strong:** Prices fully reflect all information, whether publicly or privately available

In a weak-form efficient market, an individual cannot outperform the market by analysing historic prices alone. In a semi-strong form efficient market, abnormal returns can be generated with inside information however this

advantage is nullified in a strong form efficient market where it is already accounted for in the market values.

Efficient markets aggregate information disseminated amongst their participants to accurately price securities. Prediction markets, capitalising on the presumed efficiency of markets, employ this mechanism to determine the probability of an event occurring. Traders in these markets wager on the probability of an event occurring and the market allocates "market values to make predictions about specific future events" (Berg & Rietz 2003)[p. 79]. If a contract, paying £10 if it rains tomorrow and nothing otherwise, is being traded at £1, the market has determined that the probability of rain tomorrow is 10%.

Prediction markets have been used for forecasting trends and outcomes including sales volumes (Chen & Plott 2002, Hopman 2007), initial public offerings (Cowgill et al. 2009, Berg et al. 2009) and elections and geopolitical events (Wolfers & Zitzewitz 2004, Forsythe et al. 1992, Berg et al. 2008, Wolfers & Leigh 2002, Atanasov et al. 2017). Evidence of their effectiveness has led to prediction markets increasingly influencing decision and policy making (Healy et al. 2010, Cowgill et al. 2009).

Despite the theoretical merit and widespread application of the Efficient Market Hypothesis, evidence has emerged of inefficiencies in markets. Studies have identified cases where there have been delays updating prices and information has not been fully captured by the market (Schwert 2003, Barberis & Xiong 2009, Imas 2016, Basu 1977). These inefficiencies, whilst not widespread (Jensen 1978), are evidence that the presumption of rational behaviour in markets does not always hold.

Deviations from rationality are a well documented phenomenon in psychology as models governing the behaviour of rational decision makers have been found lacking (Simon 1979, Arthur 1994, Smith & von Winterfeldt 2004, Doyle 1999*a*, Doyle 1999*b*). One prominent example of irrational decision making is an experiment by Tversky & Kahneman (1981) in which subjects were asked if they would rather:

(i) gain $240 or risk gaining $1000 with a 75% chance of gaining nothing

(ii) lose $750 or risk losing $1000 with a 25% chance of losing nothing

The majority of subjects were risk averse in a financial gain context, accepting the certain gain, and risk seeking in a loss context, gambling for the chance to lose nothing. Both choices, in different ways, demonstrated that subjects failed to maximise their subjective expected utility and make decisions in a rational manner.

Non-rational decision making arises from failures to use information effectively in maximising subjective expected utility. The complexity of a decision problem is affected by the number of possible outcomes that require consideration ('alternative-based complexity') (Timmermans 1993, Payne et al. 1993) and the ease with which differences between them can be discerned ('attribute-based complexity') (Sung et al. 2009). Attribute-based complexity increases as the relationships between attributes become more complicated (Klein & Yadav 1989, Sung & Johnson 2007) and the outcomes become fundamentally more similar (Biggs et al. 1985, Bockenholt et al. 1991).

Increased levels of complexity place additional computational demands on the cognitive resources of the decision maker (Ballou & Pazer 1985, Wand & Wang 1996). When the computational demands of a task approach or exceed the limits of the cognitive resources available to the individual, individuals experience cognitive strain. If a task persists in inducing cognitive strain, maintaining focus on the task depletes a limited mental resource and leaves the individual more susceptible to giving up on the task (Baumeister et al. 1998, Sweller 1988). This was illustrated in a study by Frederick (2005) which found that only 17% of participants answered the following three questions correctly:

(i) A bat and ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?

(ii) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

(iii) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take fo the patch to cover half of the lake?

The complexity of this task induced cognitive strain in the participants, resulting in them providing intuitive but incorrect answers.

In an attempt to alleviate cognitive strain, individuals employ simplifying heuristics to reduce the complexity of the task and lower its computational demands to manageable levels (Tversky & Kahneman 1974, Slovic & Lichtenstein 1971, Cosmides & Tooby 1994, Gigerenzer & Gaissmaier 2011). Although heuristics are useful and often necessary in judgement and decision making, these simplifications can be imperfect and introduce cognitive biases, systematic errors in reasoning, into the process (Allais 1953, Kahneman & Tversky 1984, Tversky & Kahneman 1986, Prelec & Loewenstein 1991, Schumpeter 1976).

Numerous cognitive biases have been identified since Kahneman & Tversky (1973) demonstrated that human behaviour can deviate from the normative standards of rationality including:

i) **Availability**

The availability heuristic estimates frequencies and likelihoods of events "by the ease with which instances or associations could be brought to mind" (Tversky & Kahneman 1973)[p. 208]. The easier it is to recall instances of an event, the more likely the event is deemed to be.

Whilst availability simplifies the process, the ease with which instances can be recalled is not necessarily indicative of the likelihood of an event and can be affected by factors such as vividness and recency (Schwarz et al. 1991, Nisbett & Ross 1980). Does the letter K appear more frequently in the first or third position of a word? It is certainly easier to recall words beginning with K than words with K appearing in another position. Individuals believe that words are more likely to begin with the letters K,L,N,R and V than have those letters in their third position however the opposite is true (Tversky & Kahneman 1973).

ii) **Representativeness** This heuristic estimates the likelihood of an event by the "degree to which it (a) is similar in essential characteristics to its parent population; and (b) reflects the salient features of the process by which it is generated." (Kahneman & Tversky 1972)[p. 431]. An event is considered more likely if it is closer in similarity to the average or expected event.

Individuals applying the representativeness heuristic often fail to account for variability in samples from a population. The sequence of coin tosses H-T-H-T-H-T is thought more likely to occur than the sequences H-H-H-T-T-T and H-H-H-H-T-H, even though all three sequences are equally likely (Tversky & Kahneman 1974).

Kahneman & Tversky (1973) explored the existence and failings of the representativeness heuristic by presenting participants with a description of Tom W., a fictional first-year graduate student, and asking them to estimate the likelihood he studied one of nine courses.

> Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

The experiment showed that participants over-estimated the likelihood of Tom W. studying a course when he fit the image of the stereotypical student for that course, not accounting for the face that some courses are more popular than others and a random student is more likely to study a more populated course.

iii) **Framing Effects** Framing effects are a cognitive bias where individuals respond "differently to different but objectively equivalent descriptions of the same problem" (Levin et al. 1998)[p. 150].

In an experiment establishing the existence of framing effects (Tversky & Kahneman 1981), subjects were presented with a fictional epidemic scenario and two responses to it.

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed:
>
> (a) If Program A is adopted, 200 people will be saved
>
> (b) If Program B is adopted, there is a 1 in 3 chance that 600 people will be saved and a 2 in 3 chance that no-one will be saved

A separate group were given the same problem with the responses framed in terms of lives lost instead of saved. Although both decision problems were objectively equivalent, each groups' responses demonstrated that "choices involving gains are often risk averse and choices involving loses are often risk taking" (Tversky & Kahneman 1981)[p. 453].

Although cognitive biases have been observed in human decision making, their origin remains somewhat contentious. Whilst Tversky & Kahneman (1974)[p. 1124] believe that cognitive biases are "severe and systematic errors" indicative of failings of human decision making, evolutionary psychologists claim that the "human mind is not worse that rational.. but may often be better than rational" (Cosmides & Tooby 1994)[p. 329]. Gigerenzer & Gaissmaier (2011) argue that rationality is only a fair benchmark for decision making in a limitless environment where the aim is to maximise subjective expected utility. Heuristics are not error-prone design flaws but features, aware of time and information processing restraints, which enable people to solve natural adaptive problems reliably (Cosmides & Tooby 1994).

Regardless of their construction, cognitive biases are observed occurrences of non-rational behaviours in decision making (Haselton et al. 2005). These biases affect individual decision making but are also known to pervade financial markets with favourite long shot biases (Ali 1977, Snyder 1978, Thaler & Ziemba 1988) and disposition effects (Suhonen & Saastamoinen 2018, Barberis & Xiong 2009, Andrikogiannopoulou & Papakonstantinou 2018, Imas 2016) having been identified in market prices. Cognitive biases are therefore not only systematic errors in information usage, but sources of potential market inefficiencies.

Data quality is an ongoing concern for agents operating in financial markets, affecting their ability to use information effectively (Redman 1996, English 1999). Wand & Wang (1996) categorised facets of data quality into external dimensions, relating to the suitability of a dataset for a given task, and internal dimensions, task-independent assessments of the intrinsic accuracy of a data.

The intrinsic accuracy of a dataset is comprised of several dimensions, reflecting issues which can arise during the production of data (Ballou & Pazer 1985, Wand & Wang 1996). Concerns with any of these dimensions indicate that the dataset represents a different real-world state from the one intended, limiting the usefulness of the data.

(i) **Precision:** How precisely the information is recorded. Imprecise data represents an incorrect world state

(ii) **Consistency:** Level of agreement amongst the data. Inconsistent data is representative of multiple world states

(iii) **Objectivity:** How unbiased, unprejudiced and impartial the information is. Subjective data represents a perceived world state rather than the real-world state

(iv) **Completeness:** Amount of information recorded. Missing entries prevent the dataset from fully representing a world state

(v) **Timeliness:** How quickly changes to a real-world state are recorded. Delayed data represents a previous real-world state

Poor quality data is not "fit for use by data consumers" (Wang & Strong 1996)[p. 6] and impedes them from making good decisions. Accounting for data quality increases the attribute-based complexity of decision problems, with inaccurate data often removed (Allison 2001) or replaced by estimates (Rubin 2004, Schafer 1999, Sterne et al. 2009).

Conflicting data is often produced by uncertain events where there is no clear real-world state and repetitions of an event can lead to different outcomes. It is also a common feature of decision problems where the possible choices can be ranked rather than individually evaluated (Tversky & Kahneman 1974). Ranking larger sets of alternatives in a coherent fashion is a more complex task and is often simplified by considering pairs of alternatives in isolation (Saaty 1990). In doing so, intransitive patterns of preferences can emerge where alternative A is preferred to B and B is preferred to C, yet C is preferred to A (Tversky 1969).

Intransitive patterns of preferences are a deviation from normative models which can lead to individuals acting as 'money-pumps'. Given the above pattern of preferences, it is reasonable to assume that an individual would pay to replace alternative C by B. Similarly they would pay to replace B by A and A by C. The net result is that the individual has paid a sum of money for no gain, replacing the original alternative by itself.

Ranking data, data on pairs of alternatives that is used to rank the alternatives themselves, which is produced from uncertain events can often be inconsistent, containing intransitive patterns of preferences. Such inconsistent ranking data promotes a tendency in individuals to avoid making judgements, deferring the task (Tversky & Kahneman 1992) or accepting the status quo (Luce 1998). Where they do form judgements in the presence of conflicting data, individuals often exhibit a confirmation bias, rejecting information which disagrees with their pre-established internalised mental models (Nickerson 1998, Jonas et al. 2001), or express uncertainty in their judgements and later contradict them (Fischer et al. 2000).

Participants in financial markets which surround uncertain events, in particular prediction markets, are faced with the difficulty of making decisions from inconsistent ranking data containing intransitive patterns of preferences.

It has been shown that individuals struggle with this type of data and therefore there is a very real risk that these markets violate the Efficient Market Hypothesis and that prices fail to capture all of the information available in datasets containing inconsistent ranking data.

# 2 Research Problem

Inconsistent ranking data poses a risk to the efficiency of financial markets, increasing the complexity of decision problems and the likelihood that information will not be entirely utilised for maximal expected utility by individuals. This project assesses whether this risk materialises in real world experiences, with inconsistent ranking data leading to poorer decision making and inaccurate market prices.

**RQ1**: Does the presence of inconsistent ranking data prevent decision makers from fully utilising the available consistent information?

**RQ2**: Do decision makers capture all of the information contained within inconsistent ranking data?

**RQ3**: Is there an economic cost to inconsistent ranking data? Do semi-strong form market inefficiencies exist as a result?

## 2.1 Evaluating Market Efficiency

If there is a discrepancy between the market price and intrinsic value of a security, there is an opportunity for risk-free gains and market participants have an incentive to trade. This incentive lasts until the discrepancy is resolved and the market price reflects the intrinsic value of the security.

In efficient financial markets, prices fully reflect all available information and should align with intrinsic values, although this may not always be the case. Differences between market prices and intrinsic values occur when information is inaccessible, however they disappear when this information becomes available.

Prices in an efficient market are the best estimate of the intrinsic value of a security, given the available information, and thus there is no trading strategy which can consistently outperform the market and generate excess risk-adjusted returns (Basu 1977). There is, however, an inherent randomness in market prices, which take time to react to new information and settle on a valuation, and this randomness allows investors to generate excess returns by the sole virtue of being lucky. Nonetheless it is impossible to systematically

outperform an efficient market by utilising the same information available to the market.

Determining whether the efficient market hypothesis holds across different financial markets has been a popular pursuit of economists (Malkiel 2003, Fama 1969). Different methodologies are employed for each form of the efficient market hypothesis, testing necessary conditions for each.

### 2.1.1 Weak Form

A weak efficient market fully accounts for past price movements in the prices it sets. A necessary condition for weak form efficiency is that there is no auto-correlation, correlation between a series and a delayed copy of itself, in the prices series of securities. If auto-correlation did exist, future price movements could be somewhat predicted from historical prices, presenting an opportunity for excess risk-adjusted returns.

Price series which exhibit no level of auto-correlation behave as a random walk, with historical prices having no bearing on future prices. A substantial body of work has tested whether there is auto-correlation in security prices with a range of time lags, concluding that financial markets are weak form efficient and prices series exhibit no auto-correlation of significance (Cootner 1964, Kendall & Hill 1953, Moore 1962, Granger & Morgenstern 1963, Godfrey et al. 1964).

### 2.1.2 Semi-Strong Form

In semi-strong efficient markets, all publicly available information is accounted for, leaving no opportunity for trading strategies to systematically earn excess risk-adjusted returns by using this information more effectively. Tests of semi-strong form efficiency often attempt to outperform the market, using information which is hypothesised to be fully unaccounted for.

Determining what constitutes an excessive risk-adjusted return is a necessary step in semi-strong efficiency tests. There is no unequivocal risk-adjusted return on an investment and experimental returns are instead compared to returns generated by an asset pricing model. In doing so, the experiment

tests the joint hypothesis that the market is semi-strong efficient and that the asset pricing model is appropriate and not solely whether the market is semi-strong efficient (Fama 1976, Timmermann & Granger 2004). Despite this flaw in testing semi-strong market efficiency, the potential for excess risk-adjusted returns is used as evidence for or against the efficiency of financial markets (Jensen 1978).

Basu (1977) showed that price-to-earnings (P/E) ratios, commonly regarded as indicating whether securities are under- or over-priced, were not fully account in market prices. Securities with low P/E ratios produced higher risk-adjusted returns than those with high P/E ratios.

Public announcements of earnings are often followed by anomalous security prices, allowing for "systematic excess returns in post-announcement periods" (Ball 1978)[p. 103]. This phenomenon was confirmed by Watts (1978), even after accounting for the effects of the joint hypothesis problem.

Despite evidence that semi-strong form inefficiencies exist (Smith 1986, Brunnermeier & Nagel 2004, Jensen & Ruback 1983), there is a consensus that financial markets are highly semi-strong efficient and that excess risk-adjusted returns are not systematically achievable (Malkiel 2003). Where semi-strong form inefficiencies do exist, markets often adapt and quash the opportunity for abnormal returns (Schwert 2003).

### 2.1.3 Strong Form

A strong form efficient market is required to account for all information in its prices, regardless of public availability. Strong efficiency is an extension of semi-strong efficiency with the additional requirement that inside information cannot be used to achieve excess risk-adjusted returns.

Strong form efficiency is considered less a reasonable criteria for assessing the performance of financial markets, and more of a theoretical completion of the Efficient Market Hypothesis (Fama 1969). There is limited research into strong market efficiency, with the most notable being a study by Finnerty (1976) showing that known insiders were able to achieve abnormal returns before they were identified.

## 2.2 Project Outline

The project develops a tech broadly consists of three stages:

1. Develop a technique for analysing inconsistent data
    (i) Represent a dataset of pairwise comparisons containing intransitive patterns of preferences
    (ii) Identify consistent and inconsistent ranking data within the dataset
    (iii) Determine the underlying preference for each alternative from the consistent ranking data
    (iv) Extract information from the inconsistent ranking data and redetermine the underlying preferences with regard to this information
2. Evaluate the decision making ability of individuals in the presence of inconsistent ranking data
    (i) Identify an appropriate setting to evaluate decision making
    (ii) Model decision making in this setting
    (iii) Evaluate whether the presence of inconsistent ranking data prevents decision makers from using the consistent data to its fullest
    (iv) Evaluate whether decision makers account for information contained in inconsistent ranking data
3. Identify market inefficiencies resulting from the presence of inconsistent ranking data
    (i) Determine market prices accounting for consistent and inconsistent ranking data
    (ii) Model a strategy for seeking returns from the market
    (iii) Evaluate the potential for excess risk-adjusted returns and the existence of market inefficiencies

HodgeRank, a topologically-inspired ranking algorithm (Jiang et al. 2011), is extended and employed to separate ranking data containing intransitive patterns of preference into consistent and inconsistent ranking data, extract information from the dataset as a whole, and measure the underlying preferences for each alternative.

A case study is undertaken to assess how well individuals make decisions in the presence of inconsistent ranking data. Following established methods for testing the efficiency of financial markets, this case study models the expected outcomes of decision makers and assesses whether the model can be improved by the inclusion of information extracted by the HodgeRank algorithm. If the model can be improved, there is evidence that this information has not been fully accounted for in the decisions made by market participants.

Assessing the efficiency of financial markets in regards to inconsistent ranking data requires an appropriate setting which satisfies three requirements:

(i) **Conflicting Data**: Expectation that the available ranking data is significantly inconsistent

(ii) **Quantifiable**: Decisions made by market participants can be modelled

(iii) **Verification**: Accuracy of market prices in agreeing with intrinsic values can be established

# 3 Topology

The techniques employed throughout this project to separate, understand and exploit consistent and inconsistent parts of a dataset are derived from HodgeRank, a topological ranking algorithm. This section introduces the concepts and theories required to understand the topological underpinnings of HodgeRank, however the reader is directed to Hatcher (2001) for a thorough treatment of topology.

## 3.1 Simplicial Complex

Topology is a field of mathematics which studies properties of shapes and spaces that are invariant under continuous deformations such as stretching and twisting. These shapes are often represented by discrete (hence computationally tractable) structures called simplicial complexes, encoding the 'mesh' of the shape. Before we can define a simplicial complex, we have to introduce the notion of affine independence.

**Definition 1.** *Elements* $v_0, v_1, ..., v_n \in \mathbb{R}^m$ *are* ***affinely independent*** *if the vectors* $v_1 - v_0, v_2 - v_0, ..., v_n - v_0 \in \mathbb{R}^m$ *are linearly independent.*

Affine independence is the concept that a set of points in $\mathbb{R}^m$ are linearly dependent but the direction vectors from any fixed point in the set to any other point in it are linearly independent. Thus if you move any $v_i$ to the origin in $\mathbb{R}^m$, the remaining elements will be linearly independent.

To motivate an understanding of this definition, consider the case where three vectors, $0 \neq v_0, v_1, v_2 \in \mathbb{R}^3$ form a triangle with $v_1 = 2v_0$. Obviously these vectors are not linearly independent, however it is clear that they are affinely independent since the two direction vectors from any fixed $v_i$ to the remaining $v_j$ must be linearly dependent or else the three vectors form a line instead of a triangle.

**Definition 2.** *Let $v_0, v_1, ..., v_n$ be a set of affinely independent vectors in $\mathbb{R}^m$. An **n-simplex** $\sigma$ is a subset of $\mathbb{R}^m$ defined as:*

$$\sigma = \left\{ \sum_{i=0}^{n} \lambda_i v_i \mid \sum_{i=0}^{n} \lambda_i = 1 \text{ and } \lambda_i \geq 0, \lambda_i \in \mathbb{R}, \forall i = 0, ..., n \right\}$$

*and the **vertices** of $\sigma$ are $v_0, ..., v_n$.*

An $n$-simplex over the affinely independent $n+1$ vertices $v_0, v_1, ..., v_n$ can be alternatively defined as the convex hull of the $n+1$ vertices. Intuitively, the $n$-simplex is the smallest $n$-dimensional object in Euclidean space which contains every segment between every pair of vertices.

**Example 1.** *A 0-simplex is a point. A 1-simplex is the line-segment from $v_0$ to $v_1$.*

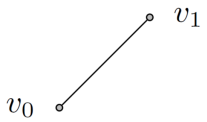


Figure 1: 1-simplex

*A 2-simplex is the triangle spanned by $v_0, v_1, v_2$.*

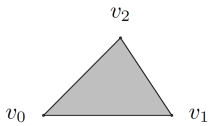


Figure 2: 2-simplex

*A 3-simplex is the solid tetrahedron spanned by $v_0, v_1, v_2, v_3$.*

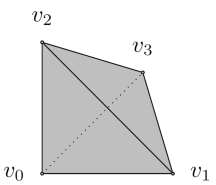


Figure 3: 3-simplex

By taking a subset of the vertices of a $k$-simplex, we can form a lower dimensional simplex which is contained in the $k$-simplex. Such a simplex is referred to as a **face** of the $k$-simplex. Formally, a face $\tau$ of a $k$-simplex is the $m$-simplex formed from $m+1$ of the $m \leq k$ vertices written as $\tau \subseteq \sigma$.

The 2-simplex above is bounded by edges. As these are 1-simplices, these edges are faces of the 2-simplex. Similarly the 3-simplex has triangles and edges as faces since these are 2-simplices and 1-simplices respectively.

**Definition 3.** *A finite collection $K$ of simplices in $\mathbb{R}^n$ is called a* ***simplicial complex*** *if the following conditions hold:*

1. $\sigma \in K$ *and* $\tau \subseteq \sigma \implies \tau \in K$

2. $\sigma, \tau \in K$ *and* $\sigma \cap \tau \neq \emptyset \implies \sigma \cap \tau \subseteq \sigma$ *and* $\sigma \cap \tau \subseteq \tau$

The first part of this definition is a transitivity condition stating that any face of a simplex in $K$ is also itself a simplex of $K$. A simplicial complex contains all the faces of the simplices it is a collection of. Secondly, the definition states that the intersection of two simplices is a simplex in $K$ and is a face of both simplices. The following example illustrates this point.

**Example 2.** *The first object is a simplicial complex whilst the second is not as the intersection of its two 2-simplices is not simplex.*

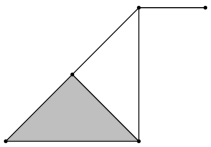

Figure 4: Simplicial complex

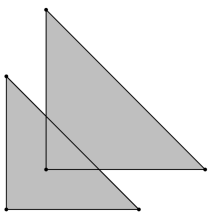

Figure 5: Not a simplicial complex

Essentially an $n$-simplex is the $n$-dimensional analogue of the triangle and a simplicial complex is a combinatorial representation of a space made by 'gluing' simplices along faces.

## 3.2 Orientation

A $(k-1)$-simplex has $k$ vertices which can be ordered in $k!$ different ways, describing the order to traverse the vertices. A permutation alters the ordering of the vertices and the set of all permutations of the $k$ vertices is denoted by $S_k$. The reader is directed to Dixon & Mortimer (1996) for a detailed understanding of permutations.

It is a standard result from group theory that every permutation can be written in terms of transpositions. A permutation is referred to as *odd* if it is the product of an odd number of transpositions or *even* otherwise and for a permutation $\sigma$,

$$\operatorname{sgn}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is an even permutation,} \\ -1 & \text{otherwise} \end{cases}$$

The decomposition of a permutation into transpositions is not unique; however, the parity is unaffected by the choice of decomposition.

**Example 3.** *Consider the set* $\{1,2,3,4\}$ *and two permutations* $\sigma, \tau \in S_k$ *which permute the set to* $\{1,3,4,2\}$ *and* $\{1,2,4,3\}$*, in this order, respectively.* $\sigma$ *can be expressed as transposing the elements* 2 *and* 3 *before transposing the elements* 2 *and* 4 *and so* $\sigma$ *is an even permutation.* $\tau$ *transposes* 3 *and* 4 *and so* $\tau$ *is an odd permutation. Alternatively,* $\tau$ *can be expressed as transposing the elements* 1 *and* 3*, transposing* 3 *and* 4 *and finally transposing* 1 *and* 4*.*

An orientation of the $(k-1)$-simplex is the equivalence class of orderings of the $k$ where two orderings are equivalent if one is an even permutation of the other. Since every permutation is either even or odd, there are precisely two orientations of any simplex of dimension greater than zero, although the precise ordering of the vertices can vary. In Example 3, the two orderings

$\{1, 2, 3, 4\}$ and $\{1, 3, 4, 2\}$ are in the same equivalence class whilst the ordering $\{1, 2, 4, 3\}$ is in the opposite equivalence class.

**Example 4.** *Orientation is an intuitive notion for simplices of dimension less than 3.*

- 0*-simplices are points and only have one orientation*
- *Orientated* 1*-simplices are directed edges:*


Figure 6: Orientations of a 1-simplex

- 2*-simplices are orientated by rotational direction (clockwise or anti-clockwise)*

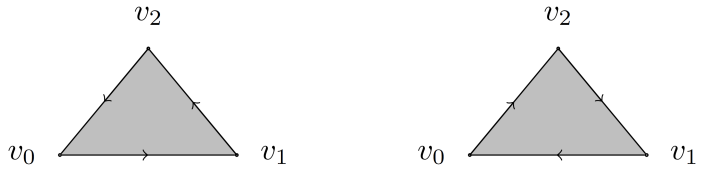


Figure 7: Orientations of a 2-simplex

A $k$-simplex of the simplicial complex $K$, with vertices ordered $v_0 < v_1 < ... < v_{k-1}$, is denoted by $[v_0, v_1, ..., v_k]$. An ordering of the vertices of $k$ induces a natural orientation on each $k$-simplex, given by $[v_{i_0}, v_{i_1}, ..., v_{i_{k-1}}]$ with $i_0 < i_1 < ... < i_{k-1}$, and a natural indexing of the set of $k$-simplices is given by sorting the simplices coordinate-wise in an ascending order. Throughout the project, this natural orientation and indexing will be used.

## 3.3 Cohomology

Cohomology is the dualization of homology and is most often defined in this way. In this section we will introduce cohomology distinctly from homology although an understanding of homology is beneficial. The reader is directed

to Hatcher (2001), from which the section is informed, for a thorough explanation of simplicial homology and cohomology.

In cohomology, functions act on oriented simplices. Thus it is important to fix an orientation of the simplices and to remain consistent with this orientation throughout. We will order the vertices $v_0, v_1, ..., v_n$ of a simplicial complex by their index so $v_0 < v_1 < ... < v_n$ which induces an ordering on all simplices contained within the complex.

**Definition 4.** *Let $K$ be a simplicial complex and $\Sigma_k$ be the corresponding set of oriented $k$-simplices. A **$k$-cochain**, $f$, is a real-valued function on the vertices of the $k$-simplices whose value alternates sign according to the orientation of the vertices, that is,*

$$f([v_{\sigma_{i_0}}, v_{\sigma_{i_1}}, ..., v_{\sigma i_k}]) = \text{sgn}(\sigma) f([v_{i_0}, v_{i_1}, ..., v_{i_k}])$$

*for all $k$-simplices $[v_{i_0}, v_{i_1}, ..., v_{i_k}] \in \Sigma_k$ and all $\sigma \in S_{k+1}$, the permutation group of the $k+1$ vertices.*

The set of all $k$-cochains is denoted by $C^k(K, \mathbb{R})$ although we will abuse notation and abbreviate it to $C^k$. Each $C^{k \geq 0}$ is a vector space over $\mathbb{R}$ and is, in particular, an abelian group.

A $k$-cochain can also be considered a real-valued function on the $k$-simplices themselves and so 0-cochains are **vertex functions**, 1-cochains are **(directed) edge functions**, 2-cochains are oriented triangle functions etc. For example a 2-cochain, $f_2 \in C^2$ assigns a real number to each 2-simplex $[v_i, v_j, v_k]$ such that

$$\begin{aligned} f_2([v_i, v_j, v_k]) &= f_2([v_j, v_k, v_i]) = f_2([v_k, v_i, v_j]) \\ &= -f_2([v_i, v_k, v_j]) = -f_2([v_k, v_j, v_i]) = -f_2([v_j, v_i, v_k]) \end{aligned}$$

as $\{j, k, i\}$ and $\{k, i, j\}$ are even permutations of $\{i, j, k\}$ whilst the remaining permutations are odd.

Following the natural oredering of the $k$-simplices (explained above), we can pointwise identify $f \in C^k$ with a vector $u \in \mathbb{R}^m$ by $u_i = f(m_i)$ where $m_i$

is the $i$-th $k$-simplex (Edelsbrunner & Harer 2010). We will abuse notion and refer to $f \in C^k$ both as a $k$-cochain and as the vector it is identified with.

**Example 5.** *Consider the following simplicial complex and cochains. These*

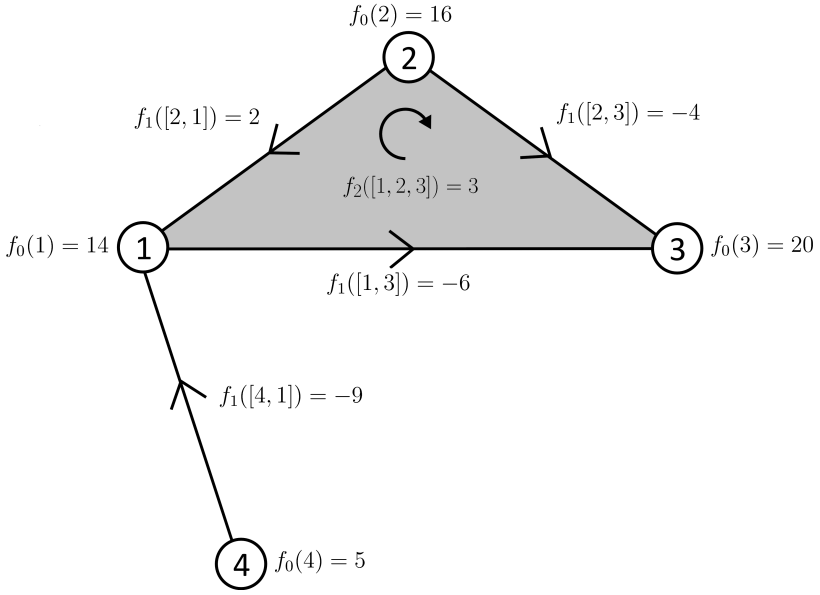


Figure 8: Cochains on a simplicial complex

*cochains can be identified the following vectors*

$$f_0 = \begin{pmatrix} 14 \\ 16 \\ 20 \\ 5 \end{pmatrix}, \; f_1 = \begin{pmatrix} -2 \\ -6 \\ 9 \\ -4 \end{pmatrix}, \; f_2 = \begin{pmatrix} 3 \end{pmatrix}$$

**Definition 5.** *The* ***k-th coboundary operator*** *on the simplicial complex $K$ is the linear map $\delta_k : C^k \to C^{k+1}$ that takes a $k$-cochain, $f \in C^k$, to a $(k+1)$-cochain, $\delta_k f \in C^{k+1}$, defined as*

$$(\delta_k f)([v_0, ..., v_{k+1}]) := \sum_{i=0}^{k+1} (-1)^i f([v_0, ..., v_{i-1}, v_{i+1}, ..., v_{k+1}])$$

The coboundary operators are linear maps, with coefficients in $\mathbb{R}$, connecting the cochain groups of a simplicial complex by extending cochains to

act on higher dimensional simplices.

**Example 6.** *Consider the set of vertices* $\{0, 1, 2\}$ *and let* $f \in C^0$ *be defined by* $f(v_i) = i$*. By applying the* 0*-th coboundary operator, we can derive a* 1*-cochain in* $C^1$ *as*

$$(\delta_0 f)([v_i, v_j]) = j - i$$

*We can also derive a* 2*-cochain by applying the* 1*-th coboundary operator to this* 1*-cochain, giving*

$$\begin{aligned}(\delta_1\delta_0 f)([v_0, v_1, v_2]) &= \sum_{i=0}^{2}(-1)^i(\delta_0 f)([v_0, ..., v_{j-1}, v_{j+1}, ..., v_2]) \\ &= (2-1) - (2-0) + (1-0) \\ &= 0\end{aligned}$$

This holds in general: the composition of two coboundary operators is always zero (Hatcher 2001, Lemma 2.1, p. 105).

**Lemma 1.** *For any* $k \in \mathbb{N}$*,*

$$\delta_{k+1} \circ \delta_k = 0$$

In particular, the image of $\delta_k$ is contained in the kernel of $\delta_{k+1}$.

Coboundary operators are linear maps which have degree 1, mapping from a cochain group to a higher dimensional cochain group. Thus for any simplicial complex $K$ we have the following sequence of cochain groups and coboundary operators

$$0 \xrightarrow{\delta_{-1}} C^0(K, \mathbb{R}) \xrightarrow{\delta_0} C^1(K, \mathbb{R}) \xrightarrow{\delta_1} C^2(K, \mathbb{R}) \xrightarrow{\delta_2} ... \xrightarrow{\delta_{r-1}} C^r(K, \mathbb{R}) \xrightarrow{\delta_r} ...$$

Cochains can be identified with vectors and thus coboundary operators, as linear maps between sets of cochains, can be identified with matrices (after fixing an ordering of the $k$-simplices for each $k$).

A coboundary operator, $\delta_k : C^k \to C^{k+1}$, can be identified with $A \in \mathbb{R}^{n \times m}$ where $m$ is the number of $k$-simplices and $n$ is the number of $(k+1)$-simplices (Edelsbrunner & Harer 2010). The entries of $A$ are given by

- If the $i$-th $k$-simplex is a face of the $j$-th $(k+1)$-simplex, with the $k$-simplex maintaining its orientation in the $(k+1)$-simplex, $A_{ji} = 1$;
- If the $i$-th $k$-simplex is a face of the $j$-th $(k+1)$-simplex, with the $k$-simplex reversing its orientation in the $(k+1)$-simplex, $A_{ji} = -1$;
- If the $i$-th $k$-simplex is not a face of the $j$-th $(k+1)$-simplex, $A_{ji} = 0$.

A 0-simplex is considered to have its orientation preserved if it is the starting vertex of a 1-simplex and reversed if it is the end. Again we will abuse notation and refer to $\delta_k$ as both a coboundary operator and the matrix it is identified with.

**Example 7.** *Consider the simplicial complex and cochains from Example 5.*

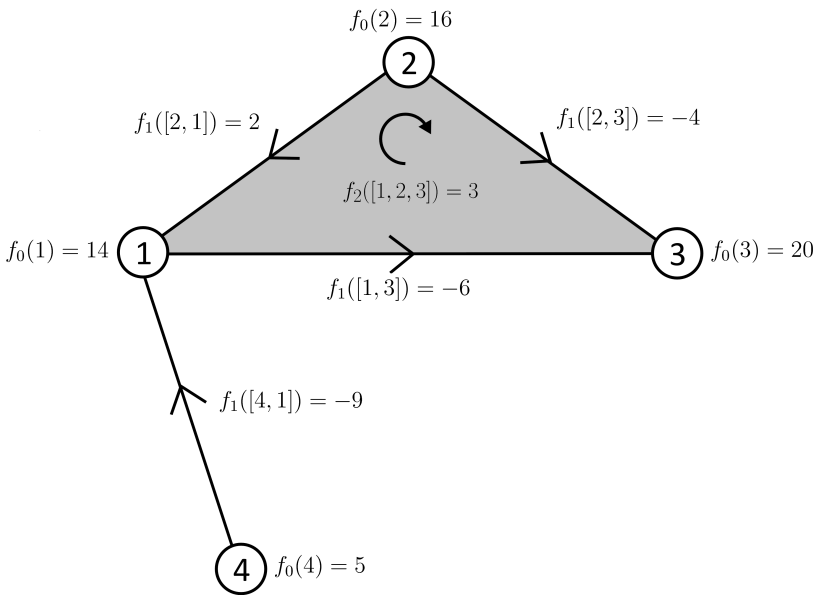


Figure 9: Cochains on a simplicial complex

*The cochains can be identified with vectors and the* 0*-th and* 1*-th cobound-*

*ary operators can be identified with the following matrices*

$$\delta_0 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}, \ \delta_1 = \begin{pmatrix} 1 & -1 & 0 & 1 \end{pmatrix}$$

Given a choice of inner product for each cochain group, which we denote by $\langle \ , \ \rangle_{C^k}$, we can also introduce a linear map from a cochain group to a lower dimensional cochain group.

**Definition 6.** *Given a choice of inner products on each $C^k$, we define the* ***adjoint $\delta_k$ of the coboundary operator*** *$\delta_k : C^{k+1} \to C^k$ as the only linear operator satisfying*

$$\langle \delta_k f_k, g_{k+1} \rangle_{C^{k+1}} = \langle f_k, \delta_k^* g_{k+1} \rangle_{C^k}$$

*for all $f_k \in C^k$ and for all $g_{k+1} \in C^{k+1}$.*

The adjoint of a linear map from one inner product space to another is itself a linear map from the latter inner product space to the former. The adjoint to a coboundary operator exists and is unique (Hatcher 2001). It should be stressed that the adjoint of an operator is not the inverse of the operator but rather the generalisation of conjugate transposes.

Similarly to the coboundary operators, the adjoint operators also form a sequence of cochain groups and linear maps, although these maps have degree $-1$.

$$0 \xleftarrow{\delta_{-1}^*} C^0(K, \mathbb{R}) \xleftarrow{\delta_0^*} C^1(K, \mathbb{R}) \xleftarrow{\delta_1^*} C^2(K, \mathbb{R}) \xleftarrow{\delta_2^*} ... \xleftarrow{\delta_{r-1}^*} C^r(K, \mathbb{R}) \xleftarrow{\delta_r^*} ...$$

An adjoint operator, $\delta_k^* : C^{k+1} \to C^k$, can be identified with a matrix $A^* \in \mathbb{R}^{m \times n}$ where $m$ is the number of $k$-simplices and $n$ is the number of $(k+1)$-simplices (Edelsbrunner & Harer 2010). The entries of this matrix vary according to the inner products attached to the relevant cochain spaces. We will denote both the adjoint operator and its identified matrix by $\delta_k^*$.

Clearly the maps $\delta_k^* \circ \delta_k$ and $\delta_{k-1} \circ \delta_{k-1}^*$ are operators from $C^k$ to itself and we refer to these as the *k-dimensional combinatorial up Laplace operator* and the *k-dimensional combinatorial down Laplace operator* respectively.

**Definition 7.** *Let $K$ be a simplicial complex. The following are linear operators on $C^k(K, \mathbb{R})$:*

*i)* ***k-dimensional up Laplace operator***

$$\Delta_k^{up} = \delta_k^* \circ \delta_k$$

*ii)* ***k-dimensional down Laplace operator***

$$\Delta_k^{down} = \delta_{k-1} \circ \delta_{k-1}^*$$

*iii)* ***k-dimensional combinatorial Laplacian***

$$\Delta_k = \Delta_k^{up} + \Delta_k^{down}$$

The 0-dimensional combinatorial Laplacian is more commonly known as the graph Laplacian and encodes the structure of the underlying graph (restricting the complex to vertices and edges) (Chung 1997). The $k$-dimensional combinatorial Laplacian is a generalisation of the graph Laplacian to higher dimensions.

# 4 HodgeRank

This project addresses the question of whether the presence of inconsistent information in available ranking data affects decision making processes and whether market inefficiencies are created as a result. Following the approach of previous literature in testing semi-strong market efficiency (Basu 1977, Ball 1978, Figlewski 1979, Bolton & Chapman 1986, Snyder 1978, Johnson et al. 2006, Sung et al. 2009), the project attempts to understand inconsistencies in ranking data and exploit this understanding to achieve excess risk-adjusted returns.

In this project a topologically-inspired technique for modelling, separating and exploiting ranking data containing inconsistent preferences for alternatives is developed and employed. The technique is derived from HodgeRank, an algorithm for ranking alternatives from observed pairwise comparisons by modelling the comparisons as a simplicial complex (Jiang et al. 2011).

This section describes the HodgeRank framework and improvements that have been made to it. Unless otherwise stated, theorems in this section are from Jiang et al. (2011) with original proofs.

Several contributions have been made to the HodgeRank algorithm, developing it further and enhancing its ability to exploit information contained within inconsistent ranking data. Although it cannot be said that the available data has been utilised to its fullest, the improvements made to the algorithm have increased the amount of information extracted from the data. This has allowed the project to more thoroughly assess the impact of inconsistent ranking data on decision making and semi-strong market efficiency in real world scenarios.

(i) **Measuring underlying preference**: HodgeRank was conceived as an algorithm for ranking alternatives from pairwise comparisons containing inconsistencies and missing entries, and has been applied as such (Jiang et al. 2011, Xu et al. 2012, Yang et al. 2014). In doing so, however, a more nuanced understanding of the underlying preference for each alternative has been disregarded.

HodgeRank not only provides a ranking of the preference for each alternative, but also captures the degree to which one alternative is preferred over another. This project argues, and successfully demonstrates, that a granular understanding of the underlying preference improves the quality of the output and is of economic significance.

(ii) **Edge weights**: Inconsistent ranking data commonly occurs when information is gathered from a range of sources and there is disagreement between them. These sources may vary in quality with some being regarded as more trustworthy or relevant than others. It is therefore important to capture the perceived quality of information for each pairwise comparison and adjust its contribution to the output accordingly.

In its simplicial complex representation of a pairwise comparison matrix, HodgeRank assumes no prior knowledge of the important of each edge, weighing them all equally. The algorithm presented here has been extended to account for the perceived quality of the information forming each pairwise comparison, producing more credible and valuable outputs from HodgeRank.

(iii) **Measuring inconsistent features**: By nature, alternatives cannot be coherently ranked from intransitive patterns of preferences. HodgeRank identifies transitive patterns of preference in pairwise comparison matrices and separates them from the intransitive patterns, deriving ranking solutions solely from this consistent ranking data.

Although coherent rankings of alternatives cannot be found from intransitive patterns of preferences, their location can provide insights into the relationships between alternatives. In any intransitive pattern of preferences, it may not be the case that every pair of alternatives precludes a ranking but rather a subset of the alternatives. Alternatives found together in many intransitive patterns of preference may possess fundamental characteristics which naturally produce these patterns (such as tennis players whose style of play is strong against certain players but less effective against others). Identifying which subsets of

alternatives cannot be easily ranked provides a more nuanced understanding of the data, and any ranking solution, which can better inform applications.

Some intransitive patterns of preferences may arise not from inherent relationships between alternatives, but from noise in the ranking data. Noisy data should be considered less reliable or informative in any conclusions drawn from the data and therefore it is important to establish whether an intransitive pattern is a feature of the alternatives being ranked or noise in the dataset.

By counting the number of intransitive patterns of preferences that a subset of alternatives are part of, and measuring how far those intransitive patterns are from satisfying the transitive property, the contribution of this subset to overall inconsistency in the ranking data can be measured. If a subset of alternatives contributes significantly to inconsistency in the ranking data, it indicates the inconsistent ranking data they produce is a signal that these alternatives do not admit a clear ranking and is not simply noise in the ranking data.

The value of this measure is established in this project, being applied to weight pairwise comparisons related to historical horse races. Weighting pairwise comparisons in this way, the HodgeRank algorithm is able to extract more valuable information which is then used to improve a statistical model forecasting the outcomes of future races.

## 4.1 Representing a dataset as a simplicial complex

Following Jiang et al. (2011), a pairwise comparison matrix (PCM) is modelled as a simplicial complex. The alternatives are represented as vertices in the complex, forming the 0-skeleton, and each pair of alternatives is connected by a directed edge (1-simplex) if there is a pairwise comparison between them, pointing towards the preferred alternative. The sets of vertices and directed edges are denoted by $V$ and $E$ respectively. Any $k$-tuple of vertices, where every pair of vertices are connected by a directed edge, forms a $(k-1)$-simplex.

Pairwise comparisons are real-valued functions on pairs of alternatives whose value changes sign if the alternatives are flipped. They are equivalent to 1-cochains, directed edge functions, and can similarly be represented as vectors. The vector space $C^1$ contains every possible collection of pairwise comparisons for the pairs of alternatives which have been compared (pairs with non-zero entries in the pairwise comparison matrix). Similarly $C^0$ is the vector space of all possible vertex functions (0-cochains), $C^2$ of all possible oriented triangle functions (2-cochains) and so on.

**Example 8.** *The pairwise comparison matrix*

$$P = \begin{pmatrix} 0 & -3 & 0 & 0 & -5 \\ 3 & 0 & -4 & -6 & 0 \\ 0 & 4 & 0 & -3 & 0 \\ 0 & 6 & 3 & 0 & 2 \\ 5 & 0 & 0 & -2 & 0 \end{pmatrix}$$

*produces the following simplicial complex*

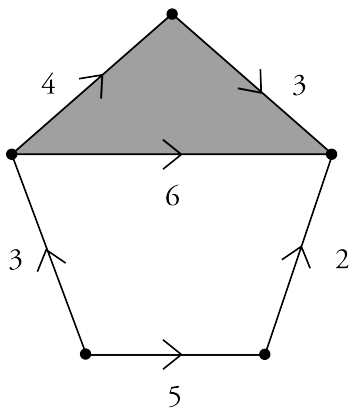


Figure 10: Simplicial complex representation of pairwise comparisons

Removing the higher dimensional features of the complex (simplices above dimension 1), the complex is equivalent to a network representation of the pairwise comparisons. The presence of higher dimensional features presents

opportunities for more detailed analysis of the set of comparisons using topological tools.

## 4.2 Extracting consistent ranking data

One of the biggest advantages offered by a network model of pairwise comparisons is the ability to infer incomplete information from the observed comparisons. Whilst the directed edges provide a direct comparison between pairs of alternatives, every longer path between a pair of alternatives provides an indirect comparison via a series of intermediary nodes. Multiple paths may exist between any pair of vertices, providing different indirect comparisons between the alternatives.

There are different ways to combine observed pairwise comparisons along a path to form an indirect measurement of the relative preference for pairs of alternatives, most commonly by adding or multiplying their values along the path. A teacher grading tests may ask whether one student deserve five marks more than another student. On the other hand, in tennis it is often more useful to question whether one player is twice as good as another. In this project it is assumed that the dataset is comprised of additive pairwise comparisons, noting that multiplicative comparisons can be transformed in additive ones by a logarithm transformation.

In a dataset containing no contradictory information, it is expected that all comparisons between the same pair of alternatives agree with each other. In particular, every indirect comparison should agree with the observed direct comparison. A path $\{v_0, ..., v_k\}$ is said to be **transitive** (with respect to the pairwise comparisons $f \in C^1$) if $f(v_0, v_1) + ... + f(v_{k-1}, v_k) = f(v_0, v_k)$. Wherever a pair of alternatives can be compared both directly and indirectly, a cycle exists in the network formed by joining the two respective paths. If the indirect comparison is transitive then

$$\begin{aligned} f(v_0, v_1) + ... + f(v_{k-1}, v_k) + f(v_k, v_0) &= f(v_0, v_k) + f(v_k, v_0) \\ &= f(v_0, v_k) - f(v_0, v_k) \\ &= 0 \end{aligned}$$

and so the sum of the direct comparisons around the cycle is zero. Any such cycle is **consistent** and every indirect comparison of length one less than the length of the cycle is transitive.

Contradictions in a dataset of pairwise comparisons produce intransitive paths in the network and therefore can be understood by the behaviour of cycles in the simplicial complex. Contradicting pairwise comparisons, direct and indirect, are represented by inconsistent cycles and the further from zero the sum around the cycle is, the greater the disparity between the comparisons.

Non-contradictory information can be extracted from the dataset by finding a collection of pairwise comparisons which are close to the observed comparisons and are consistent on every cycle in the complex. This can be achieved by solving the optimisation problem

$$\min_{f \in M} ||f - O||_2^2 \tag{1}$$

where $O$ are the observed pairwise comparisons and $M$ is the subset of $C^1$ containing 1-cochains which do not produce any inconsistencies in the complex.

In the above optimisation problem, equal emphasis is placed on each of the edges of the complex and so the solution will not favour matching a particular comparison. This assumes that each comparison in the dataset is equally reliable in representing the competition between the alternatives, however the provenance of each data point is likely to vary in practice. Each pairwise comparison can be weighted according to a chosen measure or estimate of the reliability of the information which contributes to it (for instance the trustworthiness of the sources, the recency of the information). These weights are given by a symmetric matrix $W$ and the extracted consistent pairwise comparisons are found by solving the weighted optimisation problem

$$\min_{f \in M} ||f - O||_{2,W}^2 = \min_{f \in M} \sum_{\{i,j\} \in E} W_{ij}(f_{ij} - O_{ij})^2 \tag{2}$$

An inner product on $C^1$ can be constructed from the reliability weights

with

$$\langle f, g \rangle_{C^1} = \sum_{\{i,j\} \in E} W_{ij} f(v_i, v_j) g(v_i, v_j)$$

Inner products can similarly be defined on $C^0$ and $C^2$ however, since we are concerned with pairwise comparison datasets, there is no rationale to emphasise certain vertices or triangles in the complex and so unweighted Euclidean inner products are chosen for these vector spaces. With these choices of inner products, cohomological theories can be brought to bear upon the complex. Coboundary operators and adjoint operators exist between the low dimensional vector spaces of cochains producing the cochain complex

$$C^0 \underset{\delta_0^*}{\overset{\delta_0}{\rightleftarrows}} C^1 \underset{\delta_1^*}{\overset{\delta_1}{\rightleftarrows}} C^2$$

The Hodge Decomposition Theorem orthogonally decomposes cochain groups by their relationships with neighbours in the cochain complex. The space of possible pairwise comparisons can be decomposed, with each orthogonal subspace exhibiting different characteristics. Projecting the observed pairwise comparisons into each subspace, the characteristics of the dataset can be evaluated.

**Theorem 1** (Hodge Decomposition Theorem)**.** *$C^k(K, \mathbb{R})$ admits an orthogonal decomposition*

$$C^k(K, \mathbb{R}) = \operatorname{im}(\delta_{k-1}) \oplus \ker(\Delta_k) \oplus \operatorname{im}(\delta_k^*)$$

*and*

$$\ker(\Delta_k) = \ker(\delta_k) \cap \ker(\delta_{k-1}^*).$$

*Proof.* It is a standard result that for any subspace $U$ in an inner product space $V$, with the inner product $\langle \cdot, \cdot \rangle_V$, it holds that $V = U \oplus U^\perp$, where $U^\perp$ denotes the orthogonal complement of $U$ defined as

$$U^\perp := \{v \in V \mid \langle v, u \rangle_V = 0, \ \forall u \in U\}$$

and $\oplus$ denotes a direct sum.

As $\text{im}(\delta_{k-1}) \subseteq C^k$, it follows that $C^k = \text{im}(\delta_{k-1}) \oplus \text{im}(\delta_{k-1})^\perp$.

$$\begin{aligned}\text{im}(\delta_{k-1})^\perp &= \{f_k \in C^k \mid \langle \delta_{k-1} g_{k-1}, f_k \rangle_{C^k} = 0,\ \forall g_{k-1} \in C^{k-1}\} \\ &= \{f_k \in C^k \mid \langle g_{k-1}, \delta_{k-1}^* f_k \rangle_{C^{k-1}} = 0,\ \forall g_{k-1} \in C^{k-1}\}\end{aligned}$$

For all $f_k \in C^k$, $\delta_{k-1}^* f_k \in C^{k-1}$, and so if $f_k \in \text{im}(\delta_{k-1})^\perp$, then

$$\langle \delta_{k-1}^* f_k, \delta_{k-1}^* f_k \rangle_{C^{k-1}} = 0$$

By the definition of an inner product, if $\langle x, x \rangle = 0$ then $x$ must necessarily equal 0. As such, we have that $\delta_{k-1}^* f_k = 0$ and so $\text{im}(\delta_{k-1})^\perp = \ker(\delta_{k-1}^*)$ Therefore $C^k = \text{im}(\delta_{k-1}) \oplus \ker(\delta_{k-1}^*)$. Similarly, $C^k = \text{im}(\delta_k^*) \oplus \ker(\delta_k)$.

For any $f_{k-1} \in C^{k-1}$ and any $g_{k+1} \in C^{k+1}$,

$$\begin{aligned}\langle (\delta_k \circ \delta_{k-1}) f_{k-1}, g_{k+1} \rangle_{C^{k+1}} &= \langle \delta_k (\delta_{k-1} f_{k-1}), g_{k+1} \rangle_{C^{k+1}} \\ &= \langle \delta_{k-1} f_{k-1}, \delta_k^* g_{k+1} \rangle_{C^k} \\ &= \langle f_{k-1}, \delta_{k-1}^* (\delta_k^* g_{k+1}) \rangle_{C^{k-1}} \\ &= \langle f_{k-1}, (\delta_{k-1}^* \circ \delta_k^*) g_{k+1} \rangle_{C^{k-1}}\end{aligned}$$

and so $(\delta_k \circ \delta_{k-1})^* = \delta_{k-1}^* \circ \delta_k^*$.

By Lemma 1, $\delta_k \circ \delta_{k-1} = 0$ and so $\delta_{k-1}^* \circ \delta_k^* = (\delta_k \circ \delta_{k-1})^* = 0$. This implies that $\text{im}(\delta_k^*) \subseteq \ker(\delta_{k-1}^*)$. From this it follows that

$$\begin{aligned}\ker(\delta_{k-1}^*) &= C^k \cap \ker(\delta_{k-1}^*) \\ &= (\text{im}(\delta_k^*) \oplus \ker(\delta_k^*)) \cap \ker(\delta_{k-1}^*) \\ &= (\text{im}(\delta_k^*) \cap \ker(\delta_{k-1}^*)) \oplus (\ker(\delta_k) \cap \ker(\delta_{k-1}^*)) \\ &= \text{im}(\delta_k^*) \oplus (\ker(\delta_k) \cap \ker(\delta_{k-1}^*))\end{aligned}$$

and therefore $C^k = \text{im}(\delta_{k-1}) \oplus (\ker(\delta_k) \cap \ker(\delta_{k-1}^*)) \oplus \text{im}(\delta_k^*)$. It remains to show that $\ker(\Delta_k) = \ker(\delta_k) \cap \ker(\delta_{k-1}^*)$

Let $f_k \in \ker(\delta_k) \cap \ker(\delta_{k-1}^*)$. As $\Delta_k = \delta_k^* \circ \delta_k + \delta_{k-1} \circ \delta_{k-1}^*$,

$$\begin{aligned}
\Delta_k f_k &= (\delta_k^* \circ \delta_k)(f_k) + (\delta_{k-1} \circ \delta_{k-1}^*)(f_k) \\
&= \delta_k^*(\delta_k f_k) + \delta_{k-1}(\delta_{k-1}^* f_k) \\
&= \delta_k^*(0) + \delta_{k-1}(0) \\
&= 0
\end{aligned}$$

Hence $\ker(\delta_k) \cap \ker(\delta_{k-1}^*) \subseteq \ker(\Delta_k)$.

It remains to show that $\ker(\Delta_k) \subseteq \ker(\delta_k) \cap \ker(\delta_{k-1}^*)$. Let $f_k \in \ker(\Delta_k)$. It follows that

$$0 = \Delta_k f_k = (\delta_k^* \circ \delta_k) f_k + (\delta_{k-1} \circ \delta_{k-1}^*) f_k$$

and so

$$\delta_k^*(\delta_k f_k) = -\delta_{k-1}(\delta_{k-1}^* f_k)$$

Consider the inner product of $\delta_k^*(\delta_k f_k)$ with itself.

$$\begin{aligned}
\langle \delta_k^*(\delta_k f_k), \delta_k^*(\delta_k f_k) \rangle_{C^k} &= \langle \delta_k^*(\delta_k f_k), -\delta_{k-1}(\delta_{k-1}^* f_k) \rangle_{C^k} \\
&= -\langle \delta_k^*(\delta_k f_k), \delta_{k-1}(\delta_{k-1}^* f_k) \rangle_{C^k} \\
&= -\langle \delta_k f_k, (\delta_k \circ \delta_{k-1})(\delta_{k-1}^* f_k \rangle_{C^{k+1}}
\end{aligned}$$

By Lemma 1, $(\delta_k \circ \delta_{k-1})(\delta_{k-1}^* f_k) = 0$, so

$$\begin{aligned}
\langle \delta_k^*(\delta_k f_k), \delta_k^*(\delta_k f_k) \rangle_{C^k} &= -\langle \delta_k f_k, (\delta_k \circ \delta_{k-1})(\delta_{k-1}^* f_k \rangle_{C^{k+1}} \\
&= -\langle \delta_k f_k, 0 \rangle_{C^{k+1}} \\
&= 0
\end{aligned}$$

and therefore $\delta_k^*(\delta_k f_k) = 0$. Now consider the inner product of $\delta_k f_k$ with

itself.

$$\begin{aligned}\langle \delta_k f_k, \delta_k f_k \rangle_{C^{k+1}} &= \langle f_k, \delta_k^*(\delta_k f_k) \rangle_{C^k} \\ &= \langle f_k, 0 \rangle_{C^k} \\ &= 0\end{aligned}$$

Hence $\delta_k f_k = 0$ and so $f_k \in \ker(\delta_k)$. Similarly, as $0 = \delta_k^*(\delta_k f_k) = -\delta_{k-1}(\delta_{k-1}^* f_k)$, it follows that

$$\begin{aligned}\langle \delta_{k-1}^* f_k, \delta_{k-1}^* f_k \rangle_{C^{k-1}} &= \langle f_k, \delta_{k-1}(\delta_{k-1}^* f_k \rangle_{C^k} \\ &= \langle f_k, 0 \rangle_{C^k} \\ &= 0\end{aligned}$$

Again, we have that $\delta_{k-1}^* f_k = 0$ and so $f_k \in \ker(\delta_{k-1}^*)$. Therefore $\ker(\Delta_k) \subseteq \ker(\delta_k) \cap \ker(\delta_{k-1}^*)$. □

The space of all possible pairwise comparisons can be orthogonally decomposed as $C^1 = \text{im}(\delta_0) \oplus \ker(\delta_0^*)$. The cochains in these two subspaces behave differently on the cycles of the complex, providing a useful tool for separating a dataset of pairwise comparisons into consistent and inconsistent parts.

i) **Consistent**: Directed edge functions in the image of $\delta_0$ can be expressed in the form $\delta_0 g$ for some vertex function $g$. Every cycle is consistent since, for any path $\{v_0, ..., v_k\}$,

$$\begin{aligned}(\delta_0 g)(v_0, v_1) + ... + (\delta_0 g)(v_{k-1}, v_k) &= -g(v_0) + g(v_1) + ... + g(v_k) \\ &= g(v_k) - g(v_0) \\ &= (\delta_0 g)(v_0, v_k)\end{aligned}$$

These are **consistent** collections of pairwise comparisons and contain no contradictory information.

**Lemma 2.** *Any two cochains $g, h \in C^0$ satisfy $\delta_0 g = \delta_0 h$ if and only if*

*g and h differ by an additive constant on each connected component of the complex.*

*Proof.* If $g, h \in C^0$ differ by an additive constant on each connected component of the complex then $g_i = h_i + c_i$ where $c_i$ is the additive constant on the connected component which vertex $i$ belongs to. For each directed edge $\{i, j\} \in E$, it follows that $c_i = c_j$ and so

$$\begin{aligned}(\delta_0 g)_{i,j} &= g_j - g_i \\ &= (h_j + c_j) - (h_i + c_i) \\ &= h_j - h_i \\ &= (\delta_0 h)_{i,j}\end{aligned}$$

Let $g, h \in C^0$ be such that $\delta_0 g = \delta_0 h$. For all $i, j \in V$ it follows that $g_j - g_i = h_j - h_i$. Fixing an initial vertex $v_{0_k}$ for each connected component of the complex, for every vertix in component $k$ the following holds

$$\begin{aligned}g_i &= g_i - g_{0_k} + g_{0_k} \\ &= h_i - h_{0_k} + g_{0_k}\end{aligned}$$

and so $g_i - h_i = g_{0_k} - h_{0_k}$ for every vertex in the $k$-th component. This holds for each connected component and therefore

$$h = g + \sum_k \alpha_j c_j$$

where $j$ indexes the connected components of the complex, $\alpha_j \in \mathbb{R}$ and $(c_j)_i = 1$ if vertex $i$ belongs to the component $j$ and 0 otherwise. $\square$

**Example 9.** *Decomposing the* 1*-cochain in Example 8, the consistent* 1*-cochain is represented by the following simplicial complex*

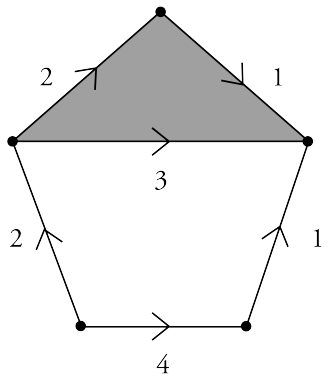


Figure 11: Consistent 1-cochain component

ii) **Inconsistent**: The adjoint of $\delta_0$ satisfies $\langle f, \delta_0^* g\rangle_{C^0} = \langle \delta_0 f, g\rangle_{C^1}$ for any $f \in C^0$ and $g \in C^1$. Expanding these two inner products gives

$$\langle f, \delta_0^* g\rangle_{C^0} = \sum_{i\in V} f_i(\delta_0^* g)_i$$

and, noting that $W_{ij} = 0$ if there is no comparison between alternatives $i$ and $j$,

$$\begin{aligned}
\langle \delta_0 f, g\rangle_{C^1} &= \sum_{\{i,j\}\in E} W_{ij}(\delta_0 f)_{ij} g_{ij} \\
&= \sum_{\{i,j\}\in E} W_{ij} f_j g_{ij} - \sum_{\{i,j\}\in E} W_{ij} f_i g_{ij} \\
&= -\sum_{\{i,j\}\in E} W_{ji} f_j g_{ji} - \sum_{\{i,j\}\in E} W_{ij} f_i g_{ij} \\
&= -\sum_{i\in V}\sum_{j\in V} W_{ij} f_i g_{ij}
\end{aligned}$$

Therefore $(\delta_0^* g)_i = -\sum_{j\in V} W_{ij} g_{ij}$ for all $i \in V$.

The kernel of $\delta_0^*$ contains directed edge functions such that $\sum_{j\in V} W_{ij} f_{ij} = 0$ for every alternative $i$. This subspace can be understood in the framework of vector calculus. In the language of vector calculus, preference

'flows' from less to more preferable alternatives and $\delta_0^*$ is known as the **divergence**.

At every vertex in the complex, the total weighted in-flow equals the total weighted out-flow for directed edge functions in this subspace. Wherever flow leaves a vertex, indicating that it is less preferable than a neighbouring alternative, there must be a path for some of that flow to return to the initial vertex, forming a cycle. Therefore any directed edge function in the kernel of $\delta_0^*$ is **inconsistent**, only producing inconsistent cycles and assigning a 0 to every edge which does not participate in a cycle.

**Example 10.** *The inconsistent* 1*-cochain component of the pairwise comparison matrix in Example 8 is represented by the following simplicial complex*

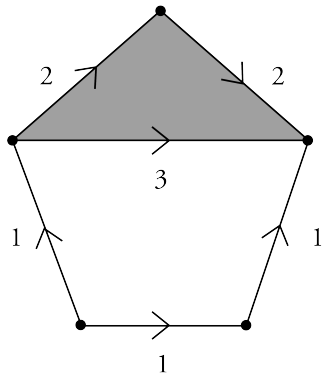


Figure 12: Consistent 1-cochain component

Any choice of pairwise comparisons on the complex can be written as the addition of a consistent cochain and an inconsistent cochain. Therefore $\text{im}(\delta_0)$ is the only subspace of $C^1$ to contain pairwise comparisons which produce no inconsistencies nor contradictory information.

The remaining problem is how to find this decomposition into consistent and inconsistent cochains. The optimisation problem (2) extracting consistent pairwise comparisons from a dataset of pairwise comparisons can be rewritten as

$$\min_{f \in \text{im}(\delta_0)} ||f - O||^2_{2,W} \tag{3}$$

Any function in the image of $\delta_0$ can be expressed in terms of a vertex function in $C_0$ and therefore the optimisation problem is equivalent to

$$\min_{g \in C^0} ||\delta_0 g - O||^2_{2,W} = \min_{g \in C^0} \sum_{\{i,j\} \in E} W_{ij}(g_j - g_i - O_{ij})^2 \tag{4}$$

The consistent pairwise comparisons which minimise the weighted difference with the observed pairwise comparisons can be found by applying the boundary operator $\delta_0$ to the solution of this optimisation problem.

**Theorem 2** (Jiang et al., 2011)**.** *There exist minima of the optimisation problem (4) which are given by the solutions of*

$$\Delta_0 f = \delta_0^* O \tag{5}$$

*Any two minima differ by an additive constant on each connected component.*

*Proof.* The cost function associated with the optimisation problem (4) is given by

$$\begin{aligned} J(f) &= \langle \delta_0 f - O, \delta_0 f - O \rangle_{C^1} \\ &= \sum_{\{i,j\} \in E} W_{ij}(f_j - f_i - O_{ij})^2 \end{aligned}$$

$J(f)$ is a continuous real-valued function which is bounded below by zero. If this function does not have a minimum which is achieved by a finite cochain, then $J(f)$ must tend to an infimum as $f$ tends to infinity in some direction. There are two cases under which a cochain can tend to infinity: either the value on each vertex in a connected component changes at the same rate, or there exists at least one vertex in a connected component whose value changes at a faster rate than the others.

Consider $\lim_{h \to \infty} J(g + h\hat{u})$ in a direction $\hat{u}$ from a finite cochain $g \in C^0$. In the first case, $\hat{u}$ is constant on every vertex in the same connected component. By Lemma 2, $\delta_0(g + h\hat{u}) = \delta_0 g$ and so $J(g + h\hat{u}) = J(g)$. If $J(f)$ tends to an infimum as $f$ tends to infinity, with the value of each vertex in a connected component changing at the same rate, then there exists a

minimum achieved by a finite cochain.

In the second case, $\hat{u}$ varies across at least one connected component. Let $v \in C^0$ be such that $v_i$ is the minimum value of $\hat{u}$ across the connected component that vertex $i$ belongs to. This vector is constant on every vertex in the same connected component so $J(g+h\hat{u}) = J(g+h(\hat{u}-v))$ by Lemma 2. Every element of $\hat{u}-v$ is non-negative and there must be at least one non-zero element, therefore there exists at least one edge $\{\alpha, \beta\}$ in the complex with $(g+h(\hat{u}-v))_\beta - (g+h(\hat{u}-v))_\alpha$ tending to (positive or negative) infinity as $h$ tends to infinity. Thus

$$J(g+h\hat{u}) = \sum_{\{i,j\}\in E} W_{ij}((g+h(\hat{u}-v))_\beta - (g+h(\hat{u}-v))_\alpha - O_{ij})^2$$

tends to infinity as $h$ tends to infinity. There must therefore exist a finite cochain which minimises $J(f)$.

At any critical point of a differentiable function, the directional derivative in every direction is zero. In an arbitrary direction $\hat{u}$

$$\begin{aligned} J(f+h\hat{u}) &= \langle \delta_0(f+h\hat{u}) - O, \delta_0(f+h\hat{u}) - O\rangle_{C^1} \\ &= \langle \delta 0 f - O + h\delta_0\hat{u}, \delta_f - O + h\delta 0\hat{u}\rangle_{C^1} \\ &= J(f) + h\langle \delta_0 f - O, \delta_0\hat{u}\rangle_{C^1} + h^2\langle \delta_0\hat{u}, \delta_0\hat{u}\rangle_{C^1} \end{aligned}$$

and so the directional derivative of $J(f)$ in the direction of $\hat{u}$ is given by

$$\begin{aligned} \nabla_{\hat{u}} J(f) &= \lim_{h\to 0} \frac{J(f+h\hat{u}) - J(f)}{h} \\ &= \lim_{h\to 0} \langle \delta_0 f - O, \delta_0\hat{u}\rangle_{C^1} + h\langle \delta_0\hat{u}, \delta_0\hat{u}\rangle_{C^1} \\ &= \langle \delta_0 f - O, \delta_0\hat{u}\rangle_{C^1} \end{aligned}$$

This limit exists and so $J(f)$ is differentiable. The directional derivative is zero in every direction precisely when $f \in C^0$ satisfies $\Delta_0 f = \delta_0^* O$.

Let $f_1, f_2$ be critical points of $J(f)$. Both cochains satisfy (5), thus

$$\delta_0^* O = \Delta_0 f_1 = \Delta_0 f_2$$

and $(f_1 - f_2) \in \ker(\Delta_0)$. Since $\Delta_0(f_1 - f_2) = 0$, the cochain $(f_1 - f_2)$ is an eigenvector of $\Delta_0$ corresponding to an eigenvalue $\lambda_k = 0$. It follows that

$$\sum_{\{i,j\}\in E} W_{ij}((f_1 - f_2)_j - (f_1 - f_2)_i)^2 = (f_1 - f_2)'\Delta_0(f_1 - f_2) = 0$$

and so $(f_1 - f_2)_j = (f_1 - f_2)_i$ if $\{i, j\}$ is a directed edge in the complex. Hence $(f_1 - f_2)$ is constant across each connected component of the complex. By Lemma 2, $\delta_0 f_1 = \delta_0 f_2$ and so both cochains have the same value of $J(f)$. Therefore every critical point has the same cost associated to it and, since there exists at least one minimum, every critical point is a minimum of $J(f)$. □

Theorem 2 reduces the optimisation problem to a system of linear equations (5). If $\Delta_0$ is non-singular then a unique solution to the optimisation problem can be found by inverting this matrix. However it follows from Lemma 2 that (3) has infinitely many solutions and so $\Delta_0$ must be singular. Indeed it is well-known that the graph Laplacian has 0 as an eigenvalue and that this eigenvalue has multiplicity given by the number of connected components in the graph (Chung 1997, Von Luxburg 2007).

Although (5) cannot be solved by inverting $\Delta_0$, it is possible to find a solution (and therefore all solutions) by constructing a matrix inverse-like object. A **generalised inverse** is any matrix satisfying the condition

$$AXA = A$$

A variety of generalised inverses can be constructed which are applicable to different problems, depending on the properties they exhibit. The **Penrose equations** are four conditions used to categorise generalised inverses (Penrose 1955):

$$AXA = A \tag{6}$$

$$XAX = X \tag{7}$$

$$(AX)^H = AX \tag{8}$$

$$(XA)^H = XA \tag{9}$$

where $X^H$ denotes the complex transpose of $X$.

**Theorem 3** (Penrose, 1955)**.** *A unique generalised inverse exists satisfying all four of the Penrose equations for any* $A \in \mathbb{R}^{m\times n}$.

*Penrose, 1955.* If a matrix satisfies the first and fourth Penrose equations then

$$XAA^H = (XA)^H A^H = (AXA)^H = A^H$$

Conversely if $XAA^H = A'$ then

$$(XA)^H = A^H X^H = XAA^H X^H = (XA)(XA)^H = ((XA)(XA)^H)^H = XA$$

and

$$A = (A^H)^H = (XAA^H)^H = AA^H X^H = A(XA)^H = AXA$$

Therefore the first and fourth Penrose equations are equivalent to the condition

$$XAA^H = A^H \tag{10}$$

Similarly, given the symmetry of the Penrose equations, it follows that the second and third Penrose equations are equivalent to the condition

$$XX^H A^H = X \tag{11}$$

The set of matrices $\{(A^H A)^k | k \in \mathbb{N}\}$ must be linearly dependent, since $A \in \mathbb{R}^{m\times n}$, so there exists $p \in \mathbb{N}$ with

$$\lambda_1 A^H A + \lambda_2 (A^H A)^2 + ... + \lambda_p (A^H A)^p = 0$$

where $\lambda_1, ..., \lambda_p$ are not all zero. If $\lambda_r$ denotes the first non-zero constant then

$$(A^H A)^r = -\frac{1}{\lambda_r}(A^H A)^{r+1}(\lambda_{r+1} I + \lambda_{r+2} A^H A + ... + \lambda_p (A^H A)^{p-r-1}) \tag{12}$$

Observing that $BAA^H = CAA^H$ implies $BA = CA$ and $BA^H A = CA^H A$

implies $BA^H = CA^H$, and letting

$$W = -\frac{1}{\lambda_r}(\lambda_{r+1}I + \lambda_{r+2}A^HA + ... + \lambda_p(A^HA)^{p-r-1})$$

it follows from equation (12) that $WA^HAA^H = A'$ and hence that $WA^H$ is a solution to equation (10). Applying the first and fourth Penrose equations to (10),

$$\begin{aligned} A^H &= W^HA^HAA^H \\ &= WA^H(AWA^HA)A^H \\ &= WA^HA(A^HAW^H)A^H \\ &= (A^H)AW^HA^H \end{aligned}$$

and so $WA^H = (WA^H)(WA^H)^HA^H$. Therefore $WA^H$ is a solution to all four of the Penrose equations.

If $X$ and $Y$ are solutions to all four of the Penrose equations then

$$\begin{aligned} X &= XX^HA^H \\ &= XX^H(YAA^H) \\ &= XX^H(A^HY^H)A^H \\ &= XX^HA^H(AY) \\ &= (X)AY \\ &= XA(YY^HA^H) \\ &= XAY(AY) \\ &= XA(A^HY^H)Y \\ &= (A^H)Y^HY \\ &= (YA)Y \\ &= Y \end{aligned}$$

Hence there is a unique solution of the Penrose equations for any given matrix. □

The unique matrix satisfying all four of the Penrose equations is the **Moore-Penrose pseudoinverse** and denoted by $A^\dagger$. For any invertible matrix, it is precisely the inverse of the matrix. The Moore-Penrose pseudoinverse can be used to solve overdetermined systems of equations, such as those given by (5).

**Lemma 3.** *The overdetermined system of equations $Ax = y$ has solutions if and only if*

$$AA^\dagger y = y$$

*In particular $A^\dagger y$ is a solution.*

*Proof.* If $AA^\dagger y = y$ then $A^\dagger y$ is a solution to $Ax = y$. Conversely if $x$ is a solution to $Ax = y$ then

$$\begin{aligned} y &= Ax \\ &= AA^\dagger Ax \\ &= AA^\dagger y \end{aligned}$$

□

Theorem 2 establishes that optimisation problem (4) has solutions and hence (5) also has solutions. Applying Lemma 3, the solutions to the optimisation problem are given by

$$\hat{s} = \Delta_0^\dagger \delta_0^* O + \hat{u} \tag{13}$$

where $\hat{u} \in C^0$ is constant on each connected component of the complex. The consistent pairwise comparisons extracted from observed pairwise comparisons are therefore

$$S = \delta_0 \Delta_0^\dagger \delta_0^* O \tag{14}$$

The solution given by (13) is a real-valued function measuring the underlying preference of the consistent pairwise comparisons for each alternative. The HodgeRank algorithm derives a ranking of the alternatives by the rule that $i \leq j$ if and only if $\hat{s}(i) \leq \hat{s}(j)$.

After extracting the consistent ranking data from the observed pairwise comparisons, the remaining ranking data, the **residual**, is a directed edge function which is intransitive, or trivial, on every cycle. This inconsistent ranking data, given by $R = O - S$, produces intransitive patterns of preferences from which a consistent ranking of the alternatives cannot be found.

## 4.3 Inconsistencies

In their paper detailing the HodgeRank algorithm, Jiang et al. (2011) identify inconsistent ranking data as an issue for ranking algorithms, producing intransitive patterns of preferences which do not permit a coherent ranking of the underlying alternatives. The presence of inconsistent ranking data is overcome by applying discrete Hodge theory to a simplicial complex representation of the observed pairwise comparisons, identifying and separating out inconsistencies, and forming a ranking of the alternatives from the remaining ranking data.

Although this approach allows for a coherent ranking of the alternatives from pairwise comparisons containing contradictions, it disregards any information contained within the inconsistent ranking data. It is tempting to suggest that inconsistent ranking data reflects errors in the underlying data however, as the authors note, inconsistent ranking data is "not necessarily due to error or noise in the data by may very well be an inherent characteristic of the data" (Jiang et al. 2011)[p. 206]. There is therefore a need to understand the size, location and nature of inconsistencies within the ranking data, and establish whether the inconsistent ranking data is noise in the dataset or a fundamental feature of the ranking data arising from the relationships between the alternatives.

Jiang et al. (2011) examined the inconsistent ranking data, the residual, to assess the degree to which the pairwise comparisons admit a coherent ranking. If the residual is small, the pairwise comparisons contain few inconsistencies and the ranking found by HodgeRank is representative of the underlying competitiveness of the alternatives. On the other hand, if the residual is large then there is a high level of inconsistency in the pairwise comparisons

and any ranking found from them should be considered unreliable.

Interpreting the residual directly is difficult as it is neither scale- nor translation-invariant and thus the size of the residual naturally varies across different pairwise comparison matrices. Contextualising the size of the residual in relation to the pairwise comparisons it is extracted from, the **cyclicity ratio**

$$c = \frac{||R||_{2,W}}{||O||_{2,W}} \tag{15}$$

measures the level of inconsistency in the dataset, approaching one as the level of inconsistency increases.

The cyclicity ratio provides a measure by which the validity of the ranking solution found by the HodgeRank algorithm can be assessed, however it does not provide insights into the nature of the inconsistencies in the ranking data. Hodge theory provides tools to understand the type of inconsistencies present in the ranking data. By the Hodge Decomposition Theorem (Theorem 1), the space of inconsistencies can be orthogonally decomposed into

$$\ker(\delta_0^*) = \ker(\Delta_1) \oplus \operatorname{im}(\delta_1^*)$$

with $\ker(\Delta_1) = \ker(\delta_0^*) \cap \ker(\delta_1)$.

Any directed edge function which belongs to $\ker(\delta_0^*)$ can be expressed as the sum of a directed edge function in $\ker(\Delta_1)$ and a directed edge function in $\operatorname{im}(\delta_1^*)$, and so inconsistent ranking data can be understood by its projection onto these two subspaces. Understanding the behaviour of directed edge functions in both of these subspaces on cycles in the simplciail complex provides a window to understanding the inconsistent ranking data contained in the observed pairwise comparisons.

(i) **Type I:** The coboundary operator $\delta_1$ maps directed edge functions to oriented triangle functions by summing (with respect to the orientation) the value of the directed edge function around the triangle. Given an oriented triangle

$\{v_0, v_1, v_2\}$ and a directed edge function $f \in C^1$,

$$(\delta_1 f)(v_0, v_1, v_2) = f(v_0, v_1) + f(v_1, v_2) - f(v_0, v_2)$$

Hence directed edge functions in $\ker(\delta_1)$ behave consistently on oriented triangles in the complex (cycles of length three).

Directed edge functions in $\ker(\Delta 1)$ also belong to the subspace $\ker(\delta_0^*)$ and therefore produce inconsistencies on cycles of length greater than three. These Type I inconsistent directed edge functions produce transitive patterns of preferences on any triple of alternatives but intransitive patterns of preferences on larger subsets of alternatives.

**Example 11.** *Decomposing the inconsistent* 1*-cochain component of the pairwise comparison matrix in Example 8, the Type I inconsistent* 1*-cochain is represented by the following simplicial complex*

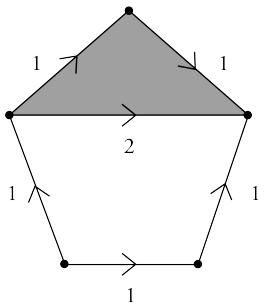


Figure 13: Type I inconsistency

(ii) **Type II:** The adjoint of the 1-*th* coboundary operator is a linear operator which satisfies $\langle f, \delta_1^* g\rangle_{C^1} = \langle \delta_0 f, g\rangle_{C^2}$ for any $f \in C^1$ and $g \in C^2$. Expanding these two inner products gives

$$\begin{aligned}
\sum_{\{i,j\}\in E} W_{ij} f_{ij} (\delta_1^* g)_{ij} &= \langle f, \delta_1^* g\rangle_{C^1} \\
&= \langle \delta_1 f, g\rangle_{C^2} \\
&= \sum_{\{i,j,k\}\in T} (\delta_1 f)_{ijk} g_{ijk} \\
&= \sum_{\{i,j,k\}\in T} g_{ijk}(f_{ij} + f_{jk} + fki) \\
&= \sum_{\{i,j\}\in E} \Big( \sum_{\{i,j,k\}\in T} f_{ij} g_{ijk} - \sum_{\{i,p,j\}\in T} f_{ij} g_{ipk} \Big) \\
&= \sum_{\{i,j\}\in E} \sum_{\{i,j,k\}\in T'} f_{ij} g_{ijk} \\
&= \sum_{\{i,j\}\in E} f_{ij} \cdot \sum_{\{i,j,k\}\in T'} g_{ijk}
\end{aligned}$$

where $T'$ denotes the set of unoriented triangles in the complex. It follows that

$$(\delta_1^* g)_{ij} = \frac{1}{W_{ij}} \sum_{\{i,j,k\}\in T'} g_{ijk} \tag{16}$$

The adjoint of the 1-th coboundary operator, $\delta_1^*$, diffuses values on oriented triangles to the directed edges which bound them. These Type II directed edge functions produce intransitive patterns of preferences on subsets of alternatives of any size.

**Example 12.** *The Type II inconsistent* 1*-cochain component of the pairwise comparison matrix in Example 8 is represented by the following simplicial complex*

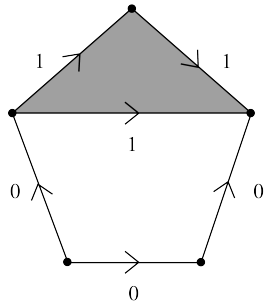


Figure 14: Type II inconsistency

Both Type I and Type II directed edge functions produce intransitive patterns of preferences, forming inconsistent cycles in the simplicial complex. They differ in their behaviour on oriented triangles, with Type I functions providing a coherent ranking of triples of alternatives whereas Type II functions cannot be used to rank any subset of the alternatives.

By orthogonally decomposing the residual into Type I and II directed edge functions, its level of inconsistency on a local level can be determined. If the inconsistent ranking data is mostly projected onto the subspace of Type I functions, there is non-contradictory information regarding relationships between triples of alternatives which is excluded by the HodgeRank algorithm.

**Theorem 4.** *Given a directed edge function $O$, the residual $R$ can be orthogonally decomposed as*

$$R = P_{\ker(\Delta_1)}O + P_{\mathrm{im}(\delta_1^*)}O \tag{17}$$

*where the projections are given respectively by $P_{\ker(\Delta_1)} = I - \Delta_1$ and $P_{\mathrm{im}(\delta_1^*)} = \Delta_1 - \delta_0\Delta_0^\dagger\delta_0^*$*

*Proof.* The space of directed edge function $C^1$ can be orthogonally decomposed with

$$C^1 = \mathrm{im}(\delta_0) \oplus \ker(\delta_0^*)$$

Any directed edge function $O$ can be orthogonally decomposed into two cochains $S$ and $R$, where $S$ is a consistent cochain given by (14) and $R$ is the residual given by (**??**). Projections are linear operators and so the projection of $O$ onto the subspace $\ker(\delta_0^*)$ can be expressed as

$$P_{\ker(\delta_0^*)}O = P_{\ker(\delta_0^*)}S + P_{\ker(\delta_0^*)}R$$

The consistent cochain $S$ belongs to the $\mathrm{im}(\delta_0)$, a subspace which is orthogonal to $\ker(\delta_0^*)$, from whence it follows that $P_{\ker(\delta_0^*)}S = 0$. In addition, $P_{\ker(\delta_0^*)}R = R$ since

$$\begin{aligned}
\delta_0^* R &= \delta_0^*(O - S) \\
&= \delta_0^* O - \delta_0^* \delta_0 s \\
&= \delta_0^* O - \Delta_0 s \\
&= 0
\end{aligned}$$

and thefore $R = P_{\ker(\delta_0^*)}O$. Under the Hodge Decomposition Theorem (Theorem 1), the subspace $\ker(\delta_0^*)$ can be further decomposed as $\ker(\Delta_1) \oplus \mathrm{im}(\delta_1^*)$ and so $R = P_{\ker(\Delta_1)}O + P_{\mathrm{im}(\delta_1^*)}O$ as claimed.

It remains to derive expressions for the projections onto $\ker(\Delta_1)$ and $\mathrm{im}(\delta_1^*)$. For any $x \in C^1$ and $y \in \ker(\Delta_1)$,

$$\begin{aligned}
\langle x - (I - \Delta_1)x, y\rangle_{C^1} &= \langle \Delta_1 x, y\rangle_{C^1} \\
&= \langle (\delta_0\delta_0^* + \delta_1^*\delta_1)x, y\rangle_{C^1} \\
&= \langle \delta_0\delta_0^* x, y\rangle_{C^1} + \langle \delta_1^*\delta_1 x, y\rangle_{C^1} \\
&= \langle x, \delta_0\delta_0^* y\rangle_{C^1} + \langle x, \delta_1^*\delta_1 y\rangle_{C^1} \\
&= \langle x, (\delta_0\delta_0^* + \delta_1^*\delta_1)y\rangle_{C^1} \\
&= \langle x, \Delta_1 y\rangle_{C^1} \\
&= \langle x, 0\rangle_{C^1} \\
&= 0
\end{aligned}$$

Hence $(I - \Delta_1)$ is the orthogonal projector onto the subspace $\ker(\Delta_1)$.

Any $x \in C^1$ can be expressed as the sum of its projection onto the three orthogonal subspaces given by the Hodge Decomposition Theorem (Theorem 1). The projection of a directed edge function onto $\text{im}(\delta_0)$ is given by (14) and so

$$\begin{aligned} x &= P_{\text{im}(\delta_0)}x + P_{\ker(\Delta_1)}x + P_{\text{im}(\delta_1^*)}x \\ &= (\delta_0\Delta_0^\dagger\delta_0^*)x + (I - \Delta_1)x + P_{\text{im}(\delta_1^*)}x \end{aligned}$$

Therefore the orthogonal projection onto the subspace $\text{im}(\delta_1^*)$ is given by $\Delta_1 - \delta_0\Delta_0^\dagger\delta_0^*$. □

As with the residual itself, determining whether the inconsistencies in the ranking data are largely of the Type I or II nature is difficult. This project proposes the **local inconsistency ratio**

$$L = \frac{||R - \Delta_1 R||_{2,W}}{||R||_{2,W}} \tag{18}$$

as a measure of the proportion of inconsistency in the pairwise comparison matrix which is consistent on local-scale features of the simplicial complex. If the ranking data has a high local inconsistency ratio, it potentially contains information in its inconsistent component which can be used to rank, in isolation, triples of alternatives.

## 4.4 Measuring Inconsistency

Although (Jiang et al. 2011) acknowledge that inconsistent ranking data may be produced as a consequence of the fundamental characteristics of the underlying alternatives, the HodgeRank algorithm they devise excludes it for the purpose of ranking the alternatives. This is a sensible approach as the inconsistent ranking data contains intransitive patterns of preferences, from which a coherent coherent ranking of the alternatives cannot be found. The authors propose the cyclicity ratio (15) as a means of evaluating how readily a pairwise comparison matrix admits a coherent ranking of the alternatives. If the observed pairwise comparisons are largely inconsistent, the ranking

is produced from a fraction of the available data and should be considered unreliable.

The cyclicity ratio provides a high level understanding of inconsistency in the ranking data, measuring inconsistency across the entire dataset. Whilst this has value in assessing the suitability of HodgeRank for ranking alternatives from a given pairwise comparison matrix, it gives no insights into how, where or why inconsistency arises.

One of the main contributions of this project is exploring the space of inconsistent directed edge functions, measuring inconsistency at different scales, and identifying the features of the complex which produce inconsistency. This more nuanced understanding of inconsistency exploits the information contained in the inconsistent ranking data, and the project goes on to demonstrate the value of this information.

Inconsistencies are cycles in the simplicial complex around which the sum, with respect to orientation, of the directed edge function encoding the observed pairwise comparisons is non-zero. These cycles produce intransitive patterns of preferences, pairwise comparisons which contradict themselves and indicate that an alternative is more preferable than itself.

The degree to which the cycle of alternatives $\{v_0, ..., v_k\}$ is inconsistent (equivalently the extent to which the underlying pairwise comparisons contradict themselves), is given by $\Psi_O(\{v_0, ..., v_k\}) = |\psi_O(\{v_0, ..., v_k\})|$ with

$$\psi_O(\{v_0, ..., v_k\}) = O_{v_k v_0} + \sum_{i=1}^{k} O_{v_{i-1} v_i} \tag{19}$$

where $O$ is the directed edge function corresponding to the pairwise comparisons. A consistent cycle $\{u_0, ..., u_m\}$ satisfies the transitivity criteria and so $\psi_O(\{u_0, ..., u_m\}) = 0$. The more inconsistent a cycle is, the further away this value is from 0, however this can be positive or negative depending on the relative orientation of the directed edges around the cycle, with $\psi_O(\{v_k, ..., v_0\}) = -\psi_O(\{v_0, ..., v_k\})$. Regardless of orientation, the cycle is equally far from satisfying the transitivity criteria and so it is necessary to measure its inconsistency by $\Psi_O(\{v_0, ..., v_k\})$.

By measuring how inconsistent each cycle in the simplicial complex is using the above measure, a distinction can be made between cycles which are close to satisfying transitivity and those which are far from it. This matters because the inconsistent cycles which are close to being consistent are likely a result of noise in the data and thus represent well the underlying preferences for the participating alternatives. On the other hand, highly inconsistent cycles more likely encode fundamental inconsistency arising as a consequence of the relationships between the alternatives.

Information regarding direct pairwise comparisons between pairs of alternatives which participate in many, highly inconsistent cycles can be considered unreliable for the purpose of ranking the entire set of alternatives. These pairwise comparisons are outliers in the pairwise comparison matrix, disagreeing strongly with the majority of the observed pairwise comparisons.

The contribution of a pairwise comparison to inconsistency in the ranking data can be measured as the total of its contribution to inconsistency in each cycle in the simplicial complex representation. In isolation the cause of inconsistency in a cycle, $\sigma$, cannot be identified since subtracting $\psi(\sigma)$ from any of its constituent pairwise comparisons transforms an inconsistent cycle into a consistent one. Instead, each pairwise comparison is deemed equally responsible for producing inconsistency in the cycle. This gives $\Psi_O(\sigma)(i,j) = |\psi_O(\sigma)(i,j)|$ as a measure of the contribution of the pairwise comparison between alternatives $i$ and $j$ to inconsistency in the cycle $\sigma = \{v_0, ..., v_k\}$ with the directed edge function $O$, with

$$\psi_O(\sigma)(i,j) = \begin{cases} {}^{\psi_O(\sigma)}/_{k+1} & \text{if } i,j \in \sigma \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

and $\Psi_O^\infty(i,j) = |\psi_O^\infty(i,j)|$ as a measure of the contribution of the pairwise comparison to inconsistency in the simplicial complex representation, with

$$\psi_O^\infty(i,j) = \sum_\sigma \psi_O(\sigma)(i,j) \tag{21}$$

Again, as with measuring the inconsistency in a cycle, the absolute value is

required to account for the arbitrary choice of orientation.

These measures discount the inconsistency in each cycle by the length of the cycle, accounting for the larger impact that perturbations in the pairwise comparisons have on inconsistency in longer cycles. Consider a cycle $\{v_0, ..., v_{k-1}\}$ with directed edge functions $f$ and $\widetilde{f}$ where $\widetilde{f}_{ij} = f_{ij} + \epsilon$ for every pairwise comparison. Although the inconsistency in the cycle differs with the two directed edge functions,

$$\begin{aligned}\psi_{\widetilde{f}}(\{v_0, ..., v_{k-1}\}) &= \widetilde{f}_{v_k v_0} + \sum_{i=1}^{k} \widetilde{f}_{v_{i-1} v_i} \\ &= f_{v_k v_0} + \epsilon + \sum_{i=1}^{k} f_{v_{i-1} v_i} + \epsilon \\ &= \psi_f(\{v_0, ..., v_{k-1}\}) + k\epsilon\end{aligned}$$

the contribution of each pairwise comparison to inconsistency in the cycle is equal.

Evaluating the contribution of pairwise comparisons to inconsistency across a simplicial complex can be a computationally expensive task, particularly as the number and size of cycles grows. It can, however, be approximated by generalising from paths, in which vertices cannot be repeated, to walks which allow backtracking.

It is well known in graph theory that, as a consequence of matrix multiplication, the number of walks in an undirected graph of length $k$ between vertices $i$ and $j$ is $A^k_{ij}$, where $A$ is the adjacency matrix of the graph. It follows that the number of walks, of any length, between vertices $i$ and $j$ is $\sum_{k=0}^{\infty} A^k_{ij}$. In general this sum diverges however the matrix exponential of the adjacency matrix, the Estrada Index (Estrada & Rodríguez-Velázquez 2005),

$$\exp A(i, j) = \sum_{k=0}^{\infty} \frac{A^k_{ij}}{k!} \tag{22}$$

scales the contribution of walks by the factorial of their length, ensuring the sum converges.

This approach can be extended to approximate the level of inconsistency contained in every cycle of the complex by evaluating the Estrada index of the observed pairwise comparisons however there are three important considerations: (i) inconsistency is defined in terms of the sum, not product, of values on edges; (ii) edges in simplicial complexes have an associated orientation and are not undirected; (iii) the measure should be symmetric.

The first issue can be dealt with by using the element-wise exponential of the observed pairwise comparisons $e^O$, whereby multiplication of the values on the edges in the Estrada index effectively becomes addition. The second issue, the orientation of edges, is accounted for by the skew-symmetry of pairwise comparison matrices since edges in the complex can be traversed in either direction by negating the value on the edge. Lastly, whilst the Estrada index of the element-wise exponential of the observed pairwise comparisons is not symmetric by virtue of the exponential function, a symmetric measure can be formed by combining the value for the directed pair $\{i, j\}$ and $\{j, i\}$.

Under the expectation of transitivity, the sum of the pairwise comparisons around any walk between alternative $i$ and $j$ should equal $O_{ij}$. Thus if a pairwise comparison does not contribute to any inconsistency in the simplicial complex (every cycle it participates in is consistent), $\exp e^O(i, j) = e^{O_{ij}} \exp e^A(i, j)$. An approximation of the contribution of the pairwise comparison between alternatives $i$ and $j$ to inconsistency in the simplicial complex can be defined as $\widetilde{\Psi}_O(i, j) = \frac{1}{2}\left(|\widetilde{\psi}_O(i, j)| + |\widetilde{\psi}_O(j, i)|\right)$ with

$$\widetilde{\psi}_O(i, j) = \exp e^O(i, j) - e^{O_{ij}} \exp e^A(i, j) \tag{23}$$

where $A$ is the adjacency matrix of the undirected graph underlying the complex. Pairwise comparisons which only participate in consistent cycles have a $\widetilde{\Psi}$ value of 0 with this value increasing as a pairwise comparison contributes more to inconsistency in the simplicial complex.

The contribution of a pairwise comparison to inconsistency in the simplicial complex can also be approximated by studying local-scale inconsistency, limiting the cycles under consideration to those of length three. There are two main motivations behind this approach. Firstly, longer cycles are formed

of more pairwise comparisons and so noise in the ranking data is more likely to produce inconsistency in the cycle. Secondly, cycles of length three have a natural topological equivalent in the form of the 2-simplices in the complex and as such, topological tools already exist for examining them.

Any inconsistent directed edge function can be orthogonally decomposed as the sum of a Type I function which belongs to the subspace $\ker(\delta_1)$, and a Type II function which belongs to $\mathrm{im}(\delta_1^*)$, the orthogonal complement of $\ker(\delta_1)$. These functions differ in their behaviour around 2-simplices, with Type I functions, unlike Type II functions, satisfying the transitivity criteria around these local-scale cycles.

The 1-st coboundary operator $\delta_1$ is a map from $C^1$ to $C^2$ which assigns a value to each oriented triangle by summing the pairwise comparisons on the directed edges which bound it with respect to their orientation. It follows that for any cycle of length three $\{v_0, v_1, v_2\}$

$$\begin{aligned}\delta_1(\{v_0, v_1, v_2\}) &= O_{v_2v_0} + O_{v_0v_1} + O_{v_1v_2} \\ &= \psi(\{v_0, v_1, v_2\})\end{aligned}$$

As such, inconsistency in local-scale cycles can be measured by $|\delta_1(\sigma)|$.

Local-scale inconsistency can be diffused back to the pairwise comparisons bounding each local-scale cycle by the adjoint of the 1-st coboundary operator $\delta_1^*$. Since $\Delta_1^{up} = \delta_1^*\delta_1$, this gives a measure of the contribution of each pairwise comparison to local-scale inconsistency as

$$\Phi_O(i,j) = |(\Delta_1^{up} O)_{ij}| \tag{24}$$

This measure differs from $\Psi_O(\sigma)(i,j)$ in that it does not assign equal responsibility for the inconsistency to each pairwise comparisons, rather it diffuses the inconsistency back to the pairwise comparisons with respect to the a priori edge weights. Pairwise comparisons with higher edge weights are deemed to have contributed less to the inconsistency in the cycle.

## 4.5 Incorporating Measures Of Inconsistency

HodgeRank allows reliability weights to be assigned to the pairwise comparisons between alternatives, placing greater emphasis on pairwise comparisons with large weights when finding a ranking of the alternatives. In the formulation of HodgeRank described above, these reliability weights are measures or estimates of the reliability of the information underlying each pairwise comparison, evaluated outside of the HodgeRank framework.

In the previous section, several measures were derived for the contribution of a pairwise comparison on inconsistency within the ranking data. Pairwise comparisons which contribute greatly to inconsistency in the ranking data disagree strongly with the remainder of the available data and can be considered unreliable for the purpose of ranking the underlying alternatives coherently.

This project proposes re-weighting the simplicial complex to account for both the initial reliability weights and the contribution of each pairwise comparison to inconsistency in the complex. No prescribed method is given for combining the two reliability measures as this will be highly application dependent, influenced by both the underlying data and the perceived reliability of the reliability weights (in most applications, a subjective assessment). The new reliability weights should be largest for pairwise comparisons with large initial reliability weights and a small contribution to inconsistency within the ranking data.

Implementing this extension to the HodgeRank algorithm, the simplicial complex representation is first equipped with an inner product on $C^1$, derived from the initial reliability weights, and then re-equipped with an inner product derived from both the initial reliability weights and measure of the contribution of each pairwise comparison to inconsistency in the ranking data. This project goes on to demonstrate an application of this extended HodgeRank algorithm in a case study, and to discuss the costs and benefits of including measures of inconsistency.

## 4.6 Computational Issues

HodgeRank is an algorithm for ranking alternatives from pairwise comparisons by measuring the underlying competitiveness of each alternative. The algorithm represents a pairwise comparison matrix, expressing the degree to which one alternative is preferred over another, as a simplcial complex and applies discrete Hodge theory to understand the ranking data and identify a coherent ranking of the alternatives.

This project implements the HodgeRank algorithm, and variants of, in MATLAB, a programming language often used by the Applied Mathematics community. MATLAB is designed to efficiently handle and performa calculations over matrices and lends itself nicely to the practical implementation of HodgeRank, a topologically-inspired algorithm whose ranking solution can be expressed in terms of matrices as described above. No claim is made that MATLAB represents the best choice of programming language in which to implement HodgeRank, simply that it is an appropriate choice.

The procedure for implementing HodgeRank is described below, including the additional steps required to measure the contribution of each pairwise comparison to local-scale inconsistency and re-weight the simplicial complex accordingly. These additional steps can be replaced in the case that a different choice is made to measure the contribution of pairwise comparisons to inconsistency in the ranking data, or skipped entirely if the user does not wish to incorporate inconsistency-based weights.

**Algorithm 1:** HodgeRank algorithm (including local-scale inconsistency weights)

**Input:** Skew-symmetric matrix $O \in \mathbb{R}^{n\times n}$ encoding pairwise comparisons between $n$ alternatives. Symmetric matrix $W \in \mathbb{R}^{n\times n}_{\geq 0}$ encoding initial weights for each comparison.

**Output:** HodgeRank score $\hat{s}$ measuring underlying preference for each alternative. Ranking of the $n$ alternatives.

Fix an orientation and indexing of the $m$ 1-simplices in the simplicial complex

Linearise $O$ as $\hat{O} \in \mathbb{R}^m$ and $W$ as $\hat{W} \in R^m$

**if** *re-weighting the complex by local-scale inconsistency measure* **then**

- Fix an orientation and indexing of the $p$ 2-simplices
- Compute $\delta_1$ and its adjoint $\delta_1^*$
- Compute $\delta_1$ and its adjoint $\delta_1^*$
- Calculate $\Delta_1^{up} = \delta_1^*\delta_1$
- Set $\hat{W} = \frac{1}{\Delta_1^{up}\hat{O}}$

Compute $\delta_0$ and its adjoint $\delta_0^*$

Calculate $\delta_0^*\hat{Y}$

Calculate $\Delta_0 = \delta_0^*\delta_0$

Compute the Moore-Penrose pseudo-inverse $\Delta_0^\dagger$

Calculate $\hat{s} = \Delta_0^\dagger\delta_0^*\hat{Y}$

Rank the $n$ alternatives by the rule $i \leq j \iff \hat{s}(i) \leq \hat{s}(j)$

Implementing HodgeRank for practical applications can require use of significant computational resources. In previous experiments the demand of the algorithm on the available computation resources limited the amount of ranking data that could be input into the algorithm. Therefore in order to facilitate meaningful applications of HodgeRank, it has been a requirement of this project to improve the efficiency of the algorithm.

There are three main issues which place a heavy burden on the available computational resources: the size of the simplicial complex, calculating the pseudoinverse, and iterating the algorithm. This project has gone some way to address these issues, reducing the computational burden of the algorithm

in the process. Nonetheless, more work remains to be done in this area.

From a computational perspective the worst case scenario is one in which every pair of alternatives has been compared, forming a complete graph. Given a set of $n$ alternatives, there are at most $\frac{n(n-1)}{2}$ pairwise comparisons and $\frac{n(n-1)(n-2)}{6}$ oriented triangles. The coboundary operators for such a simplicial complex are therefore $\delta_0 \in \mathbb{R}^{\frac{n(n-1)}{2} \times n}$ and $\delta_1 \in \mathbb{R}^{\frac{n(n-1)(n-2)}{6} \times \frac{n(n-1)}{2}}$. The size of these matrices increase exponentially with the number of alternatives, requiring greater memory and processing power.

Although these matrices can be large, the proportion of non-zero elements in them is small. Each row of $\delta_0$ corresponds to an oriented 1-simplex in the simplicial complex with

$$(\delta_0)_{ij} = \begin{cases} 1 & \text{if } j\text{-th oriented 1-simplex is } \{i,k\} \text{ for some } k \in V, \\ -1 & \text{if } j\text{-th oriented 1-simplex is } \{k,i\} \text{ for some } k \in V, \\ 0 & \text{otherwise.} \end{cases}$$

The number of non-zero elements in both $\delta_0$ and $\delta_0^*$ is therefore $2m \leq n(n-1)$ where $m$ is the number of directed edges in the complex. By a similar argument the number of non-zero elements in both $\delta_1$ and $\delta_1^*$ is $3p \leq \frac{n(n-1)(n-2)}{2}$ where $p$ is the number of oriented triangles in the complex.

Given that MATLAB requires 8 bytes to store each element in a matrix, storing an $r \times s$ matrix requires $8rs$ bytes. MATLAB can also store matrices as sparse matrices, concatenating the non-zero elements of each column into a single vector and storing this together with their row index and the number of non-zero elements in each column. This approach requires $8x + 16y$ bytes of memory where $x$ is the number of columns in the matrix and $y$ is the number of non-zero elements.

Given the sparsity of the coboundary operators, storing $\delta_0$ as a full matrix requires $8nm$ bytes against $8n+32m$ bytes as a sparse matrix, and storing $\delta_1$ as a full matrix requires $8mp$ bytes against $8m+48p$ bytes as a sparse matrix. Employing sparse matrices therefore reduces the memory requirements of the HodgeRank algorithm for sufficiently large $n, m, p$ ($\geq 7$). For a pairwise comparison matrix of 10,000 alternatives, in the worst case scenario where

the underlying graph is complete, storing $\delta_0$ requires 4,000 GB as a full matrix but just 0.04% of this (1.6GB) as a sparse matrix. The sparse matrix representation of $\delta_1$ is less than $2 \cdot 10^{-5}\%$ of the full matrix representation, although this is still large at 8,000 GB.

Fortunately in most applications of HodgeRank the underlying graph is likely far from complete. Nonetheless, the above example illustrates the significant improvement made by employing sparse matrices in the algorithm, reducing the computational requirements drastically. This reduction made an analysis of local-scale inconsistency, and subsequent re-weighting of the simplicial complex, feasible in a case study involving pairwise comparisons on up to 35,000 alternatives.

Another problematic issue in implementing the HodgeRank algorithm is computing the Moore-Penrose pseudoinverse of the 0-dimensional combinatorial Laplacian, a necessary step in solving the weighted least squares problem given by (13). This step proves to be the largest single bottleneck in the algorithm, accounting for over 80% of its run-time.

MATLAB includes a built-in function, *pinv*, for computing the Moore-Penrose pseudoinverse of a matrix using singular value decomposition (SVD). The complexity of computing the pseudoinverse of an $m \times n$ matrix using this approach is $O(mn \cdot \min(m, n))$. Given that the 0-dimensional combinatorial Laplacian is an operator from $C^0$ to itself, computing its pseudoinverse has complexity $O(n^3)$.

In their paper, Chen & Feng (2014)[p.183] identify QR decomposition as the "best choice for computing the Moore-Penrose pseudoinverse." Although the complexity of the QR decomposition approach is also $O(n^3)$, empirical evidence suggest this approach requires fewer floating point operations per second (FLOPS) than the SVD approach, and runs between 2-3 times faster.

In this project, the QR decomposition approach has been followed however, instead of computing the Moore-Penrose pseudoinverse of $\Delta_0$, a Moore-Penrose pseudoinverse factorisation object, PIF[1], has been created such that $\Delta_0^\dagger x = PIF(\Delta_0)x$. Whilst not directly computing the pseudoinverse, this

[1]B. Luong (2009). Pseudo-inverse (https://www.mathworks.com/matlabcentral/fileexchange/25453-pseudo-inverse), MATLAB Central File Exchange

technique is sufficient for the purpose of solving the weighted least squares problem. Implementing the HodgeRank algorithm in a case study, this project did indeed see the run-time of the algorithm halve by following the QR decomposition method.

MATLAB also provides an in-built function, *lscov*, for solving the weighted least squares problem without the need for computing the pseudoinverse of $\Delta_0$. This function uses a variant of QR decomposition to return the weighted least squares solution to $Ax = b$. By following this approach and directly solving the weighted least squares problem given by (4) the HodgeRank algorithm can be simplified. Although this project has not employed this approach to solving the weighted least squares problem, simulations suggest that it reduces run-times by $\sim 12\%$. The algorithm detailed below is recommended for future implementations of HodgeRank.

**Algorithm 2:** Improved HodgeRank algorithm (including local-scale inconsistency weights)

**Input:** Skew-symmetric matrix $O \in \mathbb{R}^{n\times n}$ encoding pairwise comparisons between $n$ alternatives. Symmetric matrix $W \in \mathbb{R}^{n\times n}_{\geq 0}$ encoding initial weights for each comparison.

**Output:** HodgeRank score $\hat{s}$ measuring underlying preference for each alternative. Ranking of the $n$ alternatives.

1 Fix an orientation and indexing of the $m$ 1-simplices in the simplicial complex

2 Linearise $O$ as $\hat{O} \in \mathbb{R}^m$ and $W$ as $\hat{W} \in R^m$

3 **if** *re-weighting the complex by local-scale inconsistency measure* **then**

4 Fix an orientation and indexing of the $p$ 2-simplices

5 Compute $\delta_1$ and its adjoint $\delta_1^*$

6 Compute $\delta_1$ and its adjoint $\delta_1^*$

7 Calculate $\Delta_1^{up} = \delta_1^* \delta_1$

8 Set $\hat{W} = \frac{1}{\Delta_1^{up}\hat{O}}$

9 Compute $\delta_0$

10 Compute $\hat{s} = lsvoc(\delta_0, \hat{O}, \hat{W})$

11 Rank the $n$ alternatives by the rule $i \leq j \iff \hat{s}(i) \leq \hat{s}(j)$

The last significant issue faced in implementing HodgeRank is that practical applications often require many iterations of the algorithm. In the case study undertaken by this project, the evolving underlying preference of alternatives is measured across time by inputting a sliding window of the available ranking data into the HodgeRank algorithm. This requires running the algorithm multiple times, once for each time the window slides (the case study requires over 1000 iterations). Given the high memory requirements and long run-time of the algorithm, performing multiple iterations of the algorithm on a single computing node can quickly overcome the available resources.

Parallel processing can be applied to overcome this issue since each iteration of the algorithm can be performed independently. This project has employed the IRIDIS High Performance Computing Facility[2], specifically IRIDIS 4 consisting of 750 compute nodes, each with dual 2.6GHz processors, 16 processor cores and 64GB of memory. In addition, IRIDIS 4 is equipped with four high memory nodes, each with four 2.4GHz processors, 32 processor cores and 252GB of memory. In the case study, each iteration of the algorithm was run on pairwise comparison matrices of around 25,000 alternatives which were approximately 98.5% sparse. Including the local-scale inconsistency re-weighting, each iteration of the algorithm required around 52GB of memory and 4.5 hours of run-time.

The improvements made to the HodgeRank algorithm have reduced the computational resources required, facilitating its application in practical problems as illustrated in the case study below. There remains scope for improving the efficiency of the algorithm, in particular the choice of programming language (C/C++ may be more appropriate choices) and method for solving the weighted least squares problem. Increasing the efficiency of the algorithm above its current level is however beyond the scope of this project which has focused on making necessary improvements to practically implement HodgeRank.

[2]https://cmg.soton.ac.uk/research/categories/computational-platforms/iridis/

# 5 Case Study: How well does the market handle inconsistent information?

Inconsistent ranking data poses a challenge to the decision making ability of individuals, increasing the complexity of decision problems and the likelihood that these decisions will not reflect all of the available information (Tversky 1969, Tversky & Kahneman 1992, Luce & Raiffa 1957). Techniques such as HodgeRank can be used to analyse and account for inconsistencies within ranking data, however doing so is computationally expensive.

When faced with complex mental tasks which strain their available resources, individuals often employ heuristics to simplify and reduce the cognitive load of these tasks (Tversky & Kahneman 1974, Slovic & Lichtenstein 1971). This project examines how well decision makers handle and exploit inconsistent ranking data in a real world scenario, balancing the competing desires to improve decision making and reduce cognitive strain.

Competitive events arise when participants compete against each other for rewards. Notable examples of competitive events include political elections and sporting events (Lessmann et al. 2012). A variety of approaches have been taken to predict the outcome of competitive events with varying levels of success, however a large degree of uncertainty remains reflecting the underlying randomness of the event. This randomness leads to inconsistent outcomes of events, producing conflicting data which individuals attempt to use in their decision making.

The unpredictable nature of competitive events gives rise to associated speculative markets. Participants in these markets place wagers on outcomes based on their assessment of the likelihood of each outcome. These speculative markets therefore function as prediction markets for the competitive events, aggregating the predictions of their participants with market prices quantifying the predictions and decisions of individual bettors (Figlewski 1979, Asch et al. 1982). These decisions can then be easily verified against the observed outcome of an event, resulting in a payoff for the bettor if their prediction was correct.

Competitive events are a natural setting to study the impact of incon-

sistent information in decision making and financial markets with a high expectation of conflicting data, quantifiable decisions and verifiable market prices. They are especially appealing for studying issues regarding market efficiency as "wagering markets are especially simple financial markets, in which the scope of the pricing problem is reduced" (Sauer 1998)[p. 2021].

Horse racing markets are considered highly semi-strong efficient with prices quickly and effectively incorporating all available information (Sung & Johnson 2008, Edelman 2007). A large body of work exists examining the informational efficiency of these markets, providing a framework for assessing whether inconsistent data is fully accounted for in market prices.

These markets are valuable case studies of market efficiency, exhibiting similar features to stock markets (Gabriel & Marsden 1990, Ali 1998) yet providing "a clear view of pricing issues which are more complicated elsewhere" (Sauer 1998)[p. 2021]. Moreover they are of economic significance themselves with an estimated turnover of £11 billion in 2016.

This case study evaluates the impact of inconsistent ranking data produced from past performance data on the semi-strong efficiency of parimutuel win markets surrounding horse racing. Results from 6 years (2009 to 2014 inclusive) of UK horse racing betting markets is analysed to test the impact of inconsistent ranking data on decision making in financial markets. The data consists of 54,346 races across all 34 UK racetracks competed in by 59,719 horses.

If the winning probabilities, encoded as market prices, for competing horses estimated by bettors can be improved by including a variable output by the HodgeRank algorithm, then the inconsistent data, has not been fully accounted for in the decisions made by bettors. The economic cost of a market inefficiency relating to inconsistent information is estimated as the excess returns made by including the HodgeRank variable in an existing wagering strategy employing market prices.

This case study makes several contributions to the existing literature, exploring the effects of inconsistent information on decision making and market efficiency in a real world setting. Although the study has focused exclusively on parimutuel horserace win markets, their similarity to wider financial mar-

kets suggest that the findings are likely to be replicated in other settings with access to inconsistent data, examples of which include other prediction markets and understanding consumer choice behaviour.

i) **Topological methods**: This case study represents a novel application of the HodgeRank algorithm to modelling and understanding information in financial markets. The application of a topologically-inspired technique is shown to have merit, capturing information which markets themselves have failed to.

   Not only does this demonstrate how the study of financial markets can benefit from the adoption of topological and network based methods, but it also demonstrates the value of this approach to data analysis. The succesful application of HodgeRank to parimutuel horserace win markets adds to the body of evidence indicating that these methods provide new insights into understanding data.

ii) **Inconsistent data**: The inherent randomness of competitive events often produces conflicting information. This data is generally regarded as problematic, often considered either noise or erroneous, and few studies have sought to extract information from them.

   This project explores the effects of inconsistent data on decision making and market efficiency and shows that valuable information can be contained within such data. Combining a more sophisticated version of the HodgeRank algorithm with statistical forecasting methods and wagering strategies, this case study develops a technique to understand and exploit this data.

iii) **Cognitive errors**: Decision making is known to be affected by the complexity of the decision problem which in turn is affected by the ease with which conclusions can be drawn from the available data. Inconsistent data poses a challenge to decision makers, representing multiple real world states and requiring careful consideration.

   This case study demonstrates that this challenge is not fully met by bettors in parimutuel horserace win markets whose decisions are af-

fected by the presence of inconsistent data and fail to fully encapsulate the information contained within inconsistencies. These failings represent deviations from models of rational decision making and represent cognitive errors on behalf of participants in wagering markets.

iv) **Market inefficiencies**: Horserace wagering markets are widely considered highly semi-strong efficient and security prices are expected to reflect most of the available information. The case study shows that inconsistent data is not adequately accounted for by market prices and that, by both separating out consistent data and exploiting inconsistent data, excess risk-adjusted returns are achievable.

Although the findings of this case study run contrary to the common consensus regarding the efficiency of wagering markets, a number of studies have highlighted inefficiencies in wagering markets and the opportunity for abnormal returns. The application of a more sophisticated version of HodgeRank, together with conditional logit models and Kelly wagering strategies, exposes a previously unknown market inefficiency. Given the similarity between wagering markets and wider financial markets, the existence of this inefficiency raises questions about the extent to which inconsistent data is accounted for by financial markets in general.

## 5.1 Conditional Logit Modelling of Decision Making

Participants in speculative markets make decisions, ranging from investing in securities to wagering on sporting events, by assessing the likelihood of potential outcomes occurring. Modelling this decision making behaviour is a complex task, complicated by the number of factors under consideration and the difficulty in observing decision makers without influencing their behaviour.

Conditional logistic regression is a popular choice for modelling decision making in areas including demography (Radner & Miller 1970, Boskin 1974, Davies et al. 2001, Hoffman & Duncan 1988), transportation demand (Ben-Akiva & Lerman 1985, McFadden 1974) and consumer behaviour (Berry

1994, Berry et al. 1995). A key feature of these conditional logit models is that they directly account for competition between alternatives (McFadden 1973). They have been extensively applied across studies of horseracing demonstrating their suitability for modelling decision making and security pricing in parimutuel horse race win markets (Bolton & Chapman 1986, Figlewski 1979, Benter 1993, Johnson et al. 2006).

A **conditional logit model** estimates the probability that each of the $n_i$ alternatives competing in the $i^{\text{th}}$ selection event will be the sole alternative selected by the decision maker of that event (McFadden 1973). These probabilities are derived from the **representative utility**, the preference of the decision maker for each alternative (equivalently the competitiveness of each alternative). The representative utility of alternative $j$ during the $i^{\text{th}}$ selection event is given by

$$\mu_{ij} = V_{ij} + \epsilon_{ij}$$

where $V_{ij}$ is the systematic component, reflecting known or observed information about the competitiveness of the alternative, and $\epsilon_{ij}$ is a term accounting for randomness in the behaviour of the decision maker. Often the systematic component is modelled as a linear combination of $m$ characteristics

$$V_{ij} = \sum_{k=1}^{m} \beta_k x_{ijk}$$

where the $\beta$ are coefficients and $x_{ijk}$ is the evaluation of the $k^{\text{th}}$ characteristic of alternative $j$ for the $i^{\text{th}}$ selection event as perceived by the decision maker for the event.

A rational decision maker will select the alternative with the highest representative utility and so the probability that alternative $j$ will be selected during the $i^{\text{th}}$ is

$$P_{ij} = P[\mu_{ij} > \mu_{ih}] \quad \forall h \neq j$$

This probability is conditional upon the random $\epsilon$ terms and is given by

$$\begin{aligned} P_{ij|\epsilon_{ij}} &= P[V_{ij} + \epsilon_{ij} > V_{ih} + \epsilon_{ih}] \quad \forall h \neq j \\ &= P[V_{ij} - V_{ih} + \epsilon_{ij} > \epsilon_{ih}] \quad \forall h \neq j \end{aligned}$$

Marschak (1960) showed that the selection probabilities can be evaluated by assuming the random $\epsilon$ terms are identically and independently distributed according to the Gumbel (or generalised extreme value type-I) distribution. Under this assumption the probability distribution function of $\epsilon$ is given by

$$p(\epsilon_{ij}) = e^{-\epsilon_{ij} - e^{-\epsilon_{ij}}}$$

and the cumulative probability by

$$P[\epsilon_{ij} \leq \alpha] = e^{-e^{-\alpha}}$$

It follows that the probability of an alternative being selected, conditional upon the $\epsilon$ terms, is

$$\begin{aligned} P_{ij|\epsilon_{ij}} &= \prod_{h \neq j} e^{-e^{(V_{ih} - V_{ij} - \epsilon_{ij})}} \\ &= \prod_{h=1}^{n_i} e^{-e^{(V_{ih} - V_{ij} - \epsilon_{ij})}} \cdot e^{e^{-\epsilon_{ij}}} \end{aligned}$$

Applying Bayes' theorem, the probability that the $j^{\text{th}}$ alternative will be selected during selection event $i$ is

$$
\begin{aligned}
P_{ij} &= \int_{-\infty}^{\infty} P_{ij|\epsilon_{ij}}\, p(\epsilon_{ij})\, d\epsilon_{ij} \\
&= \int_{-\infty}^{\infty} \prod_{h=1}^{n_i} e^{-e^{(V_{ih}-V_{ij}-\epsilon_{ij})}} \cdot e^{e^{-\epsilon_{ij}}} \cdot e^{(-\epsilon_{ij}-e^{-\epsilon_{ij}})}\, d\epsilon_{ij} \\
&= \int_{-\infty}^{\infty} \prod_{h=1}^{n_i} e^{-e^{(V_{ih}-V_{ij}-\epsilon_{ij})}} \cdot e^{-\epsilon_{ij}}\, d\epsilon_{ij} \\
&= \int_{-\infty}^{\infty} e^{-\epsilon_{ij}} \cdot e^{(-e^{-\epsilon_{ij}})(e^{-V_{ij}})(\sum_{h=1}^{n_i} e^{V_{ih}})}\, d\epsilon_{ij} \\
&= \left. \frac{-e^{(-e^{-\epsilon_{ij}})(e^{-V_{ij}})(\sum_{h=1}^{n_i} e^{V_{ih}})}}{(e^{-V_{ij}})(\sum_{h=1}^{n_i} e^{V_{ih}})} \right|_{-\infty}^{\infty}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{(e^{-V_{ij}})(\sum_{h=1}^{n_i} e^{V_{ih}})} \\
&= \frac{e^{V_{ij}}}{\sum_{h=1}^{n_i} e^{V_{ih}}}
\end{aligned}
$$

Although the representative utilities are a combination of systematic and random components, by assuming the random components are identically and independently distributed according to the Gumbel distribution, the probabilities of an alternative being selected are dependent only upon the systematic component of the representative utilities of those alternatives participating in the selection event.

There are several assumptions and features of conditional logit models which must be validated in the context of any potential application. It is only appropriate to employ conditional logit modelling if all of these conditions are met:

(i) **Single Selection:** Exactly one alternative is selected in each selection event.

(ii) **Rational Decision Maker:** In each selection event, the decision maker selects the alternative deemed most preferable or competitive.

(iii) **Dependent Probabilities:** Probabilities of alternatives being selected depend only on their competitiveness (representative utility) and the competitiveness of their competitors, as perceived by the decision maker of the selection event.

(iv) **Independence of Irrelevant Alternatives:** The probability ratio between any pair of alternatives in the same selection event

$$P_{ij}/P_{ik} = e^{V_{ij}}/e^{V_{ik}}$$

is independent of any other competing alternative (Hausman & McFadden 1984).

An important implication of this condition is that including or removing an alternative from a selection event does not permute the pre-existing order of preference. If alternative $j$ is preferred to alternative $k$ in a selection event, the inclusion or removal of another alternative does not affect this pairwise relation during the event.

(v) **Identical and Independent Distribution:** Any two random components of representative utility are identically distributed

$$P[\epsilon_{ij} \leq \alpha] = P[\epsilon_{pq} \leq \alpha] \quad \forall \alpha$$

and independently distributed

$$P[(\epsilon_{ij} \leq \gamma) \wedge (\epsilon_{pq} \leq \gamma)] = P[\epsilon_{ij} \leq \gamma] \cdot P[\epsilon_{pq} \leq \gamma] \quad \forall \gamma$$

(vi) **Gumbel Distribution:** The random components of representative utility follow the Gumbel (or generalised extreme value type-I) distribution, likely being small in magnitude and more likely to improve representative utility than reduce it.

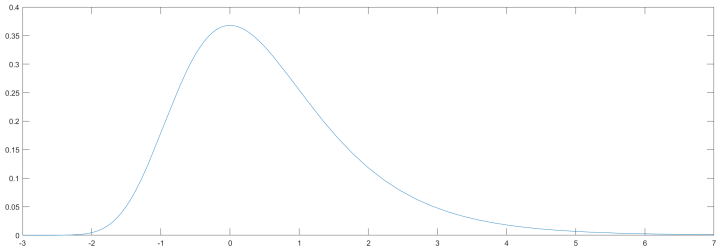


Figure 15: Probability Density of Random Component

This case study considers each race to be a selection event in which the race 'chooses' a single, most preferable horse as the winner. This framework for modelling the outcomes of horse races satisfies conditions (i) and (ii) of conditional logit modelling. Each race is a competitive event in which the outcome depends upon the relative competitiveness of the participating horses and there is minimal interaction between competing horses, hence conditions (iii) and (iv) of conditional logit modelling are also satisfied.

The remaining two conditions govern the random components of representative utility, namely that they are identically and independently distributed according to the Gumbel distribution. The inherent unpredictability of horse racing means that these conditions are unlikely to be unequivocally hold, however Henery (1983) demonstrated that they do hold in general. Noting that the problem of estimating winning probabilities is otherwise intractable (Bolton & Chapman 1986, Figlewski 1979), horserace studies have deemed that these last two conditions are sufficiently met as to not invalidate the choice of conditional logit models (Ali 1998, Benter 1994, Bolton & Chapman 1986, Johnson et al. 2006, Sung et al. 2005).

Given the uncertainty of horse races, conditional logit models are widely regarded as the most suitable choice for modelling decision making and pricing securities in parimutuel horserace win markets (McFadden 1973, Bolton & Chapman 1986, Benter 1993, Johnson et al. 2006, Sung et al. 2005). Their extensive application in the field has proven highly successful in demonstrating both a high degree of semi-strong efficiency (Figlewski 1979, Hausch & Ziemba 1990, Snyder 1978, Canfield et al. 1987) and evidence of specific market inefficiencies (Bolton & Chapman 1986, Benter 1993, Johnson

et al. 2006, Sung et al. 2009).

## 5.2 Evaluating Conditional Logit Models

Individual choice behaviour can be modelled by conditional logit models, estimating the probability that each alternative will be selected in a selection event. Comparing predictions of outcomes made by the model against the observed outcomes, the accuracy of the conditional logit model can be evaluated.

The **likelihood** of a sample of observed outcomes is the probability that the observations would occur within the conditional logit model. The greater the likelihood of the data sample, the more accurately the behaviour of the decision makers behind those selection events has been modelled.

A parameterised conditional logit model, in which some or all parameters of the systematic component of representative utility are unknown, can be fitted to a sample of observed outcomes by employing maximum likelihood estimation (MLE) techniques. These techniques estimate parameters which maximise the likelihood of the sample of observed outcomes being produced by the model (McFadden 1973).

Often it can be more convenient to consider the **log-likelihood** of a sample of observed outcomes arising from a given model $\hat{\theta}$. The log-likelihood of a sample of observations in a conditional logit model $\hat{\theta}$ is given by

$$L(\hat{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_j} \delta_{ij} \ln P_{ij}$$

where $P_{ij}$ is the probability given by the model that alternative $i$ is selected in event $j$ and

$$\delta_{ij} = \begin{cases} 1 & \text{if alternative } i \text{ is selected during event } j, \\ 0 & \text{otherwise} \end{cases}$$

The logarithm function is strictly increasing and so likelihood is maximised when log-likelihood is. In practical implementations, it is often easier to

maximise the log-likelihood of a data sample and so maximum log-likelihood estimation techniques are often employed to fit conditional logit models to observed outcomes (Hosmer Jr. et al. 2013).

When fitting a parameterised conditional logit to a training sample of observed outcomes, the objective is to choose coefficients $\hat{\boldsymbol{\beta}}$ for the explanatory variables $\boldsymbol{x}$ which maximise the log-likelihood. This is equivalent to choosing $\hat{\boldsymbol{\beta}}$ which minimise the score function $-L(\hat{\boldsymbol{\beta}})$. Maximum log-likelihood estimation techniques differentiate this score function and identify critical points where every partial derivative is 0. By evaluating the Hessian at these points, a matrix of second-order partial derivatives, it can be determined whether they are a local minimum of the score function (at a minimum, every element of the Hessian is negative) (Myung 2003).

Standard errors for these estimates are given by the inverse of the square roots of the diagonal terms of the Hessian matrix. Intuitively, the more curved the score function is at the minimum, the more certainty there is in the estimated coefficient (Gill & King 2004, Davidson & MacKinnon 2004).

A conditional logit model which has been fitted to a training sample of observed outcomes has had its parameters optimised to maximise the likelihood of those observed outcomes being produced by the model. This does not however mean that the model is a good representation of the underlying processes producing the observed outcomes. It is therefore important to cross-validate the model and evaluate its performance on samples of observed outcomes which are independent from the training sample. There are several tests and statistics available to assess whether a model is correctly specified and measure the accuracy of its predictions.

### 5.2.1 Overfitting

Conditional logistic regression models the representative utility of an alternative in a selection event as a linear combination of explanatory variables $(\boldsymbol{x_0}, ..., \boldsymbol{x_k})$. It is always possible to include additional explanatory variables in a conditional logit model, resulting in an increased likelihood of the training data being produced by the model since the corresponding coefficient can

be always set to zero. This larger model may however suffer from overfitting where it fits the training data too closely, providing a worse representation of the underlying processes generating the observations, resulting in a lower likelihood of other data samples being produced by the model.

One approach to testing for overfitting is the **Wald coefficient test**, a hypothesis test establishing whether each explanatory variable is correlated with the observed outcomes and thus whether its coefficient in the model is non-zero. For each explanatory variable, the null and alternative hypotheses are given by:

$H_0$: The variable is uncorrelated with the observed outcomes.

$H_1$: The variable is correlated with the observed outcomes.

If there is insufficient evidence to reject the null hypothesis, the coefficient corresponding to the variable should be zero. On the other hand, if the null hypothesis is rejected in favour of the alternative hypothesis, the coefficient should not be trivial.

The Wald test can be conducted to determine whether there is sufficient evidence to reject the null hypothesis. The test statistic for the $i$-th explanatory variable is given by

$$\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

with the standard errors derived from the Hessian matrix of the log-likelihood function at the point $\hat{\boldsymbol{\beta}}$, and approximates a normal distribution with $\mu = 0$ and $\sigma = 1$ (Davidson & MacKinnon 2004). If this exceeds the critical value of the chosen significant level, there is evidence that the variable is correlated with the observed outcomes and its coefficient should not be trivial.

The **Log-Likelihood Ratio (LLR) test** is another method testing for overfitting. Given a conditional logit model $\hat{\theta}_1$ with explanatory variables $(x_1, ..., x_k)$, a larger model $\hat{\theta}_2$ can be found by including additional explanatory variables. The smaller model can be recovered by setting the coefficients of the additional explanatory variables to 0 and so is nested within the larger model.

A hypothesis test can be conducted to determine whether there is information contained in the additional explanatory variables which is not captured by the smaller, nested model. The null and alternative hypotheses are given by:

$H_0$: All the information contained in the larger model $\hat{\theta}_2$ which includes additional explanatory variables is captured by the smaller, nested model $\hat{\theta}_1$.

$H_1$: There is information contained in the larger model $\hat{\theta}_2$ not captured by the smaller, nested model $\hat{\theta}_1$ which excludes the additional explanatory variables.

The test statistic for this hypothesis test is derived from the log likelihoods of the observations in both models and is given by

$$\lambda = 2(L(\hat{\theta}_2) - L(\hat{\theta}_1))$$

where $L(\theta_1)$ is the log likelihood of the testing sample in the nested model and $L(\theta_2)$ is the log likelihood in the larger model. Wilks (1938) demonstrated that the log likelihood ratio test statistic approximates a chi-squared distribution with degrees of freedom equal to the number of additional non-trivial parameters in the larger model. If the log likelihood ratio test statistic is statistically significant, there is evidence that the larger model contains information which is missing from the smaller model and that the additional variables, as a collective and not necessarily individually, improve the model.

### 5.2.2 Goodness Of Fit

Goodness of fit measures are often provided for statistical models, describing how well the model fits the observed data. These measures are generally a summary of the residuals, distances between observed values and the corresponding value expected by the model.

For continuous outcomes the residual is defined as the difference between the observed value and the expected value. The expected value produced by

conditional logit models is a continuous variable on $[0, 1]$, the probability of an alternative being selected during an event, however the observed outcome is binary, taking a value of 1 if the alternative is selected and 0 otherwise. The residuals are therefore given by $r_{ij} = \delta_{ij} - P_{ij}$. The residual plot associated with these residuals is difficult to interpret, giving two straight lines, one for each of the possible outcomes.

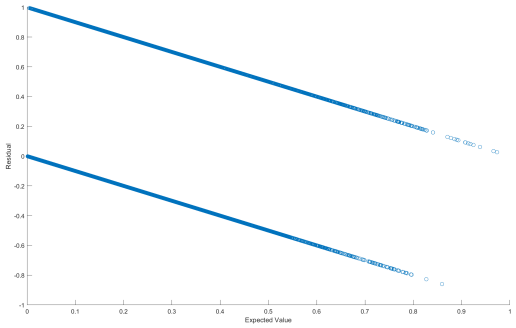


Figure 16: Residual Plot of a Conditional Logit Model

To move towards a meaningful interpretation of the residuals, a **binned residual plot** can be produced with individual residuals being binned together based on their expected value (Gelman & Hill 2007). Typically $\sqrt{n}$ bins of approximately equal size are used where $n$ is the number of observations in the sample. The average residual of each bin can be shown on a residual plot, along with $\pm 2$ standard error confidence intervals given by $2\sqrt{p(1-p)/n}$.

The binned residual plot is a method for visualising and interpreting the residuals of a conditional logit model. The number of residual bins with average residual outside of the confidence intervals can be counted. If the model is a good fit for the data, about 95% of the residual bins should fall within the confidence intervals.

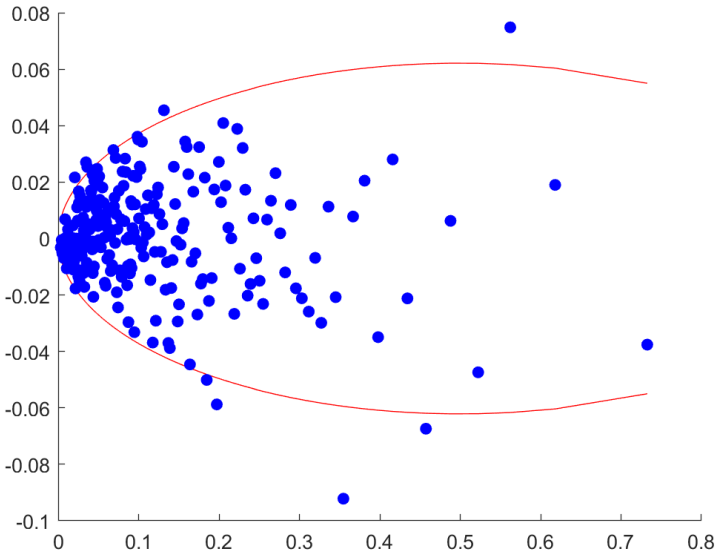


Figure 17: Binned Residual Plot of a Conditional Logit Model

The binned residual plot is a method for evaluating the absolute fit of a conditional logit model. A conditional logit model can also be evaluated in terms of its relative fit compared to other models. The **McFadden $R^2$** is a statistic which measures the goodness of fit of a model in comparison to the smallest nested model, the trivial model $\hat{\theta}_0$.

$$\widetilde{R}^2 = 1 - \frac{L(\hat{\beta})}{L(0)} \tag{25}$$

Every coefficient in the trivial model is set to 0 and so every alternative has an equal probability of being selected in each event.

This measure attempts to mimic the coefficient of determination, with better models achieving a $R^2$ closer to 1. However, unlike the coefficient of determination, $\widetilde{R}^2$ can be negative if the model is worse than the trivial model (this can only occur when the model is evaluated on a holdout sample of observations).

Another difference with the coefficient of determination is that $R^2$ is not clearly interpretable with several alternative pseudo-$R^2$'s have been proposed (Sung et al. 2016). Despite this, increases in pseudo-$R^2$'s are indicative of im-

provements in the model and increases in its predictive accuracy (Lessmann et al. 2012).

When presented with a set of candidate models and asked to determine the model which best fits a set of observed data, evaluating goodness of fit by the McFadden $\widetilde{R}^2$ is equivalent to choosing the model with the maximal log-likelihood. As discussed in the context of the Wald coefficient and LLR tests, larger models often have an increased log-likelihood due to the inclusion of additional, not always relevant, explanatory variables.

The **Akaike Information Criteria (AIC)** is a measure of goodness of fit which penalises models with more explanatory variables. The test statistic is given by

$$AIC(\hat{\theta}) = 2k - 2L(\hat{\theta})$$

where $k$ is the number of independetly adjusted parameters in the model.

Unlike the overfitting tests described above, AIC can be used to compare non-nested models. Given a set of candidate models, the model with minimal AIC value is considered the best fit for the data. This does not mean that the chosen model is a good fit for the data (this must be determined via other methods), simply that it is the best fit.

## 5.3 Economic Modelling of Decision Making

More accurately estimating the likelihood of potential outcomes allows individuals to make decisions which have a greater chance of producing desirable effects. Whilst this can improve decision making under uncertainty, the primary motivation of decision makers in speculative markets is not to make accurate decisions but rather to make profitable ones. Thus an evaluation of decision making in speculative markets must assess both its statistical accuracy and economic impact.

The decision makers under consideration operate in a parimutuel market, in which prices are set by the decisions made by its participants, and have access to the information contained within the dataset. Thus if decision makers in the speculative markets surrounding horse racing fully account for both the consistent and inconsistent information in this dataset, the market

itself should be semi-strong efficient with regards to this dataset.

In line with techniques for testing semi-strong form efficiency in financial markets, an economic evaluation of decision making is conducted by assessing the potential for achieving abnormal returns by incorporating information from the consistent and inconsistent parts of the dataset. If individuals make use of all the information contained within the dataset, including its inconsistent parts, it should not be possible to consistently outperform the market (Bolton & Chapman 1986, Basu 1977).

### 5.3.1 Kelly Wagering Strategy

Decision makers in parimutuel win markets are presented with a range of alternatives participating in selection events and can invest a portion of their capital in each alternative, with these investment decisions paying dividends if the alternative is the single alternative selected during the event. Each alternative is assigned odds of $\gamma_i - 1 : 1$ by the market reflecting the perceived probability of the alternative being selected and a successful wager returns a multiple of $\beta_i$ of the initial wager. These odds correspond with the estimated winning probabilities of each alternative, given by

$$p_i = \frac{\sum_j {}^1/_{\gamma_j}}{\gamma_i}$$

A range of strategies can be employed to determine the level of capital, if any, that should be invested in each alternative, guided by a variety of underlying principles. Naive wagering strategies include level stake betting, which assumes that outcomes of selection events are random and places an equal wager on each of the competing alternatives, and proportional stake betting, which assumes the market odds accurately reflect the real selection probabilities and places a sufficient wager on each alternative to earn a fixed return. It is possible for level stake betting to make a profit if the market performs worse than random guessing and for proportional stake betting to be profitable if the sum of the market perceived probabilities is less than 1, however it has been shown that these conditions are not met in practice and

that these naive strategies are unprofitable (Sung & Johnson 2010, Figlewski 1979, Bolton & Chapman 1986).

The Kelly wagering strategy maximises the long term rate of growth of capital (Kelly 1956). Given that the logarithm function is strictly increasing, this is achieved by maximising

$$G = \lim_{n\to\infty} \frac{1}{n} \ln \frac{v_n}{v_0}$$

where $v_0$ is the starting capital of the decision maker and $v_n$ is their capital after the $n^{\text{th}}$ selection event. Let $m_n$ be the number of competing alternatives in event $n$, $v_n(i)$ be the fraction of capital wagered, and $\gamma_n(i)$ be the return per unit of capital. It follows that if alternative $i$ is selected during event $n$ then the current capital grows by a rate of $1 + v_n(i)\gamma_n(i) - \sum_{j=1}^{m_n} v_n(j)$.

Breiman (1961) showed that maximising the expected log return from each selection event

$$g(v_n) = \sum_{i=1}^{m_n} p_n(i) \ln \left( 1 + v_n(i)\gamma_n(i) - \sum_{j=1}^{m_n} v_n(j) \right) \tag{26}$$

is asymptotically optimal for maximising the rate of growth of capital. Kelly strategies wager a proportion of available capital on each alternative in a selection event, placing larger wagers on alternatives with greater estimated winning probabilities.

Although the Kelly wagering strategy maximises long term capital growth, it is accompanied by a increasing risk of catastrophic loss of capital through a series of unsuccessful wagers (Maclean et al. 2010). Implementations of Kelly strategies seek to mitigate this risk by wagering a portion of available capital on each selection event (Benter 1993, Johnson et al. 2006).

The Kelly wagering strategy is considered an optimal technique for wealth creation, outperforming other strategies for capital growth (Maclean et al. 2010, MacLean et al. 2011). It has been successfully employed in horseracing literature for seeking abnormal returns when a market ineffiency is believed to exist (Thorp 2011, Sung et al. 2005, Lessmann et al. 2009, Johnson et al. 2006)

## 5.4 Methodology

The effect of inconsistent ranking data on decision making in financial markets is analysed in the context of UK horse racing parimutuel win markets. Results are analysed from all UK racetracks over a six year period from 2009 to 2014 inclusive. There are 436,709 observations of 59,719 horses competing in 54,346 races in the dataset, and a pairwise comparison matrix is constructed by aggregating the relative performance of pairs of horses in each race.

Bettors place wagers on horses when they believe the market odds undervalue their probability of winning a race. The market odds are updated with this new information, providing new winning probabilities for the competing horses. Thus the final market odds reflect the winning probabilities of competing horses, as perceived by bettors as a population, and quantitatively encode their decisions.

If two competing horses are respectively assigned odds of $\alpha : 1$ and $\gamma : 1$ by the market, then bettors believe that the first horse is $\alpha/\gamma$ more likely to win the race. It is often advantageous to transform the market odds into the natural logarithm of the probability implied by them (i.e. odds of $\gamma - 1 : 1$ become $\ln(1/\gamma)$), an additive variable encoding the decisions made by bettors.

The additive version of the market odds can be used as a predictor of the outcome of the race. The accuracy and profitability of these predictions can be considered measures of the quality of decision making. If better predictions can be made by incorporating information extracted from consistent and inconsistent data, then this data has not been fully accounted for in bettors' decision making.

Horses are often handicapped to facilitate a more competitive betting environment. "The essence of handicapping is a well-tried proposition that the weight a horse carries ultimately affects the speed at which is can gallop" (*A Guide to Handicapping* 2014, p.3). Therefore the observed results of races measure a combination of the competitiveness of the horse and the impact of weight on its performance.

Handicappers assess the quality of horses and assign weights to reduce

their performance in a race according to the following table:

| Distance (furlong) | Handicap (lbs/length) |
|---|---|
| 5 or less | 3 |
| 6 | 2.5 |
| 7-8 | 2 |
| 9-10 | 1.75 |
| 11-13 | 1.5 |
| 14 | 1.25 |
| 15 or more | 1 |

Table 1: Handicapping Formula

Consider a seven furlong race in which the handicappers expect horse A to finish five lengths ahead of horse B. Since the handicapping value for a race of this distance is 2 lbs/length, by allotting 2.5 lbs of additional weight to A, the handicappers expect that it will finish level with B. If instead 3 lbs of weight was allocated to A, it would be expected to finish one length behind B.

Assuming the validity of this formula, it is possible to estimate the results of the race had every horse carried the same weight. This is done by multiplying t he weight carried by the relevant handicap and subtracting this value from the observed beaten lengths. Where a race falls in between distance categories, a proportional handicap value is used.

These adjusted results provide more accurate information about the underlying quality of the competing horses than the unadjusted results which have been deliberately skewed. Throughout this case study, the adjusted results are used to measure the performance of a horse in a race.

Two explanatory variables are derived from pairwise comparisons formed from past performance data, one of which exploits information in the consistent ranking data and another which accounts for information in both the consistent and inconsistent ranking data. For each race, past performance data from the three years prior is compiled into a pairwise comparison matrix by taking a recency weighted average of the pairwise comparisons between pairs of horses who compete against each other in the same race. Each race is assigned a recency weight in accordance with the time between the race

and the current day, with the recency weight of race $n$ which occurred $d_n$ days ago given by

$$w^n = exp(\frac{-d_n}{h})$$

where $h$ is the half-life of the information provided by races.

Let $\zeta_i^n$ be the performance of horse $i$ in the $n^{\text{th}}$ race (which is zero if the horse did not compete in the race) and

$$\delta_{ijn} = \begin{cases} 1 & \text{if horses } i \ \& \ j \text{ competed in the } n^{\text{th}} \text{ race,} \\ 0 & \text{otherwise} \end{cases}$$

The recency weighted average of the pairwise comparisons between horses $i$ and $j$ is

$$O_{ij} = \frac{\sum_{n=1} \delta_{ijn}(\zeta_j^n - \zeta_i^n)}{\sum_{n=1} \delta_{ijn} w^n}$$

This observed pairwise comparison matrix can be represented as a simplicial complex and analysed via the framework and techniques described in Section 4, with initial reliability weights given by $\sum_{n=1} \delta_{ijn} w^n$. A measure of the underlying competitiveness of each horse can be extracted from the consistent part of this matrix by applying the HodgeRank algorithm to this ranking data, without re-weighting the simplcial complex with regards to inconsistency measures. The real-valued function on the horses given by the solution to (13) can be used as an explanatory variable in the conditional logit model, with horses who have performed better being assigned a higher value.

A second predictive variable can also be derived from this observed pairwise comparison matrix which accounts for information contained in the inconsistent ranking data. This variable is given by applying the HodgeRank algorithm, including local-scale inconsistency weights, to the ranking data and again using the real-valued function given by the solution to (13) as an explanatory variable.

These three predictive variables (the natural logarithm of the probability implied by market odds, the underlying competitiveness derived from consistent data, and the underlying competitiveness accounting for local-scale

inconsistency) can be incorporated in conditional logit models to predict the outcome of future races. These models can be trained on a training set of observed race outcomes and evaluated on unseen races both statistically and economically. The three predictive models considered in this case study are:

(i) $\hat{\theta}_{odds}$: Using the natural logarithm of the probability implied by market odds as the only predictive variable.

(ii) $\hat{\theta}_{Hodge}$: Using both the natural logarithm of the probability implied by market odds and the underlying competitiveness derived from consistent data as predictive variables.

(iii) $\hat{\theta}_{local}$: Using both the natural logarithm of the probability implied by market odds and the underlying competitiveness accounting for local-scale inconsistency.

Two hypothesis tests are conducted to evaluate whether bettors in horse racing parimutuel win markets utilise all of the data in their decision making. In the first test, the ability of bettors to fully utilise the information contained in consistent data is assessed, with the null and alternative hypotheses respectively given as:

$H_0$: The decision making of bettors fully utilises the information available in the consistent past performance data (i.e. $\hat{\theta}_{Hodge}$ does not produce better wagers than $\hat{\theta}_{odds}$)

$H_1$: The decision making of bettors does not fully exploit the information available in the consistent past performance data (i.e. $\hat{\theta}_{Hodge}$ improves the wagers made by $\hat{\theta}_{odds}$)

The second hypothesis test examines how well bettors use information contained in inconsistent data by incorporating this information into the predictive variable formed from the consistent data. The null and alternative hypotheses are:

$H_0$: The decision making of bettors fully accounts for the information contained in both the consistent and locally inconsistent parts of the past performance data (i.e. $\hat{\theta}_{local}$ does not produce better wagers than $\hat{\theta}_{Hodge}$)

$H_1$: The decision making of bettors can be improved by exploiting the information available in both the consistent and locally inconsistent pasts of performance data (i.e. $\hat{\theta}_{local}$ improves the wagers made by $\hat{\theta}_{Hodge}$)

## 5.5 Results

The three conditional logit models utilising combinations of market odds, consistent past performance data and local-scale inconsistency were trained on a training sample of observed race outcomes between 2010 and 2013 consisting of 21915 races and 210328 observations. These predictive models were then statistically and economically evaluated on an unseen sample of observed race outcomes from 2014 consisting of 7395 races and 67567 observations, employing a fractional Kelly wagering strategy with starting capital of £1000.

### 5.5.1 Consistent Ranking Data

The ability of bettors to fully extract and utilise the information contained in the consistent ranking data produced from the past performance data was assessed by comparing models $\hat{\theta}_{odds}$ and $\hat{\theta}_{Hodge}$. This comparison evaluated whether the decisions made by bettors could be statistically and economically improved by incorporating the underlying competitiveness derived from the consistent ranking data formed from pairwise comparisons of historic race results.

| Statistic | $\hat{\theta}_{odds}$ | $\hat{\theta}_{Hodge}$ |
|---|---|---|
| Wald Coefficient p-value | N/A | $1.45847e^{-6}$ |
| LLR p-value | N/A | $8.89971e^{-5}$ |
| $\widetilde{R}^2$ | 0.165476 | 0.165699 |
| AIC | $5.66757e^4$ | $5.66625e^4$ |

Table 2: Statistical evaluation of $\hat{\theta}_{odds}$ and $\hat{\theta}_{Hodge}$

There was evidence at the 1% level that the coefficient of the predictive variable derived from the underlying competitiveness of the horses produced from the consistent ranking data should not be trivial and that the variable improved the predictive power of the conditional logit model. A log likelihood ratio test, with one degree of freedom, confirmed at the 1% significance level that there was information contained within $\hat{\theta}_{Hodge}$ which was not contained within the nested model $\hat{\theta}_{odds}$.

The $\widetilde{R}^2$ value of $\hat{\theta}_{Hodge}$ was slightly higher than $\hat{\theta}_{odds}$, indicating that $\hat{\theta}_{Hodge}$ provided a better fit for the holdout sample of observations. This was further confirmed by the AIC values of the two models, with $AIC(\hat{\theta}_{Hodge}) < AIC(\hat{\theta}_{odds})$. Although the increase in $\widetilde{R}^2$ and decrease in AIC appear small, this was expected since pairwise comparisons formed from past performance data are one of many variables available to bettors for use in their decision making.

In conclusion, there was statistical evidence that the addition of the explanatory variable derived from the consistent ranking data improved the conditional logit model and that the decisions made by bettors could be improved by incorporating this information. An economic evaluation of both models was also conducted with the betting simulation beginning with £1000 of capital and employing a fractional Kelly wagering strategy.

| Statistic | $\hat{\theta}_{odds}$ | $\hat{\theta}_{Hodge}$ |
|---:|:---:|:---:|
| Total Bet (£) | 4205.24 | 7775.66 |
| Win/Bet Ratio (%) | 40.1079 | 39.4631 |
| Rate of Return (%) | -7.80934 | -2.09111 |
| Profit (£) | -328.40 | -162.60 |

Table 3: Economic evaluation of $\hat{\theta}_{odds}$ and $\hat{\theta}_{Hodge}$

The aim of decision makers in speculative markets is to make profitable decisions however these results indicate that bettors overall do not make profitable decisions. The conditional logit model using only the market odds as a predictive variable produced a loss of £328.40 from the initial starting

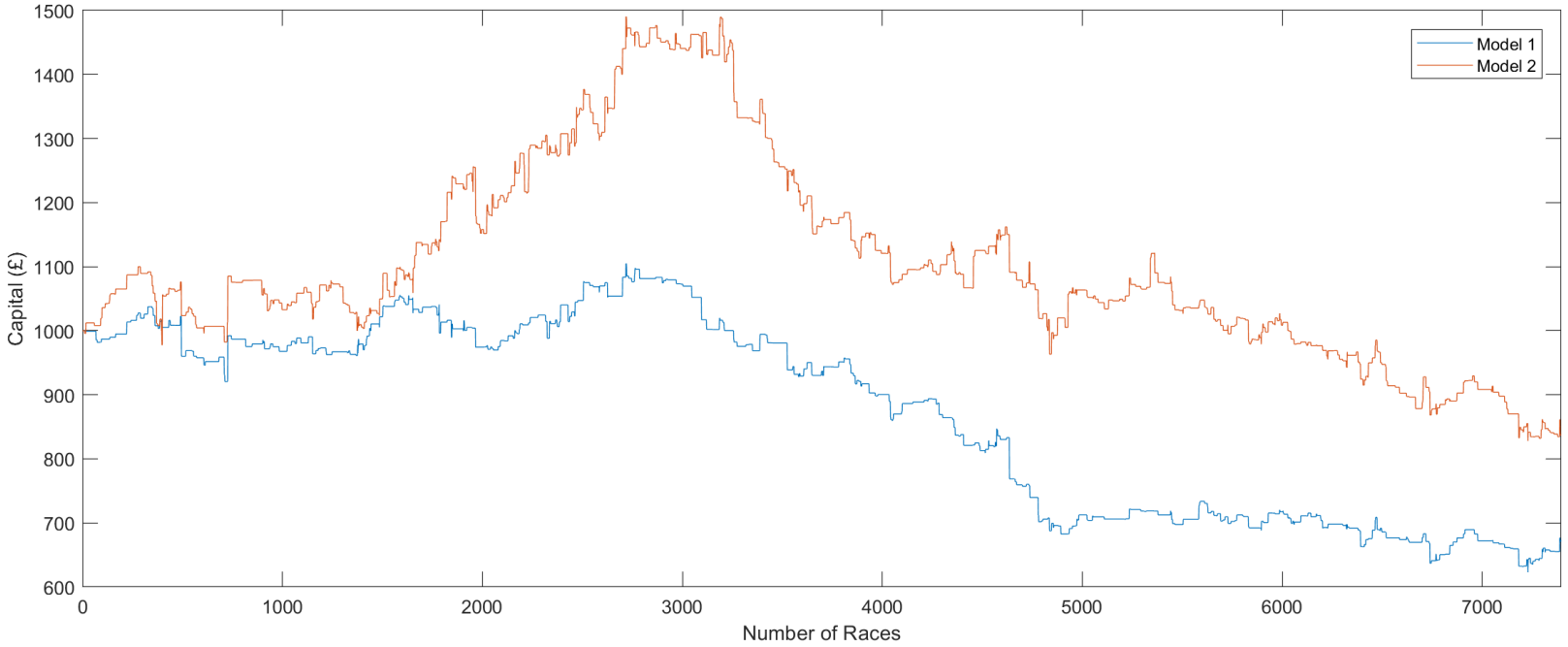


Figure 18: Running Capital of $\hat{\theta}_{odds}$ (Model 1) and $\hat{\theta}_{Hodge}$ (Model 2)

capital of £1000. The addition of the predictive variable derived from the consistent ranking data increased the return from wagers yet, whilst there were periods in the holdout sample where substantial profits were made, these predictions still resulted in a loss of £162.60.

Although there was compelling evidence that more accurate predictions could be made by incorporating information from the consistent ranking data derived from pairwise comparisons of past performances, the decisions made by including this information remained unprofitable. Nonetheless there was an improvement in the economic performance of the model by incorporating this information in the conditional logit model, suggesting the presence of a market inefficiency.

### 5.5.2 Inconsistent Ranking Data

The results of the first hypothesis test indicated that bettors fail to fully exploit the information available to them in the consistent ranking data derived from past performances. The second comparison test between $\hat{\theta}_{odds}$ and $\hat{\theta}_{local}$ evaluated whether better decisions could be made by exploiting information in the local-scale inconsistent ranking data. If $\hat{\theta}_{local}$ were to fit the holdout sample better and lead to better wagering decisions, there would be evidence that bettors fail to account for the information contained in the local-scale inconsistent ranking data. If greater risk-adjusted returns could be made by $\hat{\theta}_{local}$, there would be evidence of a market inefficiency resulting from this failure to account for information contained in inconsistent ranking data.

| Statistic | $\hat{\theta}_{odds}$ | $\hat{\theta}_{local}$ |
|---|---|---|
| Wald Coefficient p-value | N/A | $1.15638e^{-6}$ |
| LLR p-value | N/A | $6.01213e^{-7}$ |
| $\widetilde{R}^2$ | 0.165476 | 0.165848 |
| AIC | $5.66757e^{4}$ | $5.66524e^{4}$ |

Table 4: Statistical evaluation of $\hat{\theta}_{odds}$ and $\hat{\theta}_{local}$

Again there was evidence at the 1% significance level from both the Wald

coefficient and log-likelihood ratio tests that the inclusion of the explanatory variable derived from the consistent and local-scale inconsistent ranking data improved the conditional logit model. Both relevant null hypotheses are therefore rejected in favour of their alternatives that this explanatory variable is correlated with the observed outcomes and that it contains information not captured by the odds.

Both the $\widetilde{R}^2$ and $AIC$ values for $\hat{\theta}_{local}$ were better than those of $\hat{\theta}_{Hodge}$ providing evidence that this model better fit the observed outcomes in the holdout sample. This suggests that there is information contained in the consistent and inconsistent ranking data formed from past performance data which is not accounted for by bettors, and that this information can be used to improve decision making in betting markets.

The economic impact of including information extracted from both consistent and locally inconsistent data was also assessed. A betting simulation was performed starting with £1000 of capital and again employing a fractional Kelly wagering strategy.

| Statistic | $\hat{\theta}_{odds}$ | $\hat{\theta}_{local}$ |
|---|---|---|
| Total Bet (£) | 4205.24 | 9852.15 |
| Win/Bet Ratio (%) | 40.1079 | 41.7824 |
| Rate of Return (%) | -7.80934 | 0.492332 |
| Profit (£) | -328.40 | 48.51 |

Table 5: Economic evaluation of $\hat{\theta}_{odds}$ and $\hat{\theta}_{local}$

The model including both the natural logarithm of the probabilities implied by the market odds and the underlying competitiveness derived from both the consistent and local-scale inconsistent ranking data, produced a profit of £48.51 over the holdout sample and a superior win/bet ratio over $\hat{\theta}_{odds}$. There is therefore evidence of a market inefficiency in parimutuel win markets surrounding horse racing, and that this inefficiency can be exploited by including information from inconsistent ranking data related to the past performance of horses.

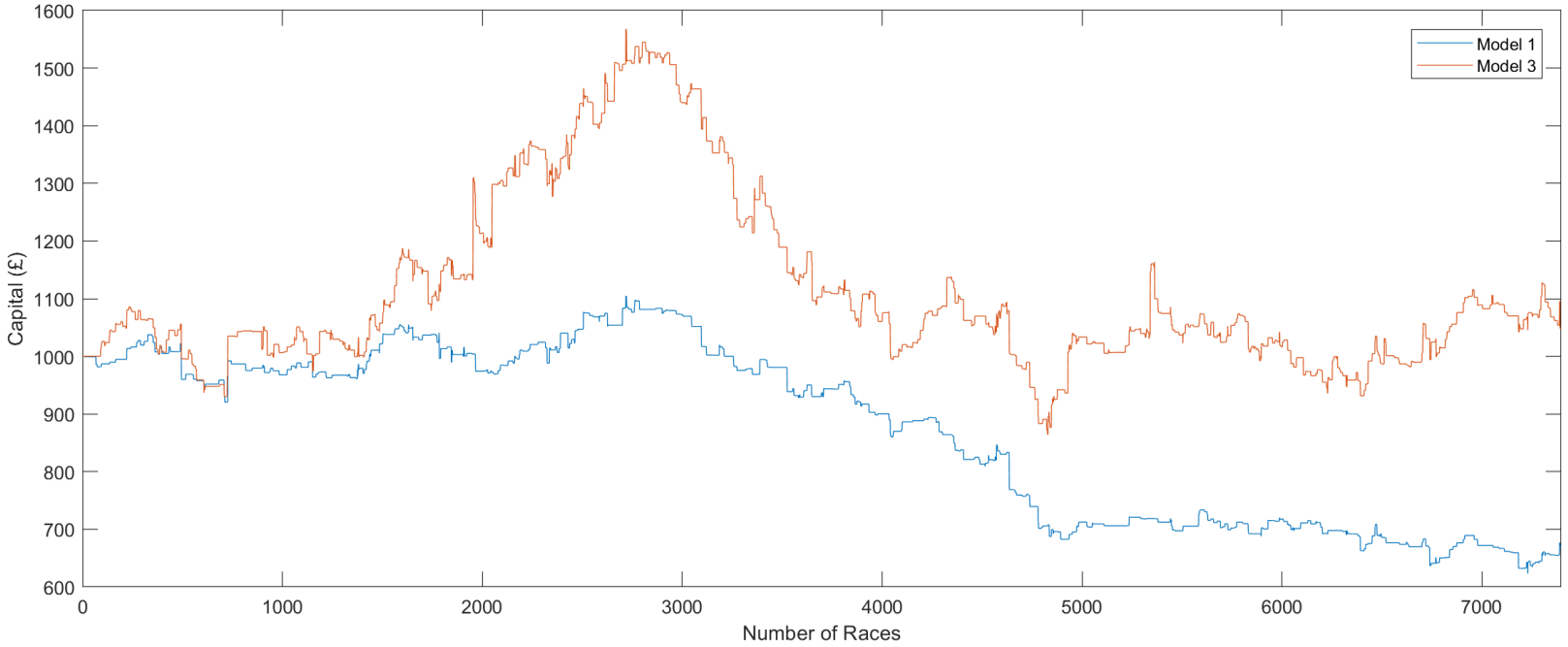


Figure 19: Running Capital of $\hat{\theta}_{odds}$ (Model 1) and $\hat{\theta}_{local}$ (Model 3)

The statistical and economic evaluation of $\hat{\theta}_{Hodge}$ and $\hat{\theta}_{local}$ demonstrates that both these models improve upon $\hat{\theta}_{odds}$. It can therefore be concluded that there is information contained in the pairwise comparisons derived from past performance data which has not been fully discounted in the odds. This suggests that the presence of inconsistencies in the ranking data inhibits the decision making ability of bettors.

In the conclusions reached so far, no direct comparison has been made of $\hat{\theta}_{Hodge}$ and $\hat{\theta}_{local}$. Instead these models have been compared to $\hat{\theta}_{odds}$ which models the decision making of bettors in these betting markets. The McFadden $\widetilde{R}^2$ value across these three models is highest for $\hat{\theta}_{local}$ suggesting that this model is the best fit for the observed outcomes in the holdout sample from amongst the three candidate models. It is expected, and observed, that $\hat{\theta}_{odds}$ has the lowest $\widetilde{R}^2$ if for no other reason than it incorporates fewer explanatory variables. It is however reassuring that this conclusion is also reached by examining the AIC values which penalises models for including additional explanatory variables which do not sufficiently improve the model. It can therefore be concluded that $\hat{\theta}_{local}$ is indeed the best choice of model for predicting winners of horse races.

Although the tests above have shown that $\hat{\theta}_{local}$ is the best choice of the candidate model, they make no judgement about whether it is a good model for the data. The absolute fit of the model is assessed by its binned residual plot (Fig. 20). There are 67572 observations in the holdout sample and therefore the residuals are grouped together into 260 $\sqrt{67572}$ bins.

This plot appears reasonable at first with the majority of residuals falling within the confidence intervals, however it is difficult to firmly conclude this given the density of bins towards the lower end of the expected outcome spectrum. The number of bins which exceed their associated $\pm 2$ standard error confidence intervals is 7, equivalent to 2.7% of the total number of bins. At a 5% significance level, it can be concluded that the model is a good fit for the observed outcomes in the holdout sample.

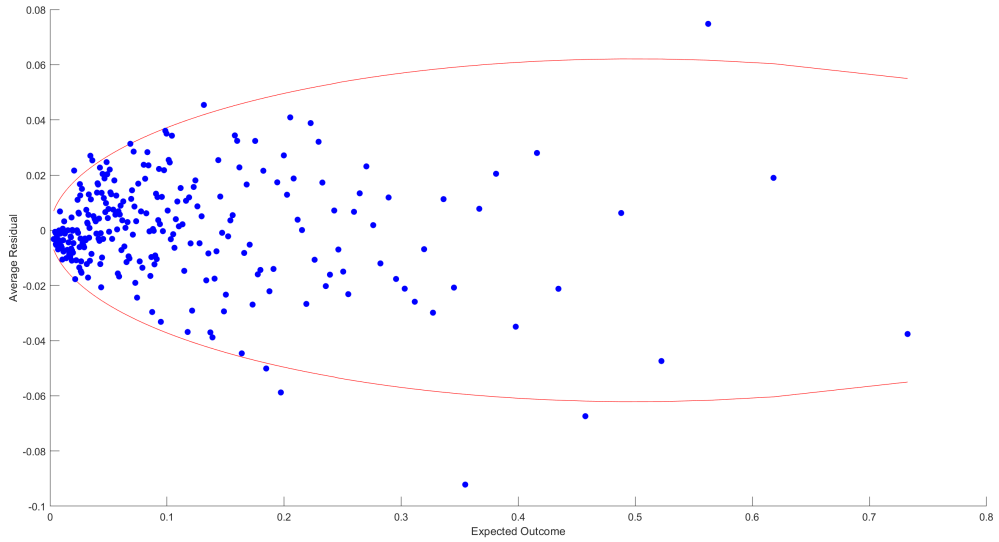


Figure 20: Binned Residual Plot of $\hat{\theta}_{local}$

## 5.6 Conclusion

This case study assessed whether the presence of inconsistencies in ranking data in past performance data inhibited bettors participating in parimutuel horserace wagering markets and, if so, whether market inefficiencies existed as a result. Bettors in these markets seek to place successful wagers and maximise their returns, however the results show that bettors were expected to lose 32.8% of their starting capital over 2014. This loss is unsurprising as these are prediction markets surrounding highly unpredictable events and prices are unlikely to correctly estimate winning probabilities for horses (Lo et al. 1995, Snyder 1978, Lessmann et al. 2010).

To evaluate the impact of inconsistent ranking data on decision making, HodgeRank was employed to separate the pairwise comparisons of past performance data into consistent and inconsistent ranking data, and extract information from the consistent ranking data. Incorporating this information into the conditional logit model, there was a statistically significant improvement in the accuracy of predictions and an improvement in the economic

performance of the fractional Kelly wagering strategy.

HodgeRank was able to better exploit consistent information in the pairwise comparisons of past performance, having first removed inconsistencies from the ranking data, demonstrating that the presence of inconsistent ranking data impacted the decision making of bettors in parimutuel horserace wagering markets. Deviations from rational decision making, fully utilising available information, present an opportunity for achieving excess returns and improved returns were made by including the HodgeRank output from the consistent ranking data.

These results illustrate a deviation from rational decision making and a resulting opportunity for excess returns. Participants in these wagering markets are, however, unlikely to regard this as an issue of much concern. Their aim is to generate economic gain and, whilst less of a financial loss is of benefit, they are unlikely to acknowledge this deficiency in decision making as a true market inefficiency unless there is a demonstrable opportunity for positive.

Although the version of the HodgeRank algorithm employed in $\hat{\theta}_{Hodge}$ resulted in a loss, the more complex version employed in the $\hat{\theta}_{local}$ was able to improve the accuracy of predictions, generate excess returns and achieve a small profit. This version exploited information contained within local-scale inconsistencies, more fully utilising the available data, by incorporating them into the simplicial complex representation of the past performance pairwise comparisons and re-weighting the edges of this complex by their contribution to local-scale inconsistency.

There is a clear impact of inconsistencies within ranking data affecting the ability of individuals to fully exploit available information and creating semi-strong inefficiencies in wagering markets. Accounting for local scale inconsistencies, HodgeRank is able to improve decision making in these areas and take advantage of these inefficiencies for economic gain.

Existing literature in the area shows that wagering markets, particularly those surrounding horse racing, are highly semi-strong efficient, with market prices dominating other variables in forecasting models (Sung & Johnson 2008, Sung et al. 2005, Edelman 2007). Market prices were also

the most significant variable in the conditional logit model with the addition of a HodgeRank variable providing a very small, but statistically significant, improvement to the $\widetilde{R}^2$ of the mode. The study provides further evidence that parimutuel horserace wagering markets are highly semi-strong efficient, despite the presence of market inefficiencies.

Despite the high degree of semi-strong efficiency in horse race wagering markets, inefficiencies have been observed (Sung et al. 2005, Sung et al. 2019, Gabriel & Marsden 1990, Gramm & Ziemba 2008). A series of studies have shown that bettors have difficulties simultaneously accounting for a range of variables derived from available information and their complex interactions (Bolton & Chapman 1986, Benter 1994, White et al. 1992, Sung & Johnson 2008). Techniques which combine and capture these complex, non-linear relationships are capable of outperforming wagering markets (Edelman 2007, Lessmann et al. 2010, Lessmann et al. 2012, Ma et al. 2016, Goddard 2005).

Systematic errors in the behaviour of market participants also contribute to inefficiency in horserace wagering markets (Sung et al. 2019). The most prominent of these errors is the favourite-longshot bias, a phenomenon where market prices underestimate the likelihood of favourites winning and overestimate the winning probabilities of longshots (Ali 1977, Sung et al. 2009, Snyder 1978, Asch et al. 1982). The origin of the favourite-longshot bias is a matter for debate, having been attributed to the behaviour of both bettors (Thaler & Ziemba 1988, Sobel & Raines 2003) and bookmakers reacting to perceived insider trading (Shin 1991, Shin 1991, Sung & Johnson 2010). Nonetheless, the existence of a favourite-longshot bias in wagering markets illustrates the imapact of cognitive errors on market efficiency.

Anchoring effects are another example of inefficiency in wagering markets caused by cognitive errors (Johnson et al. 2009). Individuals display a tendency to anchor their judgements to a previous starting point which may be internally or externally primed (Jacowitz & Kahneman 1995, Chapman & Johnson 1999). This anchoring heuristic can produce cognitive errors when insufficient adjustments are made in forming a judgement from the initial anchor (Tversky & Kahneman 1974, Furnham & Boo 2011), and nuanced anchoring effects have been observed in horserace wagering markets (Johnson

et al. 2009).

Semi-strong inefficiencies have been observed in horserace wagering markets, resulting from the complexity and range of available information, and cognitive errors in bettors themselves. The results of this case study are therefore consistent with existing literature in the field, illustrating how the complexity produced by the presence of conflicting information inhibits decision making in bettors and producing inefficiencies in parimutuel horserace win markets. HodgeRank provides a means to address this inefficiency and, together with the methodology outlined above, exploit it for economic gain.

# 6 Discussion

Financial markets depend upon the rational behaviour of their participants in order to operate efficiently. These participants are expected to immediately and fully make use of information to maximise their subjective expected utility. In doing so, they ensure that market prices are fundamentally correct and agree with intrinsic valuations.

Although rational behaviour is of paramount importance to financial markets, research has shown that rationality constitutes an ideal that is rarely observed in human behaviour(Tversky & Kahneman 1974, Kahneman & Tversky 1972, Frederick 2005). Individuals exhibit tendencies to deviate from normative models of decision making, increasingly so as the complexity of a decision problem taxes their cognitive capabilities.

Inconsistent data poses a challenge for decision makers, providing conflicting evidence for multiple world states. Inconsistencies are often found when information is gathered from multiple sources and these sources are not fully aligned, although they can also occur when criteria for evaluating alternatives shift. Conflicting data requires more effort and attention to effectively utilise, and its presence increases the complexity of decision problems.

Intransitive patterns of preferences are a form of inconsistent ranking data where alternative A is preferred to B, B is preferred to C, and C is preferred to A. These cycles of preferences indicate that an alternative is more (or less) preferable than itself, violating normative models of decision making. Intransitive patterns are a common feature of uncertain events with similar events producing highly different outcomes. For financial markets surrounding uncertain events to be efficient, their participants must effectively handle the intransitive patterns of preferences found in the available rankingdata.

HodgeRank is a topologically-inspired algorithm which models and extracts information from intransitive patterns of preferences, exploiting the natural affinity of topological models for handling cycles in pairwise comparisons. If financial markets are efficient in regards to inconsistent ranking data, market prices should fully reflect the output of HodgeRank and employing the algorithm should offer no competitive advantage over the market.

Our experiment shows, however, that in financial markets surrounding horse racing, historical performance data contains information which is not accounted for in decisions of market participants. Applying HodgeRank to th highly inconsistent ranking data produced from past performance data, the likelihood of outcomes can be more accurately estimated and better decisions made. This represents a failure to fully utilise the available information contained within inconsistent ranking data and a deviation from rational decision making.

The effect of inconsistent ranking data not only inhibits the decision making ability of individuals but also permeates the financial market, affecting the market prices of each outcome. By applying the HodgeRank algorithm to historic performance data and incorporating the output into a conditional logit model, a fractional Kelly wagering strategy is able to determine better prices for the outcomes and generate greater returns than the market. These returns still represent a loss of capital for the second model, which separates out and disregards inconsistent data, however the third model, which reweights the simplicial complex representation with regard to the local scale inconsistency measure, generates a small profit.

Combining an improved version of the HodgeRank algorithm with conditional logit models and Kelly wagering strategies, the inconsistency-based inefficiency in UK parimutuel horserace win markets is exploited to achieve a profit, illustrating that this inefficiency has a real economic cost. This profit is relatively small in comparison with that achieved by exploiting other market inefficiencies (Benter 1993, Sung et al. 2019, Lessmann et al. 2009), however this is to be expected as i) the experiment uses a small portion of the data available to bettors (Bolton & Chapman 1986) (historic performance, weight and distance) and ii) optimising the technique for maximum returns has not been an aim of the project.

It is debatable whether the returns of the second model are sufficiently large to suggest the existence a market inefficiency. It creates an opportunity for traders to reduce their loss rather than satisfy their ambition of achieving gains, and using the returns from this model as evidence of a market inefficiency arguably violates the spirit of the efficient market hypothesis.

Nevertheless, the third model achieves a small profit in the holdout sample, and is clear evidence of an inefficiency in the financial markets surrounding UK horse racing.

The findings of this experiment align with those of the existing literature in the field showing that wagering markets are highly semi-strong efficient (Sung & Johnson 2008, Edelman 2007). Market prices were the dominant variable in the conditional logit model and despite the sophistication of the HodgeRank algorithm, the information extracted by it was only able to provide a very small increase to the $\widetilde{R}^2$ of the mode. This should come as little surprise, however, as bettors consider a plethora of data beyond the historic performance, weight and distance data used in the experiment and, whilst past performances are a significant factor in forecasting winners of future races, they are not the sole determinant of market prices (Bolton & Chapman 1986, Lessmann et al. 2009).

Inefficiencies in wagering markets arise from difficulties capturing complex non-linear relationships between variables (Ma et al. 2016, Lessmann et al. 2009, Lessmann et al. 2010, Lessmann et al. 2012, Edelman 2007, Goddard 2005) and from systematic behavioural errors in market participants (Ali 1977, Sung et al. 2009, Snyder 1978, Asch et al. 1982). Security prices in parimutuel horserace win markets are affected by the presence of inconsistencies in ranking data produced from historical performance data, failing to fully account for information contained in both the consistent and inconsistent parts of the dataset. This provides evidence for a new, inconsistency-based, market inefficiency relating to the difficult of understanding complex non-linear relationships.

## 6.1 Origins of inconsistency-based inefficiencies

One area which this project has not addressed is the precise mechanism by which inconsistent ranking data causes systematic errors in decision making. The presence of inconsistent ranking data increases the complexity of decision problems, placing additional burden on the cognitive resources of the decision maker. If the decision problem persists in straining these re-

sources, focusing on the task depletes a limited pool of attention, and the individual is more likely to defer or give up on the problem (Tversky & Kahneman 1992, Baumeister et al. 1998, Sweller 1988, Frederick 2005).

Heuristics are often employed to simplify decision problems and decrease the feeling of cognitive strain to acceptable levels (Tversky & Kahneman 1974, Slovic & Lichtenstein 1971, Cosmides & Tooby 1994). These heuristics are simplifications, approximating portions of the decision making process with simpler tasks. Applying heuristics risks the integrity of the decision making process as these approximations may be inadequate and produce errors in the decisions made.

Decision makers may exhibit a degree of confirmation bias, rejecting information which disagrees with their already held internalised mental models (Nickerson 1998, Jonas et al. 2001). Where conflicting information exists, decision makers acknowledge data which agrees with the world state they perceive, bolstering their evidence in support of it, and disregard other data.

In contrast to confirmation bias which relates to internally held beliefs, Luce (1998) argued that when faced with conflicts, decision makers often opt to maintain the status quo. One explanation for this status quo bias is that individuals attempt to conserve cognitive resources and avoid making decisions (Ritov & Baron 1992). Alternatively this bias may be an extension of anchoring effects where the status quo forms an 'anchor' for judgements and decision makers require a significant incentive to move away from this initial judgement due to their risk averse tendencies (Kahneman & Tversky 1979).

An explanation for the existence of inconsistency-based inefficiencies is that decision makers display tendencies to adhere to previously held beliefs, whether externally or internally held, and the confusing nature of inconsistent data does little to alter these beliefs. "People are awfully good at fooling themselves. They're so sure they know the answer that they don't want to confuse people with ugly-looking data" (Broad 1999). It should be noted that this explanation for inconsistency-based inefficiencies is supported by observations of anchoring effects in wagering markets (Johnson et al. 2009).

## 6.2 Methodology

HodgeRank, a topologically-inspired ranking algorithm, is applied, in combination with statistical forecasting methods and Kelly wagering strategies, to seek returns from historical performance data. The findings of the experiment demonstrate the value of the HodgeRank framework, and by extension network and topological approaches, in modelling and understanding inconsistent ranking data, extracting information from the publicly available information which is not captured by the market.

It has been necessary to increase the informational efficiency of the HodgeRank algorithm itself in order to achieve abnormal returns. This has been realised by further developing the algorithm in three ways:

(i) **Measuring underlying performance**: HodgeRank finds a ranking solution by measuring underlying preference for alternatives, minimising the difference between observed and consistent pairwise comparisons. This ranking lacks a nuanced understanding of how close or far apart consecutively ranked items are.

Experiments demonstrate that the information lost by converting a measurement of underlying preference into a ranking is sufficient that the resulting output is entirely captured by market prices in wagering markets. This is not, however, true for the measurement of of underlying preference where there is information of economic importance contained within.

(ii) **Edge weights**: Information sources are not all equal in terms of their utility and reliability in a decision problem. The project has extended HodgeRank to incorporate weights on each edge in the simplicial complex representation, estimating how important the information contained in the edge should be. Again, experiments show the value of weighting edges in the complex, improving the informational efficiency of the algorithm.

(iii) **Exploiting inconsistency**: Unlike other techniques which analyse inconsistent ranking data, HodgeRank includes a framework for un-

derstanding inconsistencies. The algorithm itself, however, does not exploit this understanding, separating out consistent and inconsistent information and deriving ranking solutions from the consistent data alone.

The version of HodgeRank presented in this project, and employed in experiments, analyses the structure of these inconsistencies and identifies pairs of alternatives which contribute greatly to local-scale inconsistency in the complex. The information about these pairs conflicts with the remaining information in the dataset to a large extent. By regarding this data with scepticism, and re-weighting the complex appropriately, economically valuable is extracted by the algorithm which is not captured by the market.

There is no claim that HodgeRank has been developed to its fullest, nor that it extracts all the information available in pairwise comparisons of historical performance data of UK horseraces. It may be the case that other techniques, or a further developed HodgeRank algorithm, are more effective in finding more valuable information from the available data. Nevertheless, the development of HodgeRank outlined in this project is sufficient to answer the research questions.

## 6.3 Further work

Inconsistency-based inefficiencies are not necessarily limited to financial markets surrounding uncertain events. Datasets gathered from multiple sources are highly likely to contain conflicting information with some level of disagreement between the various sources. Examples of settings likely to produce conflicting information include uncertain events, voter aggregation and consumer choice behaviour. Evaluating the potential for inconsistency-based inefficiencies in other financial markets will provide a better picture of the impact of inconsistent data on decision making and market efficiency. In addition, a broader study of inconsistency-based inefficiencies data may shed light on their origin.

There are areas where changes to the application of the HodgeRank algorithm may result in improved performance and greater economic gains. It is important, however, that this be balanced with the substantial computational cost of running the algorithm and other approaches may ultimately be deemed more economically viable.

Some work has been conducted to optimise the parameters of HodgeRank in this environment however, due to its theoretical complexities, in particular its use of pseudo-inverses, it is exceedingly difficult to determine how perturbations in these parameters affect the output of the algorithm. Thus it is impractical to analytically optimise the parameters of the input variables, the pairwise comparisons and weights, to maximise the value of the output. Instead a grid search approach will have to be adopted at a significant cost in both computational resources and time.

Re-weighting pairwise comparisons by their contribution to local-scale inconsistency in the network improves the information captured from the data. It is intuitive to consider whether re-weighting comparisons by their contribution to global scale inconsistency in the network, cycles of any length, will further increase the effectiveness of the HodgeRank algorithm. Whilst this measure has been theoretically developed, it requires unfeasible computational resources and is unsuitable for practical applications. Approximations or significant computational resources (most likely both) will be required to assess whether this more general measure of inconsistency adds value to the HodgeRank algorithm.

HodgeRank may also be improved by incorporating more theory from topology and network studies, although many of these avenues have been exhausted. An unfortunate consequence of Lemma 1 is that there is no clear approach to extending the analysis of the 1-cochain space to dimensions higher than 2. All $k$-cochains, with $k > 2$, derived from a pairwise comparison matrix are 0. Nonetheless, there is a large body of existing literature about networks and a growing body concerning topological data analysis, and so the potential for substantive theoretical leaps remains.

## 6.4 Concluding Remarks

Inconsistent ranking data will continue to impact decision making and the efficiency of financial markets until individuals improve their handling and processing of such data. Evolutionary psychologists would argue that fully utilising inconsistent ranking data falls beyond the purview of human behaviour, exceeding reasonable computational and time restraints (a point we cannot deny given the computational requirements of the HodgeRank algorithm). We believe, however, there is scope to improve decision making and a necessary first step in this process is to understand the mechanisms which are involved in making decisions from inconsistent ranking data, and how these mechanisms fail to adhere to the standards of rationality.

# References

*A Guide to Handicapping* (2014), Technical report, British Horseracing Authority.
URL: `http://www.britishhorseracing.com/wp-content/uploads/2014/03/Guide-to-Handicapping.pdf`

Ali, M. (1977), 'Probability And Utility Estimates For Racetrack Bettors', *The Journal of Political Economy* **85**(4), 803–815.

Ali, M. M. (1998), 'Probability models on horse-race outcomes', *Journal of Applied Statistics* **25**(2), 221–229.

Allais, M. (1953), 'Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine', *Econometrica* **21**(4), 503–546.

Allison, P. (2001), *Missing Data*, Vol. 136 of *Quantitative Applications in the Social Sciences*, Sage, Thousand Oaks, CA.

Andrikogiannopoulou, A. & Papakonstantinou, F. (2018), 'Individual Reaction To Past Performance Sequences: Evidence From A Real Marketplace', *Management Science* **64**(4), 1957–1973.

Arthur, B. (1994), 'Inductive Reasoning And Bounded Rationality', *The American Economic Review* **84**(2), 406–411.

Asch, P., Malkiel, B. & Quandt, R. (1982), 'Racetrack Betting And Informed Behavior', *Journal of Financial Economics* **10**(2), 187–194.

Atanasov, P., Rescober, P., Stone, E., Swift, S., Servan-Schreiber, E., Tetlock, P., Ungar, L. & Mellers, B. (2017), 'Distilling The Wisdom Of Crowds: Prediction Markets vs. Prediction Polls', *Management Science* **63**(3), 691–706.

Ball, R. (1978), 'Anomalies In Relationships Between Securities' Yields And Yield-Surrogates', *Journal of Financial Economics* **6**(2-3), 103–126.

Ballou, D. & Pazer, H. (1985), 'Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems', *Management Science* **31**(2), 150–162.

Barberis, N. & Thaler, R. (2003), A Survey of Behavioral Finance, *in* G. Constantinides, M. Harris & R. Stulz, eds, 'Handbook of the Economics of Finance', Elsevier B.V., Amsterdam, pp. 1052–1121.

Barberis, N. & Xiong, W. (2009), 'What Drives The Disposition Effect? An Analysis Of A Long-Standing Preference-Based Explanation', *The Journal of Finance* **64**(2), 751–784.

Basu, S. (1977), 'Investment Performance Of Common Stocks In Relation To Their Price-Earning Ratios: A Test Of The Efficient Market Hypothesis', *The Journal of Finance* **32**(3), 663–682.

Baumeister, R., Bratslavsky, E., Muraven, M. & Tice, D. (1998), 'Ego Depletion: Is The Active Self A Limited Resource?', *Journal of Personality and Social Psychology* **74**(5), 1252–1265.

Ben-Akiva, M. & Lerman, S. (1985), *Discrete Choice Analysis: Theory And Application To Travel Demand*, The MIT Press, Cambridge, MA.

Benter, B. (1993), Computer based horse race handicapping and wagering systems: A report, *in* 'Operations Research Society of America Conference', Phoenix, US.

Benter, W. (1994), Computer Based Horse Race Handicapping and Wagering Systems, *in* D. Hausch, V. Lo & W. Ziemba, eds, 'Efficiency of Race Track Betting Markets', Acadmic Press, London, pp. 183–198.

Berg, J., Forsythe, R., Nelson, F. & Rietz, T. (2008), 'Results From A Dozen Years Of Election Future Markets Research', *Handbook of Experimental Economics Results* **1**, 742–751.

Berg, J., Neumann, G. & Rietz, T. (2009), 'Searching For Google's Value: Using Prediction Markets To Forecast Market Capitalization Prior To An Initial Public Offering', *Management Science* **55**(3), 348–361.

Berg, J. & Rietz, T. (2003), 'Prediction Markets As Decision Support Systems', *Information Systems Frontiers* **5**(1), 79–93.

Berry, S. (1994), 'Estimating Discrete-Choice Models Of Product Differentiation', *The RAND Journal of Economics* **25**(2), 242–262.

Berry, S., Levinsohn, J. & Pakes, A. (1995), 'Automobile Prices In Market Equilibrium', *Econometrica* **63**(4), 841–890.

Biggs, S., Bedard, J., Gaber, B. & Linsmeier, T. (1985), 'The Effects Of Task Size And Similarity On The Decision Behavior Of Bank Loan Officers', *Management Science* **31**(8), 970–987.

Bockenholt, U., Albert, D., Aschenbrenner, M. & Schmalhofer, F. (1991), 'The Effects Of Attractiveness, Dominance, And Attribute Differences On Information Acquisition In Multiattribute Binary Choice', *Organizational Behavior and Human Decision Processes* **49**(2), 258–281.

Bolton, R. & Chapman, R. (1986), 'Searching for positive returns at the track: A multinomial logit model for handicapping horse races', *Management Science* **32**(8), 1040–1060.

Boskin, M. (1974), 'A Conditional Logit Model Of Occupational Choice', *Journal of Political Economy* **82**(2), 389–398.

Breiman, L. (1961), Optimal Gambling Systems For Favorable Games, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, CA, pp. 65–78.

Broad, W. J. (1999), 'Data Tying Cancer To Electric Power Found To Be False', *New York Times* .
URL: `https://www.nytimes.com/1999/07/24/us/data-tying-cancer-to-electric-power-found-to-be-false.html`

Brunnermeier, M. K. & Nagel, S. (2004), 'Hedge Funds And The Technology Bubble', *The Journal of Finance* **59**(5), 2013–2040.

Canfield, B. R., Fauman, B. C. & Ziemba, W. T. (1987), 'Efficient Market Adjustment Of Odds Prices To Reflect Track Biases', *Management Science* **33**(11), 1428–1439.

Chapman, G. B. & Johnson, E. J. (1999), 'Anchoring, Activation, And The Construction Of Values', *Organizational Behavior and Human Decision Processes* **79**(2), 115–153.

Chen, K.-Y. & Plott, C. R. (2002), Information Aggregation Mechanisms: Concept, Design And Implementation For A Sales Forecasting Problem, Social Science Working Paper 1131, California Institute of Technology, Pasadena, CA.

Chen, Y. & Feng, J. (2014), 'Efficient Method For Moore-Penrose Inverse Problems Involving Symmetric Structures Based On Group Theory', *Journal of Computing in Civil Engineering* **28**(2), 182–190.

Chung, F. (1997), *Spectral Graph Theory*, number 92 *in* 'Regional Conference Series in Mathematics', American Mathematical Society, Providence, RI.

Cootner, P. H., ed. (1964), *The Random Character Of Stock Market Prices*, MIT Press, Cambridge, MA.

Cosmides, L. & Tooby, J. (1994), 'Better Than Rational: Evolutionary Psychology And The Invisible Hand', *The American Economic Review* **84**(2), 327–332.

Cowgill, B., Wolfers, J. & Zitzewitz, E. (2009), Using Prediction Markets To Track Information Flows: Evidence From Google, *in* S. Das, M. Ostrovsky, D. Pennock & B. Szymanksi, eds, 'Auctions, Market Mechanisms and Their Applications', Vol. 14, Springer, Berlin, p. 3.

Davidson, R. & MacKinnon, J. G. (2004), *Econometric Theory And Methods*, Oxford University Press, Oxford.

Davies, P., Greenwood, M. & Li, H. (2001), 'A Conditional Logit Approach To U.S. State-To-State Migration', *Journal of Regional Science* **41**(2), 337–360.

Dixon, J. D. & Mortimer, B. (1996), *Permutation Groups*, Graduate Texts in Mathematics, Springer-Verlag, New York.

Doyle, J. (1999*a*), Bounded Rationality, *in* R. Wilson & F. Keil, eds, 'The MIT Encyclopedia of the Cognitive Sciences', The MIT Press, Cambridge, MA, pp. 92–94.

Doyle, J. (1999*b*), Rational Decision Making, *in* R. Wilson & F. Keil, eds, 'The MIT Encyclopedia of the Cognitive Sciences', The MIT Press, Cambridge, MA, pp. 701–703.

Edelman, D. (2007), 'Adapting Support Vector Machine Methods For Horserace Odds Prediction', *Annals of Operations Research* **151**(1), 325–336.

Edelsbrunner, H. & Harer, J. (2010), *Computational Topology: An Introduction*, American Mathematical Society.

English, L. (1999), *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, John Wiley & Sons, New York, NY.

Estrada, E. & Rodríguez-Velázquez, J. (2005), 'Spectral Measures of Bipartivity in Complex Networks', *Physical Review E* **72**(4).

Fama, E. (1969), 'Efficient Capital Markets: A Review Of Theory And Empirical Work', *The Journal of Finance* **25**(2), 28–30.

Fama, E. (1976), 'Efficient Capital Markets: Reply', *The Journal of Finance* **31**(1), 143–145.

Figlewski, S. (1979), 'Subjective information and market efficiency in a betting market', *Journal of Political Economy* **87**(1), 75–88.

Finnerty, J. E. (1976), 'Insiders And Market Efficiency', *The Journal of Finance* **31**(4), 1141–1148.

Fischer, G. W., Luce, M. F. & Jia, J. (2000), 'Attribute Conflict And Preference Uncertainty: Effects On Judgment Time And Error', *Management Science* **46**(1), 88–103.

Forsythe, R., Nelson, F., Neumann, G. R. & Wright, J. (1992), 'Anatomy Of An Experimental Political Stock Market', *The American Economic Review* **82**(5), 1142–1161.

Frederick, S. (2005), 'Cognitive Reflection And Decision Making', *Journal of Economic Perspectives* **19**(4), 25–42.

Furnham, A. & Boo, H. C. (2011), 'A Literature Review Of The Anchoring Effect', *The Journal of Socio-Economics* **40**(1), 35–42.

Gabriel, P. E. & Marsden, J. R. (1990), 'An Examination Of Market Efficiency In British Racetrack Betting', *Journal of Political Economy* **98**(4), 874–885.

Gelman, A. & Hill, J. (2007), *Data Analysis Using Regression And Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press, New York.

Gigerenzer, G. & Gaissmaier, W. (2011), 'Heuristic Decision Making', *Annual Review of Psychology* **62**, 451–482.

Gill, J. & King, G. (2004), 'What To Do When Your Hessian Is Not Invertible: Alternatives To Model Respecification In Nonlinear Estimation', *Sociological Methods & Research* **33**(1), 54–87.

Goddard, J. (2005), 'Regression Models For Forecasting Goals And Match Results In Association Football', *International Journal of Forecasting* **21**(2), 331–340.

Godfrey, M. D., Granger, C. W. J. & Morgenstern, O. (1964), 'The Random-Walk Hypothesis Of Stock Market Behavior', *Kyklos* **17**(1), 1–30.

Gramm, M. & Ziemba, W. T. (2008), The Dosage Breeding Theory For Horse Racing Predictions, *in* D. B. Hausch & W. T. Ziemba, eds, 'Handbook of Sports and Lottery Markets', Handbooks in Finance, Elsevier, pp. 307–340.

Granger, C. W. & Morgenstern, O. (1963), 'Spectral Analysis Of New York Stock Market Prices', *Kyklos* **16**, 1–27.

Haselton, M., Nettle, D. & Andrews, P. (2005), The Evolution Of Cognitive Bias, *in* D. Buss, ed., 'The Handbook of Evolutionary Psychology', John Wiley & Sons.

Hatcher, A. (2001), *Algebraic Topology*, Cambridge University Press, New York.

Hausch, D. & Ziemba, W. (1990), 'Arbitrage Strategies for Cross-Track Betting on Major Horse Races', *The Journal of Business* **63**(1), 61–78.

Hausman, J. & McFadden, D. (1984), 'Specification Tests For The Multinomial Logit Model', *Econometrica* **52**(5), 1219–1240.

Healy, P., Linardi, S., Lowery, R. & Ledyard, J. (2010), 'Prediction Markets: Alternative Mechanisms For Complex Environments With Few Traders', *Management Science* **56**(11), 1977–1996.

Henery, R. (1983), 'Permutation probabilities for gamma random variables', *Journal of Applied Probability* **20**(4), 822–834.

Hoffman, S. & Duncan, G. (1988), 'Multinomial And Conditional Logit Discrete-Choice Models In Demography', *Demography* **25**(3), 415–427.

Hopman, J. W. (2007), 'Using Forecasting Markets To Manage Demand Risk', *Intel Technology Journal* **11**(2), 127–135.

Hosmer Jr., D., Lemeshow, S. & Sturdivant, R. (2013), *Applied Logistic Regression*, Vol. 398 of *Wiley Series in Probability and Statistics*, third edition edn, John Wiley & Sons, Hoboken, NJ.

Imas, A. (2016), 'The Realization Effect: Risk-Taking After Realized Versus Paper Losses', *American Economic Review* **106**(8), 2086–2109.

Jacowitz, K. E. & Kahneman, D. (1995), 'Measures Of Anchoring In Estimation Tasks', *Personality and Social Psychology Bulletin* **21**(11), 1161–1166.

Jensen, M. (1978), 'Some Anomalous Evidence Regarding Market Efficiency', *Journal of Financial Economics* **6**(2-3), 95–101.

Jensen, M. & Ruback, R. (1983), 'The Market For Corporate Control: The Scientific Evidence', *Journal of Financial Economics* **11**(1-4), 5–50.

Jiang, X., Lim, L.-H., Yao, Y. & Ye, Y. (2011), 'Statistical ranking and combinatorial hodge theory', *Mathematical Programming* **127**(1), 203–244.

Johnson, J. E. V., Liu, S. & Schnytzer, A. (2009), 'To What Extent Do Investors In A Financial Market Anchor Their Judgments? Evidence From The Hong Kong Horserace Betting Market', *Journal of Behavioral Decision Making* **22**(4), 410–434.

Johnson, J., Jones, O. & Tang, L. (2006), 'Exploring Decision Makers' Use of Price Information in a Speculative Market', *Management Science* **52**(6), 897–908.

Jonas, E., Schulz-Hardt, S., Frey, D. & Thelen, N. (2001), 'Confirmation Bias In Sequential Information Search After Preliminary Decisions: An Expansion Of Dissonance Theoretical Research On Selective Exposure To Information', *Journal of Personality and Social Psychology* **80**(4), 557–571.

Kahneman, D. & Tversky, A. (1972), 'Subjective Probability: A Judgment Of Representativeness', *Cognitive Psychology* **3**(3), 430–454.

Kahneman, D. & Tversky, A. (1973), 'On The Psychology Of Prediction', *Psychological Review* **80**(4), 237–251.

Kahneman, D. & Tversky, A. (1979), 'Prospect Theory: An Analysis Of Decision Under Risk', *Econometrica* **47**(2), 263–292.

Kahneman, D. & Tversky, A. (1984), 'Choice, Values, and Frames', *American Psychologist* **39**(4), 341–350.

Kelly, J. L. (1956), 'A New Interpretation Of Information Rate', *Bell System Technical Journal* **35**, 917–926.

Kendall, M. G. & Hill, A. B. (1953), 'The Analysis Of Economic Time-Series-Part-I: Prices', *Journal of the Royal Statistical Society* **116**(1), 11–34.

Klein, N. & Yadav, M. (1989), 'Context Effects On Effort And Accuracy In Choice: An Enquiry Into Adaptive Decision Making', *Journal of Consumer Research* **15**(4), 411–421.

Lessmann, S., Sung, M.-C. & Johnson, J. (2009), 'Identifying Winners of Competitive Events: An SVM-Based Classification Model for Horserace Prediction', *European Journal of Operational Research* **196**(2), 569–577.

Lessmann, S., Sung, M.-C. & Johnson, J. (2010), 'Alternative Methods of Predicting Competitive Events: An Application in Horserace Betting Markets', *International Journal of Forecasting* **26**(3), 518–536.

Lessmann, S., Sung, M.-C., Johnson, J. & Ma, T. (2012), 'A New Methodology For Generating And Combining Statistical Forecasting Models To Enhance Competitive Event Prediction', *European Journal of Operational Research* **218**(1), 163–174.

Levin, I., Schneider, S. & Gaeth, G. (1998), 'All Frames Are Not Created Equal: A Typology And Critical Analysis Of Framing Effects', *Organizational Behavior and Human Decision Processes* **76**(2), 149–188.

Lo, V. S. Y., Bacon-Shone, J. & Busche, K. (1995), 'The Application Of Ranking Probability Models To Racetrack Betting', *Management Science* **41**(6), 1048–1059.

Luce, M. (1998), 'Choosing To Avoid: Coping With Negatively Emotion-Laden Consumer Decisions', *Journal of Consumer Research* **24**(4), 409–433.

Luce, R. D. & Raiffa, H. (1957), *Games And Decisions: Introduction and Critical Survey*, Dover Publications, Inc., New York.

Ma, T., Tang, L., McGroarty, F., Sung, M.-C. & Johnson, J. E. V. (2016), 'Time Is Money: Costing The Impact Of Duration Misperception In Market Prices', *European Journal of Operational Research* **255**(2), 397–410.

MacLean, L. C., Thorp, E. O., Zhao, Y. & Ziemba, W. T. (2011), Medium Term Simulations Of The Full Kelly And Fractional Kelly Investment Strategies, *in* L. C. MacLean, E. O. Thorp & W. T. Ziemba, eds, 'The Kelly Capital Growth Investment Criterion: Theory and Practice', World Scientific, Singapore, pp. 543–561.

Maclean, L., Thorp, E. O. & Ziemba, W. T. (2010), 'Long-Term Capital Growth: The Good And Bad Properties Of The Kelly And Fractional Kelly Capital Growth Criteria', *Quantitative Finance* **10**(7), 681–687.

Malkiel, B. (2003), 'The Efficient Market Hypothesis And Its Critics', *Journal of Economic Perspectives* **17**(1), 59–82.

Marschak, J. (1960), 'Binary Choice Constraints on Random Utility Indicators', *Proceedings Of The First Stanford Symposium On Mathematical Methods In The Social Sciences* pp. 312–329.

McFadden, D. (1973), Conditional Logit Analysis of Qualitative Choice Behavior, *in* P. Zarembka, ed., 'Frontiers in Econometrics', Academic Press, New York, pp. 105–142.

McFadden, D. (1974), 'The Measurement Of Urban Travel Demand', *Journal Of Public Economics* **3**(4), 303–328.

Moore, A. B. (1962), A Statistical Analysis Of Common Stock Prices, Ph.D. thesis, Graduate School of Business, University of Chicago.

Myung, I. J. (2003), 'Tutorial On Maximum Likelihood Estimation', *Journal of Mathematical Psychology* **47**(1), 90–100.

Nickerson, R. S. (1998), 'Confirmation Bias: A Ubiquitous Phenomenon In Many Guises', *Review of General Psychology* **2**(2), 175–220.

Nisbett, R. & Ross, L. (1980), *Human Inference: Strategies And Shortcoming Of Social Judgment*, Century Psychology, Prentice-Hall, Englewood Cliffs, NJ.

Payne, J., Bettman, J. & Johnson, E. (1993), *The Adaptive Decision Maker*, Cambridge University Press.

Penrose, R. (1955), 'A Generalized Inverse For Matrices', *Mathematical Proceedings of the Cambridge Philosophical Society* **51**(3), 406–413.

Prelec, D. & Loewenstein, G. (1991), 'Decision Making Over Time and Under Uncertainty: A Common Approach', *Management Science* **37**(7), 770–786.

Radner, R. & Miller, L. (1970), 'Demand And Supply In U.S. Higher Education: A Progress Report', *The American Economic Review* **60**(2), 326–334.

Redman, T. (1996), *Data Quality for the Information Age*, Artech House, Norwood, MA.

Ritov, I. & Baron, J. (1992), 'Status-Quo And Omission Biases', *Journal of Risk and Uncertainty* **5**(1), 49–61.

Rubin, D. (2004), *Multiple Imputation For Nonresponse In Surveys*, John Wiley & Sons, Hoboken, NJ.

Saaty, T. (1990), 'How To Make A Decision: The Analytic Hierarchy Process', *European Journal of Operational Research* **48**, 9–26.

Sauer, R. D. (1998), 'The Economics Of Wagering Markets', *Journal of Economic Literature* **36**(4), 2021–2064.

Savage, L. (1972), *The Foundations of Statistics*, Dover Publications, Inc., New York.

Schafer, J. (1999), 'Multiple Imputation: A Primer', *Statistical Methods in Medical Research* **8**(1), 3–15.

Schumpeter, J. (1976), *Capitalism, Socialism and Democracy*, Routledge.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H. & Simons, A. (1991), 'Ease Of Retrieval As Information: Another Look At The Availability Heuristic', *Journal of Personality and Social Psychology* **61**(2), 195–202.

Schwert, G. W. (2003), Anomalies And Market Efficiency, *in* G. Constantinides, M. Harrison & R. Stulz, eds, 'Handbook Of The Economics Of Finance', Vol. 1, Elsevier, Amsterdam, pp. 939–974.

Shin, H. S. (1991), 'Optimal Betting Odds Against Insider Traders', *The Economic Journal* **101**(408), 1179–1185.

Simon, H. (1979), 'Rational Decision Making In Business Organizations', *The American Economic Review* **69**(4), 493–513.

Slovic, P. & Lichtenstein, S. (1971), 'Comparison Of Bayesian And Regression Approaches To Study Of Information Processing In Judgment', *Organizational Behavior and Human Performance* **6**(6), 649–744.

Smith, C. (1986), 'Investment Banking And The Capital Acquisition Process', *Journal of Financial Economics* **15**(1-2), 3–29.

Smith, J. & von Winterfeldt, D. (2004), 'Decision Analysis in "Management Science"', *Management Science* **50**(5), 561–574.

Snyder, W. (1978), 'Horse Racing: Testing the Efficient Markets Model', *The Journal of Finance* **33**(4), 1109–1118.

Sobel, R. S. & Raines, S. T. (2003), 'An Examination Of The Empirical Derivative Of The Favourite-Longshot Bias In Racetrack Betting', *Applied Economics* **35**(4), 371–385.

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A. & Carpenter, J. (2009), 'Multiple Imputation For Missing Data In Epidemiological And Clinical Research: Potential And Pitfalls', *BMJ* **338**, b2393.

Suhonen, N. & Saastamoinen, J. (2018), 'How Do Prior Gains And Losses Affect Subsequent Risk Taking? New Evidence From Individual-Level Horse Racing Bets', *Management Science* **64**(6), 2797–2808.

Sung, M.-C. & Johnson, J. (2007), 'The Influence Of Market Ecology On Market Efficiency: Evidence From A Speculative Financial Market', *Journal of Gambling Business and Economics* **1**(3), 185–198.

Sung, M.-C. & Johnson, J. (2010), 'Revealing Weak-Form Inefficiency in a Market for State Contingent Claims: The Importance of Market Ecology, Modelling Procedures and Investment Strategies', *Economica* **77**(305), 128–147.

Sung, M.-C., Johnson, J. & Bruce, A. (2005), Searching for Semi-Strong Form Inefficiency in the UK Racetrack Betting Market, *in* 'Information Efficiency in Financial and Betting Markets', Cambridge University Press, Cambridge, pp. 179–192.

Sung, M.-C., Johnson, J. & Dror, I. (2009), 'Complexity As A Guide To Understanding Decision Bias: A Contribution To The Favorite-Longshot Bias Debate', *Journal of Behavioral Decision Making* **22**(3), 318–337.

Sung, M.-C., McDonald, D. C. J., Johnson, J. E. V., Tai, C.-C. & Cheah, E.-T. (2019), 'Improving Prediction Market Forecasts By Detecting And Correcting Possible Over-Reaction To Price Movements', *European Journal of Operational Research* **272**(1), 389–405.

Sung, M.-C., McDonald, D. & Johnson, J. (2016), 'Probabilistic Forecasting With Discrete Choice Models: Evaluating Predictions With Pseudo-Coefficients Of Determination', *European Journal of Operational Research* **248**(3), 1021–1030.

Sung, M. & Johnson, J. E. V. (2008), Semi-Strong Form Information Efficiency In Horse Race Betting Markets, *in* 'Handbook of Sports and Lottery Markets', Handbooks in Finance, Elsevier, pp. 275–306.

Sweller, J. (1988), 'Cognitive Load During Problem Solving: Effects On Learning', *Cognitive Science* **12**(2), 257–285.

Thaler, R. H. & Ziemba, W. T. (1988), 'Anomalies: Parimutuel Betting Markets: Racetracks And Lotteries', *Journal of Economic Perspectives* **2**(2), 161–174.

Thorp, E. O. (2011), The Kelly Criterion In Blackjack Sports Betting, And The Stock Market, *in* L. C. MacLean, E. O. Thorp & W. T. Ziemba, eds, 'The Kelly Capital Growth Investment Criterion: Theory and Practice', World Scientific, Singapore, pp. 789–832.

Timmermann, A. & Granger, C. (2004), 'Efficient Market Hypothesis And Forecasting', *International Journal of Forecasting* **20**(1), 15–27.

Timmermans, D. (1993), 'The Impact Of Task Complexity On Information Use In Multi-Attribute Decision Making', *Journal of Behavioral Decision Making* **6**(2), 95–111.

Tversky, A. (1969), 'Intransitivity Of Preferences', *Psychological Review* **76**(1), 31–48.

Tversky, A. & Kahneman, D. (1973), 'Availability: A Heuristic For Judging Frequency And Probability', *Cognitive Psychology* **5**(2), 207–232.

Tversky, A. & Kahneman, D. (1974), 'Judgment Under Uncertainty: Heuristics And Biases', *Science* **185**(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1981), 'The Framing Of Decisions And The Psychology Of Choice', *Science* **211**(453-458), 4481.

Tversky, A. & Kahneman, D. (1986), 'Rational Choice And The Framing Of Decisions', *The Journal of Business* **59**(4), S251–S278.

Tversky, A. & Kahneman, D. (1992), 'Advances In Prospect Theory: Cumulative Representation Of Uncertainty', *Journal of Risk and Uncertainty* **5**(4), 297–323.

Von Luxburg, U. (2007), 'A Tutorial On Spectral Clustering', *Statistics and Computing* **17**(4), 395–416.

Von Neumann, J. & Morgenstern, O. (1944), *Theory Of Games And Economic Behavior*, Princeton University Press.

Wand, Y. & Wang, R. (1996), 'Anchoring Data Quality Dimensions In Ontological Foundations', *Communications of the ACM* **39**(11), 86–95.

Wang, R. & Strong, D. (1996), 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems* **12**(4), 5–33.

Watts, R. L. (1978), 'Systematic 'Abnormal' Returns After Quarterly Earnings Announcements', *Journal of Financial Economics* **6**(2-3), 127–150.

White, E. M., Dattero, R. & Flores, B. (1992), 'Combining Vector Forecasts To Predict Thoroughbred Horse Race Outcomes', *International Journal of Forecasting* **8**(4), 595–611.

Wilks, S. (1938), 'The Large-Sample Distribution Of The Likelihood Ratio For Testing Composite Hypotheses', *The Annals Of Mathematical Statistics* **9**(1), 60–62.

Wolfers, J. & Leigh, A. (2002), 'Three Tools For Forecasting Federal Elections: Lessons From 2001', *Australian Journal of Political Science* **37**(2), 223–240.

Wolfers, J. & Zitzewitz, E. (2004), 'Prediction Markets', *Journal of Economic Perspectives* **18**(2), 107–126.

Xu, Q., Huang, Q., Jiang, T., Yan, B., Lin, W. & Yao, Y. (2012), 'HodgeRank On Random Graphs For Subjective Video Quality Assessment', *IEEE Transactions on Multimedia* **14**(3), 844–857.

Yang, H., Lin, W., Deng, C. & Xu, L. (2014), Study On Subjective Quality Assessment Of Digital Compound Images, *in* 'IEEE International Symposium on Circuits and Systems (ISCAS)', IEEE, Melbourne, pp. 2149–2152.