Explorations in Pairwise Measures of Dependence and Pooled Significance

by

Chris Salahub

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: J Concepción Loredo-Osti
Professor, Dept. of Mathematics and Statistics,
Memorial University of Newfoundland

Supervisor: Wayne Oldford
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal Member: Ryan Browne
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal Member: Kun Liang
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal Member: Greg Bennett
Professor Emeritus, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal-External Member: Brian Ingalls
Professor, Dept. of Applied Mathematics,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the exploration of data sets with many variables, the search for interesting pairs is often the first step of analysis. This search builds a road map of the entirety of data before looking at its details, and can provide indispensable inspiration for deeper investigation. Challenges are present, however, in adjusting results to address the multiple testing problem and choosing a measure with sufficient generality to detect many forms of dependence.

This work proposes the measurement of statistical dependence by recursive binning of marginal ranks as a flexible measure of dependence. Simulation studies are used to characterize the null distribution and demonstrate the method's sensitivity to different data patterns. By splitting bins randomly, the χ^2 statistic has a null distribution conservatively approximated by the χ^2 distribution seemingly without a loss of power compared to maximized splitting rules, which has an inflated statistic value. The method is demonstrated on real S&P 500 constituent data.

To adjust for multiple testing, a new framework and coefficient are devised with appropriate proofs for analyzing pooled p-values based on their tendency to detect concentrated or diffuse evidence. This motivates a pooled p-value based on the χ^2 quantile function as a way to adjust for multiple testing while controlling the family-wise error rate and fine-tuning for the evidence pattern of interest. Simulation studies suggest this method is similarly powerful to the uniformly most powerful method while being more robust to mis-specification.

Both the recursive binning measurement of association and the χ^2 pooled p-value are then demonstrated for genetic data after a tutorial introducing the relevant genetic concepts. A method of moments adjustment of the χ^2 pooled p-value to account for correlation between tests is introduced and used with genomic and phenomic data from mice to identify regions of interest. The use of pooled p-values to combine parameter estimates in meta-analysis is also explored, establishing the concepts of evidential intervals and demonstrating their behaviour on simulated data.

Acknowledgements

Throughout my doctorate, I was helped greatly by my community of friends, family, and colleagues. I want to thank every one of them in Calgary for making the city still feel like home after all these years, those in Kitchener-Waterloo that make graduating so bittersweet, and those elsewhere who still enrich my life with their stalwart friendship.

Aside from his help refining the ideas presented here, I want to thank my supervisor, Wayne Oldford, for guiding me to some exceptionally interesting books on the history and philosophy of statistics. My perspective on the field (and on other topics) has been broadened and deepened by these recommendations, and I greatly appreciate his support of my interest in these topics. I'd also like to thank Ryan Browne, Greg Bennett, Kun Liang, Brian Ingalls, and J Concepción Loredo-Osti for reading my lengthy thesis and giving helpful suggestions. Though he changed positions and so was not on the final committee, I thank Marius Hofert for his early feedback.

For supporting this academic work, I'd like to thank all the Statistics and Actuarial Science support staff. Mary Lou Dufton deserves special mention for the guidance she gave me throughout my studies. Thanks to Shoja Chenouri for encouraging me to become more involved in the social life and administration of the faculty and Chi-Kuang Yeh for helping me to chair the student seminar series. My fellow students Marzieh Rizi, Dylan Spicker, Zhaohan Sun, Carlos Araiza, Nam-Hwui Kim, Pavel Shuldiner, Grace Tompkins, Augustine Wigle, Tatiana Krikella, Marcus Di Renzo, and Alex Bühler also deserve thanks for sharing their perspectives and time with me throughout my doctorate. They added an incredible variety, vibrancy, and vivacity to my days working in M3.

I want to thank my girlfriend, Ioana Crant, for her care and support. She went above and beyond to help when my appendix ruptured just a few weeks before my defence, and without her attentive care during my recovery I would have been much less prepared to defend.

The largest thanks go to my sister, Lisa, and my parents, Diane and Dave, who did everything they could to help my work whenever I was back in Calgary. I am incredibly grateful for this unwavering support (especially for the free room and board during my extended trips home) and for all of the lessons they taught me over the years that made this work possible.

Dedication

This is dedicated to my friends and family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The modern world is awash with data. Ubiquitous data collection throughout society has rendered data sets comprised of ever more variables and observations which must be sifted through for insight. This is true, for example, in finance and genetics, where real time measurement and the advent of complete sequencing have resulted in truly massive raw data. In finance, the identification of interesting pairwise variable relationships can motivate strategy or hedge risk, while the comparison of many genetic markers to a physical trait confers greater understanding of inheritance and the nature of some disease. How "interestingness" is measured in these and many other applications takes many different forms, such as the scagnostics of Wilkinson et al. (2005).

Perhaps the most widely used measure of interestingness is statistical dependence. A broad measure of the interestingness of a relationship between variables, it is defined as the absence of statistical independence and so includes a wide array of patterns. Consequently, many different measures devised for specific contexts and patterns of interest have been developed (e.g. Goodman and Kruskal, 1979; Liebetrau, 1983; Choi et al., 2010) since the early efforts of Galton and Pearson to characterize it with correlation and the χ^2 test (Stigler, 1989; Hald, 1998).

The rise of computers has changed the nature of these proposals drastically, resulting in more flexible and computationally complex measures than ever before. Many modern measures of dependence are products of the computer age, such as algorithmic ones based on nearest neighbours (Dümcke et al., 2014), prediction (Breiman, 2001), or the maximization of a particular data transformation (Breiman and Friedman, 1985; Reshef et al., 2011; Jiang et al., 2015; Liu et al., 2018).

This change of circumstances has precipitated similar changes in the motivation of these measures. Before computers, calculating a measure for a single data pair was a laborious task, simulation required creative physical devices, and results could only be shared slowly by modern standards. Distributional approximations were therefore necessary for early pairwise measures in their attempts to summarize the "true" dependence of one variable on another (Pearson, 1900; Greenwood and Yule, 1915; Cramér, 1924) as characterizing a measure in the null case to verify its behaviour was an almost insurmountable challenge. Indeed, it took two decades for Pearson's original formulation of the χ^2 test to be adjusted to the correct sample distribution[1].

Today, computers can calculate pairwise measures and generate null examples trivially, allowing us to explore the large volume of data presented to us more completely. This ease of application allows for a more agnostic view of the results. Rather than the bedrock of analysis, pairwise measures of dependence are computed across all pairs early in the analysis of complex data to highlight the most interesting relationships before more complex models are fit. Extreme values from among the vastly many summarized relationships are taken as an indication of a pattern to be investigated more fully.

Several problems arise with the use of pairwise measures dependence to search data in this way. First, it can be difficult to choose from the plethora of specialized measures available. Most measures perform best on patterns of a particular type, and may miss other interesting aspects of the data. This is not a trivial choice; choose incorrectly and the most pertinent aspects of the data may be overlooked. In cases where the relationships of interest have dependence patterns unknown *a priori*, this choice is particularly daunting. To measure dependence in contingency tables alone, Goodman and Kruskal (1979) outline dozens of measures of dependence to choose from.

This has led to the development of a collection of general measures of dependence. Computer-age candidates for such a general measure to detect any statistical dependence either take partitions of the sample space (Reshef et al., 2011; Jiang et al., 2015; Heller et al., 2016; Reshef et al., 2018) or apply some transformation to the data (Székely and Rizzo, 2009; Liu et al., 2018). Data can be partitioned and transformed somewhat arbitrarily, however, and so some selection critera is required. Generally, the choice is motivated by

[1] This chapter of statistical history is quite interesting, and an excellent summary is provided by Hald (1998). It is also possible to track the development of the χ^2 test directly by viewing Pearson (1900), Greenwood and Yule (1915), Yule (1922), and Fisher (1922). Following the initial χ^2 proposal in Pearson (1900), an inconsistency was noted for the 2×2 contingency table by Greenwood and Yule (1915): the same statistic computed through different intermediate steps implied different null degrees of freedom. This was resolved theoretically by Fisher (1922) at the same time as an experiment by Yule (1922) lent empirical credence to Fisher's correction. Though Pearson did not initially accept these results and used Biometrika to voice his dissent (Pearson, 1922, 1923), the broader statistical community correctly adopted Fisher's assertion that estimating moments restricts the degrees of freedom.

maximization of a statistic measuring difference between the transformed or partitioned result and that expected under independence.

Even using a general measure, the issue of incomparability between measures presents another problem. Data may be continuous or may come in ordinal or nominal categories and these different variable types, for example, can force the use of different measures. Even using general measures for each type may make comparisons unclear, as different conceptual frameworks or completely different scales can be hidden by the report of a single number commonly between 0 and 1. An example is furnished by Wilkinson and Wills (2008), who note that each scagnostic has a different distribution over $[0, 1]$ for uniform data. Absent context or experience, this makes the interpretation of a scagnostic value of 0.6 highly challenging and its comparison to other measures fraught with difficulty.

The p-value remedies this problem, and so remains an important and powerful tool to facilitate these comparisons. Not only does it measure the extremity of an observed value in context, but all p-values exist on the same scale with identical interpretations and uniform distributions over $[0, 1]$ when the null is true. It provides a value dependent on our assumptions and statistical theory which can nonetheless be compared to any other.

Finally, these many pairwise comparisons raise the issue of multiple testing. The number of pairwise comparisons grows rapidly in the number of variables, and values for a measure which would be exciting on their own become routine even under the null. When using pairwise p-values to guide analysis, care is required to avoid chasing spurious patterns down dead-end paths. This was recognized well before the modern data deluge, proposals from as early as Fisher (1932) and Pearson (1933) give simple ways to combine, or *pool*, independent p-values and assess their overall significance. Though the former has been widely adopted into practice, the following decades have seen many other proposals (Stouffer et al., 1949; Edgington, 1972; Mudholkar and George, 1977; Heard and Rubin-Delanchy, 2018; Wilson, 2019; Cinar and Viechtbauer, 2022).

1.1 Outline

The following chapters explore the very broad problem of using pairwise measures of dependence to sort and filter variables. As outlined, this requires consideration of several statistical problems, each with their own literature and conventions. Consequently, the literature and previous work will be outlined at the start of the relevant chapter or in a chapter immediately preceeding. As a highly applied topic, three chapters are fully dedicated to applications. In particular, Chapter 5 develops the motivating example of genomic studies. Though more pertinent theoretical work is presented first, this motivating example could be read before the rest.

Genetics, in particular exploratory genome-wide association studies, are a natural application for the pairwise measure of independence to filter variables. There are tens of thousands of genes that may affect a given trait or condition, and sifting through these for the most interesting ones is a common problem in, for example, genome-wide association studies. Chapter 5 clarifies the structure of this data and establishes a simple model of genetics that informs a derivation of genetic correlation. This introduction is accompanied by a package in R, `toyGenomeGen`, which allows for experimentation using the genetic model presented. Real genetic data processed from the Mouse Genome Database (Bult et al., 2019) is used to compare derived correlation under this simple model to that observed. Results displayed using a custom plot matrix communicating the observed and theoretical values of correlation along with their distribution under repeated simulation indicate a reasonable fit.

Of course, the problem of leveraging pairwise measurements applies widely beyond genetics. A brief survey of measures of interestingness from other contexts is presented at the beginning of Chapter 2 before focus is placed on dependence due to its universal interest. This motivates the introduction of functional measures of dependence evaluating departures of the joint distribution of pairs $f_{X,Y}(x, y)$ from the product of marginal distributions $f_X(x)f_Y(y)$ and the evaluation of these functional measures by modern bin-based algorithmic measures. In these, a global measure of dependence $D(X, Y)$ is split into measures applied to each bin $d(X, Y)$ and then summarized. Several measures are expressed as a sum over all partitions of a function measuring the departure of each individual partition's count from that expected under independence, with a primary difference distinguishing them the splitting logic or penalty function added to the measure of dependence.

Chapter 3 expands on bin-based measures by introducing a measure of dependence based on recursive binary partitions of the pairwise sample space modelled on the recursive binary partitioning of Rahman (2018). By first converting all variables to marginal ranks, expected counts within each bin can still be determined under this flexible binning method. This algorithm is sketched, an iterative implementation is detailed, and the corresponding R package `AssocBin` is introduced. A proof supports the consideration of splits at points alone. Extensive simulations are used to obtain true p-values for the χ^2 statistic applied to recursively partitioned bins under different settings. These indicate that random splits produce a statistic with a null distribution conservatively modelled by the χ^2 without a loss of power against different simulated data patterns compared to maximized splits. In contrast, maximizing splits produce inflated statistic values relative to the χ^2 distribution and an effective visual summary of the data. The algorithm is finally applied to S&P 500 constituent stock data and compared to a previous analysis of the pairwise relationships from Hofert and Oldford (2018).

Chapter 4 discusses pooled p-values as a way to control the family-wise error rate of $M > 1$ p-values arising from independent tests. Previous work is summarized after establishing a series of telescoping hypotheses. These telescoping hypotheses are organized by the prevalence of non-null p-values (measured by the proportion η of non-null tests in the collection) and the strength of evidence provided by each non-null test (measured the Kullback-Leibler divergence of the p-value distribution of the test from uniformity). A simulation study carried out to investigate the performance of the uniformly most powerful (UMP) test from Heard and Rubin-Delanchy (2018) using these telescoping hypotheses suggests a framework for pooled significance based on the detection of either concentrated or diffuse evidence. Several proofs develop this concept and culminate in a centrality coefficient in $[0, 1]$ which communicates the preference of a pooling function for diffuse or concentrated evidence. A pooled p-value based on the χ^2 quantile function is proposed to control this coefficient. By changing the degrees of freedom, it is proved that a pooled p-value with arbitrary centrality coefficient can be obtained for any M. Furthermore, simulation studies indicate that the χ^2 pooled p-value is more robust than the UMP to mis-specification, and can be leveraged to provide information on the plausible alternative distributions that generated a collection of p-values. Functionality to compute the centrality quotient, UMP, and χ^2 pooled p-value are implemented in the `R` package `PoolBal`. Note that Chapter 4 considers only the combination of p-values, and so the resulting method and insight for multiple testing adjustment is highly general. Any group of methods producing p-values can make use of these findings.

After the theoretical chapters, the remainder of the work considers particular examples. Chapter 5, explained earlier, establishes a genetic model which is used immediately in Chapter 6. Marginal and central rejection from Chapter 4 are shown to correspond neatly to patterns of oligogenic inheritance for linear traits in genetics. This is demonstrated in simulation studies before an investigation of real genetic data from the Mouse Genome Database (Bult et al., 2019). In both the real and simulated data, the χ^2 pooling method is adjusted to account for dependence using the method of moments with a Satterthwaite approximation and a large simulation study. The resulting adjustment seems to correct the level of the pooled p-value without impacting its conclusions, suggesting a robustness of the χ^2 pooling function to dependence.

Chapter 7 presents the most exploratory and experimental application of pooling p-values. Recognizing that the current glut of data implies a similar abundance of analyses, a short foray into meta-analysis based on the χ^2 pooling function of Chapter 4 is undertaken. A novel method of combining parameter estimates between studies is proposed

which considers many possible candidate parameters and tests each. This results in a region of plausible values where we would fail to reject the pooled p-value at threshold α that simultaneously suggests plausible parameter values and tests whether observed estimates could have arisen from the same population parameter. Evidential inference of this sort is outlined and the χ^2 pooled p-value is explored for this purpose. Simulation studies indicate that changing the degrees of freedom changes the treatment of outliers by evidential inference an impacts the coverage probability and probability of rejecting homogeneity. Power investigations under common settings suggest that this new method of combining parameter estimates performs similarly to more classical methods.

Chapter 2

Measuring Interestingness

At the core of this work is the use of pairwise measures to highlight interesting variables in a data set of many variables, and so it is natural to begin with an overview of these measures. First, notation is established and terms are defined.

In general, "interestingness" is a subjective term. Conceivably any pattern could be interesting in the right exploratory context, and so many different measures capturing many different patterns exist. These are briefly discussed in the following chapter before statistical dependence is explored in more detail. This focus on statistical dependence is motivated by the observation that it is almost always interesting in an exploratory context. When sifting through many variable pairs to select those worthy of further investigation, a statistical dependency between any two provides information about how they should be controlled or modelled in any investigation.

The exploration of statistical dependence eventually focuses on the evaluation of functionals that compare the observed joint density to the product of marginal densities, as these address the problem of dependence directly using its definition.[1] In particular, *bin-based* methods which partition the pairwise sample space are considered, as these automatically perform non-parametric estimation of the joint and product marginal densities. To discuss these methods, all are placed in a common notation which decomposes a global statistic as the sum of local statistics computed on each bin independently.

[1] This contrasts the organization of Tjøstheim et al. (2022), an up-to-date survey of meaures of dependence. While Tjøstheim et al. (2022) separate measures based on copulas, kernel functions, and partitioning, these are all cast here as particular instances of functionals comparing the joint and product marginal densities to emphasize how similar the computation of these conceptually different measures is in practice.

2.1 Preliminaries

Let X and Y be a pair of random variables with distribution functions $F_X(x)$ and $F_Y(y)$ over the domains $\mathcal{X}$ and $\mathcal{Y}$ both respectively. Denote the joint distribution of the pair X, Y as $F_{X,Y}(x, y)$, and the conditional distributions $F_{X|Y}(x|y)$ and $F_{Y|X}(y|x)$. To remain fully general, note that both of X and Y may be continuous or discrete. Let $f_X(x)$, $f_Y(y)$, $f_{X,Y}(x, y)$, $f_{X|Y}(x|y)$, and $f_{Y|X}(y|x)$ be the corresponding probability densities or probability mass functions in the continuous or discrete cases. We say that X and Y are independent and write $X \perp\!\!\!\perp Y$ if and only if their joint sample space is a Cartesian product and

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \tag{2.1}$$

or equivalently $f_{X|Y}(x|y) = f_X(x)$ and $f_{Y|X}(y|x) = f_Y(y)$.

Define a K-dimensional copula as a distribution function $C(u_1, u_2, \dots, u_K)$ over $[0, 1]^k$ with uniform marginal distributions for all u_k as in Embrechts et al. (2001). By Sklar's Theorem, for any K continuous random variables there exists a unique copula C such that

$$F_{X_1,X_2,\dots,X_K}(x_1, x_2, \dots, x_K) = C\big(F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_K}(x_K)\big), \tag{2.2}$$

that is the joint distribution can be summarized by a copula on the marginals transformed to be uniformly distributed. In other words: the relationship between the variables is uniquely determined by C and the marginals alone. Define the independence copula

$$C_I(u_1, u_2, \dots, u_K) = \prod_{k=1}^{K} u_k. \tag{2.3}$$

In particular, consider the copula of X and Y, $C\big(F_X(x), F_Y(y)\big) = C(u, v)$, and the bivariate independence copula, $C_I(u, v) = uv$.

In practice these theoretical quantities are unknown. Instead, all that is available is a sample of paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ of (X, Y). Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^\mathsf{T}$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^\mathsf{T}$ be the observed values of X and Y respectively. For X and Y which can be ordered, let a subscript (x) indicate a non-decreasing sorting of elements with respect to $\mathbf{x}$ so that $\mathbf{x}_{(x)}$ is the vector $\mathbf{x}$ in increasing order and $\mathbf{y}_{(x)}$ is the vector $\mathbf{y}$ sorted in increasing order of $\mathbf{x}$. Elementwise, follow the convention

$$\mathbf{x}_{(x)} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})^\mathsf{T}$$

to denote the elements of $\mathbf{x}_{(x)}$ and analogously for $\mathbf{y}_{(y)}$. Define the rank function on a marginal sample $\mathbf{x}$ as

$$r(x; \mathbf{x}) = \sum_{i=1}^{n} I_{(-\infty, x]}(x_i), \tag{2.4}$$

where

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

is the usual indicator function so that $r(x_i, \mathbf{x})$ gives the index of x_i in $\mathbf{x}_{(x)}$. Note that this definition assumes $x_i \neq x_j$ for all $i \neq j$, the case of no ties.

To consider different conventions to address ties, suppose $x_1 = x_2 = \cdots = x_m = x_{(1)}$ for $m < n$. In this case, $r(x_1; \mathbf{x}) = r(x_2; \mathbf{x}) = \cdots = r(x_m; \mathbf{x}) = m$, that is all are given the maximum index m. It is not clear this must be the case, as choosing the minimum index 1 seems equally valid for these first m observations. Another option is random tie-breaking, which randomly assigns the indices $1, 2, \ldots, m$ to $x_1, x_2, \ldots, x_m$. Random tie breaking is of particular interest, as it induces a uniform distribution on the ranks for tied regions. For applications where complete ranks are necessary, this is more desirable than the gaps introduced by the maximum or minimum indexing.

We can then define the empirical distribution function for these ordered cases as

$$\widehat{F}_{\mathbf{x}}(x) = \frac{1}{n} r(x; \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i), \tag{2.5}$$

with $\widehat{F}_{\mathbf{y}}(y)$ defined similarly. The empirical copula of X and Y is defined as

$$\widehat{C}(u, v) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, u]}\left(\frac{r(x_i; \mathbf{x})}{n}\right) I_{(-\infty, v]}\left(\frac{r(y_i; \mathbf{y})}{n}\right). \tag{2.6}$$

Finally, define the vector-valued rank function as

$$\mathbf{r}(\mathbf{x}) = (r(x_1; \mathbf{x}), r(x_2; \mathbf{x}), \ldots, r(x_n; \mathbf{x}))^{\mathsf{T}}. \tag{2.7}$$

For discrete X without ordered values, take an arbitrary numbering of the possible values in $\mathcal{X}$ and let $I = |\mathcal{X}|$. Otherwise, number such that the assigned values reflect the ordering. If Y is discrete, number $\mathcal{Y}$ similarly with $J = |\mathcal{Y}|$. Then, for $\mathbf{x}$ and $\mathbf{y}$, define the observed marginal and joint counts

$$o_{i+} = \sum_{j=1}^{J} \sum_{k=1}^{n} I_{\{(i,j)\}}((x_k, y_k)) = \sum_{j=1}^{J} o_{ij} \tag{2.8}$$

and o_{+j} analogously. These give corresponding probability estimates $\widehat{p}_{ij}$, $\widehat{p}_{i+}$, and $\widehat{p}_{+j}$ when divided by n.

Distinguish three different *types* of X and Y based on $\mathcal{X}$ and $\mathcal{Y}$. The three types are:

continuous some interval or collection of intervals of real numbers;

ordinal categorical discrete possibilities with an ordering (e.g. income brackets);

nominal categorical discrete possiblities without an ordering (e.g. country of birth).

The possible combinations of these lead to six unique pairwise combinations. Following the convention of Lee and Huh (2003), call comparisons of data of the same type *simple* comparisons. *Complex* comparisons refer to the unique pairings between types: continuous versus ordinal, continuous versus nominal, and ordinal versus nominal.

Introduce the function

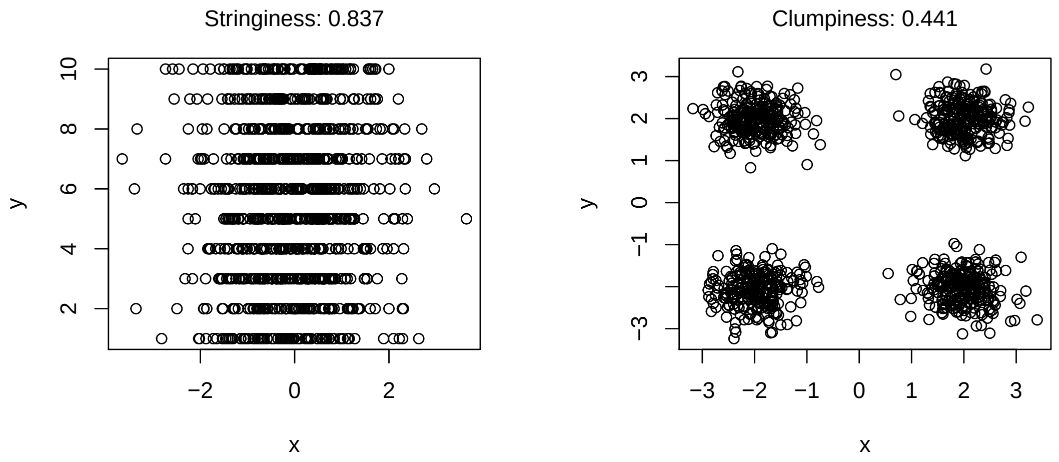
$$G : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{R}$$

quantifying some pattern of interest between X and Y with a value in $\mathcal{R} \subset \mathbb{R}$. For the purpose of this work, any such G is called a *measure of interestingness* or simply a *measure*. The term 'interestingness' is used instead of 'association' following the tradition of researchers such as John W. Tukey (Friedman and Stuetzle, 2002) and to avoid the implications of 'association,' typically used to describe *statistical dependence* as in Equation (2.1).

Consider, for example, Tukey's scagnostic measures of structure as reintroduced in Wilkinson et al. (2005). Two different constructed data sets exemplifying the scagnostics for "stringiness" and "clumpiness" respectively are shown in Figure 2.1. Both of these simulated data sets have margins generated independently of each other, and still have relatively large values for their respective measures compared to the null distributions in Wilkinson and Wills (2008). While scagnostics have been expressly developed to capture patterns other than statistical independence, calling them measures of association confuses the term.[2] Another example is λ_b from Goodman and Kruskal (1979), which is motivated by prediction and not strictly tied to dependence.

Numerous frameworks have been proposed to evaluate candidates for G. A statistical framework is given in Rényi (1959), and this framework has been subsequently updated by Schweizer and Wolff (1981) and Móri and Székely (2019). Reimherr and Nicolae (2013)

[2] See, for example Liu et al. (2018) utilizing data analogous to Figure 2.1(b) as an example of 'association' in a paper motivated by statistical independence.



(a) Striated/stringy sample

(b) Clumpy sample

Figure 2.1: Data exhibiting certain scagnostic structures despite the independent generation of x and y.

present a slightly different framework which emphasizes interpretability and delineates three different motivations for measuring association. Here, the focus is on the *ordering* G induces on $\mathcal{X} \times \mathcal{Y}$, as this can accomplish all of the goals from Reimherr and Nicolae (2013) and is critical to the search for interesting variables.

The importance of ordering restricts most measures of interestingness to a range on a finite interval $\mathcal{R} = [g_{\min}, g_{\max}] \subset \mathbb{R}$. The upper bound is obtained for X and Y exemplary of the pattern of interest, while $g_{\min}$ is less consistent. For signed measures, such as Pearson's correlation coefficient, $g_{\min}$ may still indicate perfect correspondence to a particular pattern. In unsigned measures $g_{\min}$ suggests no indication of the pattern[3]. The X and Y which lead to these extremes are typically not unique, but rather represent a family of patterns which the measure does not distinguish.

Commonly, $G(X, Y)$ is scaled by $\max\{|g_{\max}|, |g_{\min}|\}$ such that $\mathcal{R}$ is restricted to $[-1, 1]$ or $[0, 1]$. Scaling G by the magnitude of its most extreme value makes the ordering it imposes explicit. For any pair X and Y, this scaled measure communicates directly how the pair compares to the perfect example. This scaling can be misleading, however, when

[3] Any signed measure G can, of course, be made unsigned by taking $|G|$.

the analyst is unfamiliar with its distribution along its range.

2.2 Measuring dependence

As noted by Cramér (1924); Fairfield Smith (1957); Goodman and Kruskal (1979) and likely others, measures are devised to capture specific patterns. This is often desirable for interpretability, as Reimherr and Nicolae (2013) note, and can lead to descriptive adjectives which evoke a measure, as in the scagnostics of Wilkinson et al. (2005). Some common patterns and corresponding measures are:

linearity which is typically measured by Pearson's product moment correlation;

monotonicity captured by Spearman's ρ, the rank version of correlation;

concordance measured by Kendall's τ and Goodman and Kruskal's γ (Goodman and Kruskal, 1979);

predictibility measured by Goodman and Kruskal's λ (Goodman and Kruskal, 1979);

agreement quantified by Cohen's κ measure of inter-rater reliability (Cohen, 1960);

and others listed in Liu et al. (2018); Liebetrau (1983); Agresti (1981); and Goodman and Kruskal (1979).

The variety of patterns which one might want to measure and the different types of X and Y have led to a great proliferation of bivariate measures of interestingness[4], each with its own interpretation. Unfortunately, the division of measures by both type and pattern makes the search for interesting patterns in complex data far more challenging, with research such as Khamis (2008) and Lee and Huh (2003) devoted to the task of guiding practitioners to choose the most comparable measures between types. In practice, correlation is often applied to all variable pairs regardless of type, despite its well-known shortcomings (Tjøstheim et al., 2022; Reshef et al., 2011; Anscombe, 1973).

Regardless of motivation or type, however, X and Y satisfying Equation (2.1) are universally considered uninteresting. As a consequence, many complex measures of interestingness attempt to measure departures from statistical independence. These measures

[4] The 2×2 contingency table, for example, has an almost overwhelming roster of measures analyzed in Choi et al. (2010).

of statistical dependence are functionals of the form

$$D\big(f_{X,Y}(x,y), f_X(x)f_Y(y)\big) \tag{2.9}$$

which compare the joint and marginal product distributions of X and Y. Often, they will attempt to satisfy the desiderata outlined by Rényi (1959) or Schweizer and Wolff (1981), the latter of which relaxed the axioms of the former after noting they are unnecessarily restrictive and hard to apply in practice. Some examples for continuous X and Y include

$$\mathit{\Delta}(X,Y) = \int\int_{\left\{(x,y):f_{X,Y}(x,y)\geq f_X(x)f_Y(y)\right\}} \big[f_{X,Y}(x,y) - f_X(x)f_Y(y)\big]dxdy \tag{2.10}$$

from Silvey (1964) and the mutual information defined as

$$\mathcal{I}(X,Y) = D_{KL}(f_{X,Y}||f_Xf_Y) = \int_{\mathcal{Y}}\int_{\mathcal{X}} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)dxdy \tag{2.11}$$

from Shannon (1948), where $D_{KL}(F||G)$ denotes the Kullback-Leibler divergence of G from F. For discrete X and Y a classic example is the χ^2 statistic for independence from Pearson (1900),[5]

$$\mathcal{D}(X,Y) \propto \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{[f_{X,Y}(x,y) - f_X(x)f_Y(y)]^2}{f_X(x)f_Y(y)}. \tag{2.12}$$

Another possible functional based on cumulative distribution functions is given by Hoeffding (1948):

$$D_H(X,Y) = \int \left(F(x,y) - F(x,\infty)F(\infty,y)\right)^2 dF(x,y). \tag{2.13}$$

Central to all of these measures is the factorization definition of independence from Equation (2.1). In every case, some comparison is made between $f_{X,Y}(x,y)$ and $f_X(x)f_Y(y)$ which is zero when the two are equal almost everywhere.

For any function of the form in Equation (2.9), an equivalent measure exists in the copula space. By Sklar's Theorem, discussed in Sklar (1996) and summarized in Embrechts et al. (2001), the dependence between X and Y can be captured by the marginal distribution functions F_X and F_Y and their copula C. Just as in Equation (2.1), X and Y are independent only if their copula is the independence copula $C_I(u,v) = uv$. Schweizer and

[5] Note that this formulation of $\mathcal{D}(X,Y)$ is somewhat atypical, as this is a quantity usually defined on the sample $\mathbf{x}, \mathbf{y}$.

Wolff (1981) therefore show that a transform can be applied to any measure formulated in $F_X(x)$, $F_Y(y)$, and $F_{X,Y}(x,y)$ to convert it to a functional of the form

$$D^*\Big(C\big(F_X(u), F_Y(v)\big), uv\Big).$$

Such transforms are attractive because they remove the marginal distributions of the variables being compared. This allows for non-parametric estimation of quantities without presuming any particular distribution. Indeed, there are many measures which utilize empirical copulas (Ding et al., 2017; Siburg and Stoimenov, 2010; Genest and Rémillard, 2004).

Measures like the distance covariance of Székely and Rizzo (2009) instead apply a functional of the form in Equation (2.9) to the characteristic functions of X and Y. The characteristic function of X is defined as

$$\phi_X(t) = E[e^{itX}], \tag{2.14}$$

where E is the expectation operator with respect to X. The joint characteristic function, $\phi_{X,Y}(t,s)$, and characteristic function of Y, $\phi_Y(s)$, are defined similarly. This is based on an important result of Equation (2.1), that $\phi_{X,Y}(s,t) = \phi_X(t)\phi_Y(s)$ if and only if $X \perp\!\!\!\perp Y$.

Equation (2.9) is really just an evaluation of the goodness of fit of $f_X(x)f_Y(y)$ to $f_{X,Y}(x,y)$, and this creates an obvious analogy to empirical distribution function goodness of fit tests as described in Zheng et al. (2021) and Stephens (1974). Many common measures of functional distance appear in the literature of independence tests, such as the Kolmogorov-Smirnov test in Heller et al. (2016) and results for any L_P distance in Schweizer and Wolff (1981). These can be made even more general by replacing Euclidean distances by kernel distances, as in Liu et al. (2018) or the Hilbert-Schmit independence criterion (HSIC) of Gretton et al. (2007). While Liu et al. (2018) maximize over a prespecified set of kernel distances, Lopez-Paz et al. (2013) instead introduce a measure which randomly transforms X and Y and takes the maximum of the applied random transforms.

More unique applications of goodness of fit principles are found in Dümcke et al. (2014) and Heller et al. (2013). Dümcke et al. (2014) utilize the exact distribution of nearest neighbour distances under independence to develop two novel tests. Rather than comparing the joint distribution to a product of marginals, their tests are based on the deviation between the exact distribution of nearest neighbours and that observed in a sample. Heller et al. (2013) make use of local distances about each point in turn to construct a series of contingency tables and then aggregate the p-values gained.

The generation of numerous tables evokes a relevant class of measures based on the application of Equation (2.9) to partitions of the outcome space $\mathcal{X} \times \mathcal{Y}$ and aggregation of the results. Such partitioning allows for local estimation of $f_{X,Y}$ and $f_X f_Y$ without a parametric family. Rather, the estimate relies on the choice of partition. Another benefit of such partitioning is it permits the application of the same test to any data type, as partitioning continuous data produces ordinal categorical data. As a consequence of this potential and their popularity in recent literature, this work primarily focuses on these methods.

2.3 Bin-based measures

Though it has become more popular as it has become computationally feasible, partitioning, or *binning*, has always played a role in measuring association. See, for example, the early investigations into the χ^2 test outlined in Plackett (1983). Rather than using X and Y directly, a binning function can be applied to their marginal values in order to discretize them.[6] Discussing a measure which induces bins only makes sense when at least one of X and Y is continuous, though the results can be applied to categorical cases.

A univariate binning on J bins is a function $b : \mathcal{B} \mapsto \{1, \dots, J\}$ which partitions its continuous domain $\mathcal{B} \subseteq \mathbb{R}$ into J distinct parts, or *bins*. Any such b has a vector-valued version $\mathbf{b} : \mathcal{B}^p \mapsto \{1, \dots, J\}^p$ such that $\mathbf{b}(\mathbf{x}) = (b(x_1), \dots, b(x_n))^\intercal$ for $\mathbf{x} \in \mathcal{B}^p$. Consider applying $b_X : \mathcal{X} \mapsto \{1, \dots, I\}$ to X and $b_Y : \mathcal{Y} \mapsto \{1, \dots, J\}$ to Y. That is apply a binning on I bins to X and a binning on J bins to Y. This is equivalent to an $I \times J$ grid on $\mathcal{X} \times \mathcal{Y}$ so that the bins can be indexed by (i, j) to correspond with $\big(b_X(X), b_Y(Y)\big)$. Define

$$\epsilon_{ij} = n \int_{\left\{x : b_X(x) = i\right\}} dF_X(x) \int_{\left\{y : b_Y(y) = j\right\}} dF_Y(y), \tag{2.15}$$

the expected count of observations of n which fall into bin (i, j) under independence. Note that in the case of uniform X and Y, this simplifies to n times the area of the (i, j) bin.

In practice the vectors $\mathbf{x}$ and $\mathbf{y}$ are all that is observed, so define the analogous sample quantity

$$e_{ij} = \frac{1}{n} \sum_{k=1}^{n} I_{\{i\}} \left(b_X(x_k)\right) \sum_{l=1}^{n} I_{\{j\}} \left(b_Y(y_l)\right). \tag{2.16}$$

[6] In the case where one of X and Y is already categorical, this reduces to the K-sample problem. Some works, such as Heller et al. (2016), switch freely between the K-sample problem and the problem of measuring association.

The observed (i, j) bin count is given by

$$o_{ij} = \sum_{k=1}^{n} I_{\{(i,j)\}}\left(\left(b_X(x_k), b_Y(y_k)\right)\right), \tag{2.17}$$

so $e_{ij} = n\frac{o_{i+}}{n}\frac{o_{+j}}{n}$ using the notation of Equation (2.8). Under this binning, X and Y are converted to the contingency table in Table 2.1. Once so binned, the o_{ij} and e_{ij} can be used to in place of the unknown densities in Equations (2.10), (2.11), and (2.12) nonparametrically. In this way, a grid is simply a particular kind of two-dimensional histogram.[7] For Silvey's $\mathit{\Delta}$ from Equation (2.10), the analogue on the binned data in Table 2.1 is

	$b_Y(y) = 1$	$b_Y(y) = 2$	$\ldots$	$b_Y(y) = J$	
$b_X(x) = 1$	o_{11}	o_{12}	$\ldots$	o_{1J}	o_{1+}
$b_X(x) = 2$	o_{21}	o_{22}	$\ldots$	o_{2J}	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$b_X(x) = I$	o_{I1}	o_{I2}	$\ldots$	o_{IJ}	o_{I+}
	o_{+1}	o_{+2}	$\ldots$	o_{+J}	n

,

Table 2.1: The contingency table imposed on $\mathbf{x}$ and $\mathbf{y}$ by the binnings $b_X(x)$ applied to $\mathbf{x}$ and $b_Y(y)$ applied to $\mathbf{y}$.

$$\mathit{\Delta}(\mathbf{b_x}, \mathbf{b_y}) = \sum\sum_{\{(i,j):o_{ij} \geq e_{ij}\}} \frac{o_{ij} - e_{ij}}{n} = \sum\sum_{\{(i,j):o_{ij} \geq \frac{1}{n}o_{i+}o_{+j}\}} \frac{no_{ij} - o_{i+}o_{+j}}{n^2}; \tag{2.18}$$

Shannon's mutual information from Equation (2.11) becomes the multinomial log-likelihood ratio

$$\mathcal{I}(\mathbf{b_x}, \mathbf{b_y}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{o_{ij}}{n} \log\left(\frac{o_{ij}}{e_{ij}}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{o_{ij}}{n} \log\left(\frac{no_{ij}}{o_{i+}o_{+j}}\right); \tag{2.19}$$

and Pearson's χ^2 measure from Equation (2.12) becomes

$$\mathcal{D}(\mathbf{b_x}, \mathbf{b_y}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(no_{ij} - o_{i+}o_{+j})^2}{o_{i+}o_{+j}}, \tag{2.20}$$

[7] There are many other possible tessellations which produce a two-dimensional histogram density estimate, see Carr et al. (1987) and Scott (1988)

where $\mathbf{b}_\mathbf{x} = \mathbf{b}_X(\mathbf{x})$ and $\mathbf{b}_\mathbf{y} = \mathbf{b}_Y(\mathbf{y})$ are used for brevity. While $\mathbf{b}_X$ and $\mathbf{b}_Y$ can be arbitrarily defined, Equations (2.18), (2.19), and (2.20) depend only on the counts in each cell of Table 2.1. Therefore, only the values of $\mathbf{x}$ and $\mathbf{y}$ need to be considered as possible partition boundaries, or *bin edges*. Considering edges between every point, and not allowing for identical $\mathbf{x}$ and $\mathbf{y}$ values, this means are $n - 1$ possible bin edges to consider in each of $\mathbf{x}$ and $\mathbf{y}$. Noting that the bin edges can either be present or absent, there are therefore $2^{n-1}2^{n-1} = 4^{n-1}$ possible contingency tables for a given $\mathbf{x}$ and $\mathbf{y}$.

There is often no a priori exploratory reason to choose a particular binning, and so methods from Reshef et al. (2011); Jiang et al. (2015); Heller et al. (2016); and Reshef et al. (2018) have the exploration of this large number of grids at their core. Reshef et al. (2011) and Reshef et al. (2018), in the introduction of the maximal information criterion (MIC), propose computing Equation (2.19) for all grids such that $IJ \leq n^{0.6}$, scaling these values, and storing them in a matrix M. The maximal value of this matrix is then taken as the MIC.[8] In order to contextualize MIC values, p-values based on simulated null data sets are computed.

Jiang et al. (2015) propose a penalized version of Equation (2.19) to find a solution. The particular penalty is outlined in Equation (2.24). Conceptually, this penalized optimization assumes a Poisson distribution on the number of bins for one of the margins conditioned on the other and then maximizes the likelihood dynamically. As with the MIC, the null distribution of this method is determined empirically.

Hoeffding (1948) restricts the exploration only to $I = J = 2$, and proposes a sum of Equation (2.13) for all unique possible two-by-two grids, which Thas and Ottoy (2004) show is asymptotically equivalent to Equation (2.20) evaluated over all two-by-two grids with a suitable scaling. Heller et al. (2016) extend this by simply restricting the grid to m divisions on both margins. They investigate both summation and maximization for either of Equation (2.20) and (2.19) across all possible $m \times m$ grids. In the case of summation, the values over all possible $m \times m$ grids are computed and the sum is returned, while the maximization case reports only the largest of these values. The distributions of these statistics is computed empirically just as for the MIC and Jiang et al. (2015). The language used, where aggregation is compared to the Cramer-von Mises criterion and maximization to the Kolmogorov-Smirnov test, is evocative of methods for testing empirical distribution functions (Stephens, 1974).

[8] It should be noted that even this space is too large to fully explore, and so only a small number of grids are actually computed in practice. Their concept of *equitability* has also generated considerable controversy. See the discussions in Gorfine et al. (2012); Kinney and Atwal (2014a); Reshef et al. (2014); Kinney and Atwal (2014b); and Simon and Tibshirani (2014).

To implement these methods, all of Heller et al. (2016); Jiang et al. (2015); and Reshef et al. (2011) utilize similar recursive algorithms. While Heller et al. (2016) compute their summation statistic directly, as this reduces to a counting problem, they borrow the procedure of Jiang et al. (2015) with a different penalty for the maximization case. Reshef et al. (2011) use a similar concept to guide their exploration of the space of grids with $IJ \leq n^{0.6}$.

For a more detailed discussion of this recursive algorithm and its applications, introduce the notation

$$D(\mathbf{b}_\mathbf{x}, \mathbf{b}_\mathbf{y}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d(o_{ij}, e_{ij}), \tag{2.21}$$

thereby expressing D over the entire data set as the sum of d applied to each cell. Introduce the subscript notation

$$\mathbf{b}_{\mathbf{x}[1:k]} = \big(b_X(x_1), b_X(x_2), \dots, b_X(x_k)\big)^\intercal$$

to denote a binning on I bins applied to the first k elements of $\mathbf{x}$. Consider the value of Equation (2.21) applied to the first k observations of $\mathbf{x}$ and $\mathbf{y}$ with a given binning on $\mathbf{y}$, written

$$D_k(\cdot|\mathbf{b}_\mathbf{y}) : \{1, \dots, I\}^k \mapsto \mathbb{R} \tag{2.22}$$

for $I \leq k$ and a constant binning $\mathbf{b}_\mathbf{y}$ on J bins defined for all of $\mathbf{y}$ (not just the first k observations).

Equation (2.22) presents an important modification of Equation (2.21). By viewing only the first k elements with a pre-specified $\mathbf{b}_\mathbf{y}$, the problem of identifying a binning which optimizes D_k is much simpler than for D. Rather than selecting among all possible grids, only a small number at each step need to be considered. Suppose $\mathbf{b}^*_{\mathbf{x}[1:k]}$ is the binning on the first k elements of $\mathbf{x}$ which maximizes D_k, define

$$D^*_k = D_k(\mathbf{b}^*_{\mathbf{x}[1:k]}|\mathbf{b}_\mathbf{y}),$$

as the maximal value of Equation (2.22). It is therefore the maximal value of D applied to the first k observations of $\mathbf{x}$ and $\mathbf{y}$ given a known binning on $\mathbf{y}$.

Using this general notation, the recursive estimate $\widehat{D}^*_k$ used by Reshef et al. (2011); Jiang et al. (2015); and Heller et al. (2016) is

$$\widehat{D}^*_k = \max_{1 \leq i < k} \left[\widehat{D}^*_{i-1} + \sum_{j=1}^{J} d \left(\sum_{l=i}^{k} I_{\{j\}}\left(b_Y(y_l)\right), \frac{k-i}{n} \sum_{l=1}^{n} I_{\{j\}}\left(b_Y(y_l)\right) \right) \right]. \tag{2.23}$$

The edges which give these optimal binnings are given by replacing the max with an arg max. The arguments inside d are the observed counts of each Y bin from i to k and the expected counts based on the marginal distribution and the relative length of the interval from i to k, respectively.

For each $i \in \{1, \dots, k-1\}$, this estimate considers two parts. The first part, $\widehat{D}^*_{i-1}$, is the previously computed maximal measure on the first $i-1$ points with the convention $\widehat{D}^*_0 = 0$. The second part is a sum of d over all J bins of $\mathbf{b_y}$ with observed counts given by the incidence of each of the J bins from i to k and the expected counts given by the marginal distribution of $\mathbf{b_y}$ multipled by the length of the interval from i to k.

Therefore, this algorithm chooses to add a bin edge at the i which maximizes Equation (2.21) *conditional on previous bin edges.* Additionally, this estimate requires a *pre-specified* $\mathbf{b_y}$, and so it does not give a global optimum over all grids. Indeed, it still requires some ad hoc choice of binning on $\mathbf{y}$. Jiang et al. (2015) suggests using the slicing methods of Jiang and Liu (2013) to bin $\mathbf{y}$ while Reshef et al. (2011) suggests a simple equipartition on $\mathbf{y}$. The former chooses bins to optimize the difference between the conditional and unconditional variances while the latter is computationally easy to implement.

The main difference between the implementations in Jiang et al. (2015), Heller et al. (2016), and Reshef et al. (2011) is the choice of d. Jiang et al. (2015) takes a penalized measure

$$d(o_{ij}, e_{ij}) = \frac{o_{ij}}{n} \log\left(\frac{o_{ij}}{e_{ij}}\right) - \frac{\lambda_0}{J} \log n \tag{2.24}$$

where λ_0 is a penalty parameter and $\mathbf{b_y}$ is a binning on J bins. This is equivalent to a prior distribution of the bin edges giving so that the number of bins follows a Poisson distribution. Heller et al. (2016) instead take

$$d(o_{ij}, e_{ij}) = \frac{o_{ij}}{n} \log\left(\frac{o_{ij}}{e_{ij}}\right) + \frac{\lambda_0}{J} \log\binom{n-1}{k-1}, \tag{2.25}$$

where k matches the index $\widehat{D}^*_k$. This modification is equivalent to a uniform distribution over all possible marginal binnings. Under both versions of the penalized d, splits are then considered at each x_k in turn and the bins which maximize the penalized score are used. In contrast, Reshef et al. (2011) take the unpenalized

$$d(o_{ij}, e_{ij}) = \frac{o_{ij}}{n} \log\left(\frac{o_{ij}}{e_{ij}}\right)$$

and choose to restrict the search space to avoid creating bins which are too small.

Attempts to expand these procedures to optimize over both margins have also been made. Chen et al. (2016) suggest a modified version of the MIC which is guided by the significance of a χ^2 test to choose the optimal bins. Cao et al. (2021) suggests a backwards merging algorithm on $\mathbf{b_y}$ to relax the perfect equipartition of Reshef et al. (2011). In both cases, however, an equipartition is the starting point, and so will inevitably impact the final binning.

Given that a conditional and recursive optimization procedure is already the norm, it seems natural to use recursive binary splits to generate a binning as in classification and regression trees and the tree-based binning of Rahman (2018). Recursive binning has several advantages over marginal binning. For one, recursive binning optimizes the bin edge choice over both dimensions simultaneously rather than each alone. It also produces more flexible bin arrangements than marginal methods, which can only produce bins aligned along both axes. Finally, recursive splits are adaptive to patterns in the data which are hidden in projections on either axis, suggesting that the method may detect patterns marginal bins miss.

The main problem recursive binning creates is the estimation of expected counts in a bin, which can no longer proceed by taking the product of marginal distributions. However, by first taking the ranks $\mathbf{r}(\mathbf{x})$ and $\mathbf{r}(\mathbf{y})$, the expected count can be determined by the area of the bin directly rather than a marginal product. An algorithm which implements this recursive binary binning is outlined in the following chapter.

Chapter 3

A Proposed Measure

Using the notation of the previous chapter, this chapter describes an algorithm which recursively bins the marginal ranks of a pair of variables to measure association. It begins with a sketch of the algorithm that frames the more detailed discussion of spitting rules and stop criteria in Sections 3.1.1 and 3.1.2. Following this sketch it is proven that splits only need to be considered at observed points to maximize convex score functions, in particular the commonly used ones based on the χ^2 statistic and mutual information. A maximized splitting logic is outlined and stop criteria and splitting logic which put a floor on the minimal bin size are presented.

An iterative version of the algorithm is then given in pseudo-code in Section 3.2. This code outlines the three core functions which recursively bin a data set, all of which are reflected in the implementation of the `AssocBin` package for `R` outlined in Section 3.3. An outline provides the names of the main functions in the package and how they fit together to create a modular framework for custom splitting logic.

To better understand this algorithm, Section 3.4 first provides a step-by-step demonstration of maximized splitting on uniform and perfectly dependent data. In the following subsection, the null distribution of the χ^2 statistic computed on bins split to maximize the mutual information, χ^2 statistic, or at random is explored using a simulation study. A key finding is that maximized binning of either score leads to inflated statistic values, while random splits under stop criteria and splits that ensure all bins have expected counts of 5 or more lead to a statistic conservatively approximated by the χ^2 distribution.

To evaluate the ability of recursive binning to detect different patterns of dependence, the measure is applied to several example patterns in Section 3.4.4. No great difference in the relative statistic values for each pattern compared to the null was observed between randomized and χ^2 maximizing splits, suggesting both have similar power to detect dependence. Finally, recursive binning to highlight dependence in real data is demonstrated by ordering the pairs in a real data set of historical stock values for S&P 500 constituents in Section 3.5.

3.1 Recursive rank binning to measure association

Consider a pair of variables X and Y realized in a sample of n paired observations $\mathbf{x}$ and $\mathbf{y}$, as in Section 2.1. Define the vectors of marginal ranks $\mathbf{s} = \mathbf{r}(\mathbf{x})$ and $\mathbf{t} = \mathbf{r}(\mathbf{y})$ with the convention of random tie-breaking for observations with the same rank to avoid ties. Converting to the ranks is equivalent to taking empirical CDF transforms of X and Y so that the joint distribution of $\mathbf{s}$ and $\mathbf{t}$ is the empirical copula of X and Y with a uniform joint distribution under independence. Specifically, this means the expected number of points in a region of $\{1, \dots, n\}^2$ is equal to the region's area divided by n under the null hypothesis of no dependence. This conversion allows much more flexible partitions to be considered than marginal partitioning. Here, recursive binary splits are proposed on $\mathbf{s}$ and $\mathbf{t}$ to take advantage of this flexibility.

To sketch the algorithm, first consider the objects it acts upon. From the perspective of this algorithm, a bin is a rectangular subspace of $\{1, \dots, n\}^2$ which may contain some pairs from the paired vectors $\mathbf{s}$ and $\mathbf{t}$. It is defined by its lower and upper bounds in each dimension, and has implicit features such as its area, its depth, the number of pairs it is expected to contain under independence, and the number of pairs actually observed within its bounds. The depth of a bin can be understood as the number of recursive calls required to produce it from the initial state: a single bin with bounds of 0 and n in both dimensions that contains every observation.

At each step, the algorithm is presented with bins partitioning $\{1, \dots, n\}^2$ that result from the preceding splits made by the algorithm. For each bin, the algorithm must choose whether it should be split, and if so how it should be split. Under completely unrestricted splitting, splits could be made through either of the two margins along any horizontal or vertical line within the bin boundaries. Choosing among the infinite possible splits is accomplished by finding the split optimizing a score function chosen to reflect the goal of the binning. Once a split is chosen in every bin to be split, the algorithm proceeds recursively by considering the resulting bins in the same way. Two choices made by the analyst therefore dictate the final bins produced by the algorithm: the *score* used to choose splits and the *stop criteria* which determine whether a split is made at all.

Using the settings of the following section, this algorithm has a runtime proportional to $n \log n$ if bin count or area limits are used as the stop criteria. At each step, it searches through each point in each bin, meaning all n points are considered. In the worst case, this will only halve the bin count and area at each step, and so $\log n$ splits are required. In any other case, not all n points will be considered at every depth. If small bins are created early in the procedure, the points they contain will be ignored thereafter.

3.1.1 Splitting bins

Many heuristics exist to choose splits, see Garcia et al. (2012) and Rahman (2018) for surveys, but previous bin-based measures of association have focused on the χ^2 statistic against independence and the mutual information (MI) (Reshef et al., 2011; Jiang et al., 2015; Chen et al., 2016; Heller et al., 2016; Cao et al., 2021). As they are designed to measure the discrepancy of observed distributions from expected distributions with minimal assumptions, both are natural choices to measure the dependence present in a sample.

However, these previous works choose only marginal splits on both dimensions, and so must be adapted to the recursive binning framework of Rahman (2018) by defining local versions of both, the *chi score* and *mi score*, to select the optimal split within a bin. It will be proven that, for either score, the optimal split occurs at the coordinate of a point within a bin. To distinguish between the scores and the final statistics, *chi* and *mi* will be used exclusively to refer to the scores computed to determine splitting and χ^2 and MI will be used to refer to the final statistics computed over all bins.

To define these local scores recall Equation (2.21),

$$D(\mathbf{b_x}, \mathbf{b_y}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d(o_{ij}, e_{ij}),$$

which expresses a functional measuring statistical dependence over X and Y as the sum of a function evaluated over the expected (e_{ij}) and observed (o_{ij}) number of points in the partitions created by marginal binnings $\mathbf{b_x}$ and $\mathbf{b_x}$. By adapting Equation (2.21) for the χ^2 and MI statistics to the recursive binning framework, the chi and mi scores are implied by the form of $d(o_{ij}, e_{ij})$. As the bins produced by recursive binary splits are not defined by independent marginal binnings $\mathbf{b_x}$ and $\mathbf{b_y}$, they do not have obvious i, j indices. Instead, assume the total number of bins is n_{bin} and (arbitrarily) index the bins by $i \in \{1, \ldots, n_{bin}\}$. The arguments must also be changed to $\mathbf{s}$ and $\mathbf{t}$ to reflect the absence of marginal bins.

Together, this gives the modified expression

$$D(\mathbf{s}, \mathbf{t}) = \sum_{i=1}^{n_{bin}} d(o_i, e_i), \tag{3.1}$$

where o_i is the number of observations within the i^{th} bin and e_i is the number expected assuming independence. Letting the area of the i^{th} bin be a_i, this is given by $e_i = a_i/n$. The chi score of each bin is then

$$d(o_i, e_i) = \text{chi}(o_i, e_i) = \frac{(o_i - e_i)^2}{e_i} \tag{3.2}$$

and the mi score is

$$d(o_i, e_i) = \text{mi}(o_i, e_i) = \frac{o_i}{n} \log \frac{o_i}{e_i}. \tag{3.3}$$

Either of these scores can be maximized in the proposed recursive binning algorithm to determine the split coordinate.

Maximizing scores

In more detail, denote the o_i pairs of ranks in bin i as $\{(s_{i1}, t_{i1}), (s_{i2}, t_{i2}), \ldots, (s_{io_i}, t_{io_i})\}$ and its s bounds $(l_s, u_s]$ and t bounds $(l_t, u_t]$. Bin i can be split either by a vertical line at $c_s \in (l_s, u_s)$ or a horizontal line at $c_t \in (l_t, u_t)$ resulting in two new bins with two new chi or mi scores. Denote the observed and expected values for the bin above c_s as $o_{i+}(c_s)$ and $e_{i+}(c_s)$ respectively (analogously, those above c_t as $o_{i+}(c_t)$ and $o_{i+}(c_t)$), and use the subscript $i-$ in the same way to indicate the new bin below the split. A split at c changes the total score measured by $d(\cdot, \cdot)$ for the region $(l_s, u_s] \times (l_t, u_t]$ by

$$\delta_i(c, d) = d\big(o_{i+}(c), e_{i+}(c)\big) + d\big(o_{i-}(c), e_{i-}(c)\big) - d(o_i, e_i), \tag{3.4}$$

and so the maximizing split coordinate along a given dimension is

$$c^* = \arg\max_c \delta_i(c, d) = \arg\max_c \Big[d\big(o_{i+}(c), e_{i+}(c)\big) + d\big(o_{i-}(c), e_{i-}(c)\big) \Big].$$

Though $e_{i+}(c)$ and $e_{i-}(c)$ vary continuously in the split coordinate c, both of $o_{i+}(c)$ and $o_{i-}(c)$ change only when c corresponds with the coordinate of a point contained in bin i, in other words when $c_s \in \{s_{i1}, s_{i2}, \ldots, s_{io_i}\}$ or $c_t \in \{t_{i1}, t_{i2}, \ldots, t_{io_i}\}$. This has important consequences to selecting splits for d convex in the second argument.

Proposition 1 (The score-maximizing split)**.** *If the score function $d(x, y)$ is continuous and convex in y, that is*

$$\frac{d^2}{dy^2} d(x, y) \geq 0,$$

then the split coordinate c maximizing

$$\delta_i(c, d) = d\big(o_{i+}(c), e_{i+}(c)\big) + d\big(o_{i-}(c), e_{i-}(c)\big) - d(o_i, e_i)$$

is the coordinate of one of the points within the bin.

Proof. Without loss of generality, consider a split at $c_s \in (s_{ij}, s_{i,j+1})$ between the j and $j+1$ horizontal coordinates in bin i. As the split is made between point coordinates j and $j+1$, the observed number of points above is constant at $o_{i+} = o_i - j$ and the observed number below is constant at $o_{i-} = j$ by definition. The change in score is

$$\delta_i(c_s, d) = d\big(o_i - j, e_{i+}(c)\big) + d\big(j, e_{i-}(c)\big) - d\big(o_i, e_i\big).$$

But $e_{i+}(c_s) = (u_s - c_s)(u_t - l_t)/n$ and $e_{i-}(c_s) = (c_s - l_s)(u_t - l_t)/n$ so that

$$\delta_i(c_s, d) = d\left(o_i - j, \frac{(u_s - c_s)(u_t - l_t)}{n}\right) + d\left(j, \frac{(c_s - l_s)(u_t - l_t)}{n}\right) - d(o_i, e_i).$$

If $\delta_i(c_s, d)$ is convex and continuous in c_s, then its maximum must occur at one of s_{ij} or $s_{i,j+1}$. An illustration is shown in Figure 3.1.

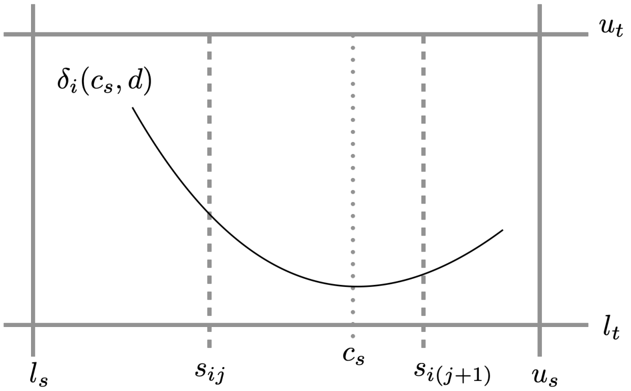


Figure 3.1: An example of a convex $\delta_i(c_s, d)$ within bin i.

Therefore, we only need to prove the convexity of $\delta_i(c_s, d)$ to prove that its maximum occurs at one of its boundaries. Consider the sign of its second derivative. The first derivative of $\delta_i(c_s, d)$ with respect to c_s is

$$\begin{aligned}
& \frac{d}{de_{i+}} d\left(o_i - j, e_{i+}\right) \frac{d}{dc_s} \frac{(u_s - c_s)(u_t - l_t)}{n} + \frac{d}{de_{i-}} d\left(j, e_{i-}\right) \frac{d}{dc_s} \frac{(c_s - l_s)(u_t - l_t)}{n} \\
= & -\frac{d}{de_{i+}} d\left(o_i - j, e_{i+}\right) \frac{u_t - l_t}{n} + \frac{d}{de_{i-}} d\left(j, e_{i-}\right) \frac{u_t - l_t}{n} \\
= & \frac{u_t - l_t}{n} \left[\frac{d}{de_{i-}} d\left(j, e_{i-}\right) - \frac{d}{de_{i+}} d\left(o_i - j, e_{i+}\right) \right].
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{d^2}{dc_s^2} \delta_i(c_s, d) &= \frac{u_t - l_t}{n} \frac{d}{dc_s} \left[\frac{d}{de_{i-}} d\left(j, e_{i-}\right) - \frac{d}{de_{i+}} d\left(o_i - j, e_{i+}\right) \right] \\
&= \frac{(u_t - l_t)^2}{n^2} \left[\frac{d^2}{de_{i-}^2} d\left(j, e_{i-}\right) + \frac{d^2}{de_{i+}^2} d\left(o_i - j, e_{i+}\right) \right],
\end{aligned}$$

which is greater than or equal to zero if $\frac{d^2}{dy^2} d(x, y) \geq 0$ for all $x \in \{0, 1, \ldots, o_i\}$.

In this case, $\delta_i(c_s, d)$ is concave up between the horizontal coordinates of the points within a bin so that any optimum within these bounds must be minimum. As these are continuous functions, this means maximum must occur at one of the boundaries of the interval $(s_{ij}, s_{i,j+1})$. As the index j was chosen arbitrarily, this same argument holds for every interval and so the global maximum must occur at one of these boundaries. These boundaries are defined by the locations of the points contained within the bin, so the maximal split must occur at the coordinate of a point within the bin. The same argument holds identically for the vertical coordinates, a fact easily seen by switching the subscripts. □

In particular, note that

$$\frac{d^2}{dy^2} \text{chi}(x, y) = \frac{d^2}{dy^2} \frac{(x - y)^2}{y} = 2\frac{x^2}{y^3}$$

and

$$\frac{d^2}{dy^2} \text{mi}(x, y) = \frac{d^2}{dy^2} \frac{x}{n} \log \frac{x}{y} = \frac{x}{ny^2}$$

are both greater than or equal to zero for all $x \geq 0$ and $y \geq 0$. This means that splits only need to be considered at the points in $\{(s_{i1}, t_{i1}), (s_{i2}, t_{i2}), \ldots, (s_{io_i}, t_{io_i})\}$ to maximize the chi and mi scores, rather than considering the continuum of potential splits in $(l_s, u_s] \times (l_t, u_t]$.

Empty bins

If splitting occurs at the coordinates of observed points within bins, the observation at the split has ambiguous bin membership. Taking the convention that these observations are counted in the bin below the split, which is consistent with the initial bin bounds $[0, n]$ in both dimensions, it is impossible for this algorithm to create empty bins below the observed points. This is despite the potential utility of empty bins when detecting association, as large regions without any observations in the rank space are a strong indication of departures from independence.[1]

To remedy this and allow empty bins to be created, a potential split coordinate below the smallest observations horizontally and vertically is added, denote these pseudo-observations as $s_{i(1)} - 1$ and $t_{i(1)} - 1$. Finally, take the convention that a point is included in the lower bin when a split occurs at one of its coordinates. Along s this gives the maximizing split coordinate

$$c_s^* = \operatorname*{arg\,max}_{c_s \in \{s_{min}-1, s_{i1}, \ldots, s_{io_i}\}} \delta_i(c_s, d)$$

and similarly

$$c_t^* = \operatorname*{arg\,max}_{c_t \in \{t_{min}-1, t_{i1}, \ldots, t_{io_i}\}} \delta_i(c_t, d).$$

Of these two maximizing splits, that giving the greater $\delta_i(c, d)$ is chosen to split bin i.

Controlling minimum bin size

Though not strictly necessary, one may want to control the minimum bin size produced by splits. This requires some balance to be struck between selecting the maximal split and keeping bins at a particular size. This is relevant, for example, if the χ^2 statistic is applied to the final bins. Supposing n_{bin} bins are created by the algorithm, the rank space $[0, n]^2$ with a presumed uniform distribution has been partitioned into n_{bin} mutually exclusive categories constrained only by the restriction that

$$\sum_{i=1}^{n_{bin}} o_i = n.$$

[1] The ability to create empty bins directly is an advantage of the recursive binning algorithm over marginal methods, which cannot do so. Empty regions marginally only reflect the marginal distribution of a variable, and converting to the ranks presents margins without any gaps for both variables.

This setting is similar to the circumstance originally considered by Pearson in his proposal of the χ^2 test, and so it is natural to assume a $\chi^2_{n_{bin}-1}$ distribution for the χ^2 statistic applied to these bins.

Care must be taken with bin size in order to apply this result, however. The $\chi^2_{n_{bin}-1}$ distribution is only asymptotically valid for the χ^2 statistic applied to this data and the fit is better the larger the expected counts in each bin. Therefore, the χ^2 distribution is typically only applied to the χ^2 statistic when the expected number of points in each partition is greater than or equal to five (Cochran, 1952). This motivates a floor on the bin size at an expected value of five.

The preceding maximization logic provides no such guarantees, and so restrictions on the candidate splits must be introduced to control the minimal bin size. Rather than change the score function, this can be accomplished by changing the $\delta_i(c, d)$ function used to evaluate splits. Consider the modified function

$$\delta_i'(c, \text{chi}, z) = I_{[z,\infty)^2}\left(\left(e_{i+}(c), e_{i-}(c)\right)^\intercal\right)\delta_i(c, \text{chi}) \tag{3.5}$$

that forces the change of score to be zero if either $e_{i+}(c) < z$ or $e_{i-}(c) < z$, where $I_A(x)$ is the indicator function of $x \in A$. This change works because $\delta_i(c, \text{chi}) \geq 0$, as

$$\begin{aligned}
\frac{(o_i - e_i)^2}{e_i} &= \frac{\left([o_{i+}(c) - e_{i+}(c)] + [o_{i-}(c) - e_{i-}(c)]\right)^2}{e_i} \\
&\leq \frac{[o_{i+}(c) - e_{i+}(c)]^2}{e_i} + \frac{[o_{i-}(c) - e_{i-}(c)]^2}{e_i} \\
&\leq \frac{[o_{i+}(c) - e_{i+}(c)]^2}{e_{i+}(c)} + \frac{[o_{i-}(c) - e_{i-}(c)]^2}{e_{i-}(c)}
\end{aligned}$$

by the triangle inequality and because $e_{i+}(c) \leq e_i$ and $e_{i-}(c) \leq e_i$. Therefore, splits producing bins that are too small will give scores less than or equal to the scores produced by all other splits. Taking $z = 5$ restricts bin splits to follow standard practice.

Controlling the bin size for $\delta_i(c, \text{mi})$ requires a different convention, as

$$\begin{aligned}
\frac{o_i}{n}\log\frac{o_i}{e_i} &= \frac{o_i}{n}\log\left[\frac{o_{i+}(c)}{o_i}\frac{o_i}{e_i} + \frac{o_{i-}(c)}{o_i}\frac{o_i}{e_i}\right] \\
&\leq \frac{o_i}{n}\left[\frac{o_{i+}(c)}{o_i}\log\frac{o_i}{e_i} + \frac{o_{i-}(c)}{o_i}\log\frac{o_i}{e_i}\right] \\
&\leq \frac{o_{i+}(c)}{n}\log\frac{o_{i-}(c)}{e_i} + \frac{o_{i+}(c)}{n}\log\frac{o_{i+}(c)}{e_i} + \frac{o_{i-}(c)}{n}\log\frac{o_{i-}(c)}{e_i} + \frac{o_{i-}(c)}{n}\log\frac{o_{i+}(c)}{e_i}
\end{aligned}$$

which may be greater than $\frac{o_{i+}(c)}{n} \log \frac{o_{i+}(c)}{e_{i+}(c)} + \frac{o_{i-}(c)}{n} \log \frac{o_{i-}(c)}{e_{i-}(c)}$. However, noting that $\text{mi}(o_i, e_i)$ and $\text{chi}(o_i, e_i)$ are both independent of the split line c, maximization of $\delta_i(c, \text{chi})$ and $\delta_i(c, \text{chi})$ depends only on the post-split scores $\text{mi}\big(o_{i+}(c), e_{i+}(c)\big) + \text{mi}\big(o_{i+}(c), e_{i+}(c)\big) \geq 0$ and $\text{chi}\big(o_{i+}(c), e_{i+}(c)\big) + \text{chi}\big(o_{i+}(c), e_{i+}(c)\big) \geq 0$. Splitting to control the minimum bin size can therefore proceed with indicators by taking the larger of

$$c_s^* = \underset{c_s \in \{s_{min}-1, s_{i1}, \ldots, s_{io_i}\}}{\arg\max} I_{[z,\infty)^2}\left((e_{i+}(c_s), e_{i-}(c_s))^\mathsf{T} \right) \left[d\big(o_{i+}(c_s), e_{i+}(c_s)\big) + d\big(o_{i+}(c_s), e_{i+}(c_s)\big) \right]$$

and

$$c_t^* = \underset{c_t \in \{t_{min}-1, t_{i1}, \ldots, t_{io_i}\}}{\arg\max} I_{[z,\infty)^2}\left((e_{i+}(c_t), e_{i-}(c_t))^\mathsf{T} \right) \left[d\big(o_{i+}(c_t), e_{i+}(c_t)\big) + d\big(o_{i+}(c_t), e_{i+}(c_t)\big) \right]$$

when $d \in \{\text{chi}, \text{mi}\}$. In the case where all splits are tied on both margins, the bin is halved on a random margin at

$$\text{ceiling}\left(\frac{l+u}{2} \right)$$

regardless of the distribution of points within the bin.[2]

3.1.2 Stop criteria

At each recursive step, a bin is split only if it fails to meet a set of stop criteria. These can include the bin area, number of points in the bin, and the depth of the bin[3]. If, for example, the criteria are a depth of 5 or $n_i < 10$, a bin with a depth of 5 or a bin with 10 or fewer points will not be split. Any bin which does not satisfy the stop criteria is split in two and the splitting algorithm is again called on both of the resulting bins.

If bin size is restricted such that $e_i \geq z \; \forall i \in \{1, \ldots, n_{bin}\}$, the corresponding stop criterion to prevent the creation of bins below this minimum size is $e_i < 2z$. When $e_i < 2z$, any split will produce at least one bin with an expected count smaller than z. Even if there is no restriction $e_i \geq z$ when splitting, it is advisable to incorporate one in the stop criteria to limit the creation of very small bins.

In all of the following examples, two stop criteria were held constant. Splitting was always stopped when $n_i = 0$ or $e_i \leq 10$. To explore the performance of the algorithm over

[2] Choosing a random split is another obvious tie-breaking convention, but can force splits which violate the minimum size restriction.

[3] Where "depth" is the number of recursive binary splits needed to create the bin from the original data.

successive splits, the depth criterion was varied from 2 to 10. For smaller sample sizes, the maximal 1024 bins implied by the depth limit of 10 was never achieved due to the stop criteria for area and empty bins.

3.2 An iterative version

Though it was conceived recursively, an iterative implementation of the algorithm outlined in Section 3.1 is detailed here. First, consider the outer wrapper function to oversee arbitrary splitting and stopping, presented in Algorithm 1. This wrapper acts on a list of `bin` objects, each with elements:

- `bounds`: named list of upper and lower bounds on s and t
- `s`: marginal ranks on x of observations in the bin
- `t`: marginal ranks on y of observations in the bin
- `e`: expected number of points in the bin
- `depth`: number of recursive splits required to create the bin.

While any bins in a list fail to satisfy the stop criteria, this wrapper calls the splitting function on those bins and combines the resulting new bins with those already stopped. The new bins are then checked against the stop criteria. To initialize, all observations are placed in a bin with bounds of $(0, n]$ in both dimensions.

Of course, the splitting and stopping logic described in Sections 3.1.1 and 3.1.2 are at the heart of this algorithm. Algorithm 2 presents pseudo-code for the splitting logic. The stopping logic is not given a pseudo-code version, as it only involves computing and checking numerous properties of each bin against a pre-determined set of thresholds. Section 3.3 discusses a particular implementation of this in the R programming language.[4]

Finally, Algorithm 3 gives the pseudocode for a chi scoring function with limited minimal bin size. As written, this function would be provided as the `score` in the `maxScoreSplit` function of Algorithm 2. Note that this algorithm has been written for the chi score but provides a framework for any scoring function. The specifics of line 14 can

[4] In this implementation, there is the ability to specify a custom initial split function with the argument `init`. Choosing `init` to randomly halve a bin along one margin coincides with the splitting logic described earlier, as every split produces a score of zero for the initial uniform data.

Algorithm 1 Iterative binning wrapper

```
Input
    x - vector of observed values in $x$
    y - vector of observed values in $y$, paired with $x$
    stopper - function which tests a list of bins against the stop criteria
    splitter - function which performs a binary split on a bin
    init - (possibly) different splitting function applied to the initial bin with all observations

function BINNER(x, y, stopper, splitter, init)
    $n \longleftarrow$ length(x)                                  ▷ Compute preliminaries
    s $\longleftarrow$ rank(x)
    t $\longleftarrow$ rank(y)
    iniBin $\longleftarrow$ makeBin(s = s, t = t, bounds = list(s = $(0, n)$, t = $(0, n)$), e = $n$, depth = 0)   ▷ Construct initial bin
    binList $\longleftarrow$ init(iniBin)                           ▷ Initialize bin list
    stopStat $\longleftarrow$ stopper(binList)                      ▷ Initial stop check
    while any stopStat are FALSE do              ▷ Continue as long as bins can be split
        oldBins $\longleftarrow$ binList[stopStat]                  ▷ Stopped bins
        oldStop $\longleftarrow$ stopStat[stopStat]                 ▷ All TRUE
        newBins $\longleftarrow$ {}                  ▷ Variable to store splitting results
        for $bin$ in binList[!stopStat] do
            append splitter($bin$) to newBins                   ▷ Add split results
        end for
        newStop $\longleftarrow$ stopper(newBins)    ▷ Check stop criteria on new bins
        binList $\longleftarrow$ append newBins to oldBins
        stopStat $\longleftarrow$ append newStop to oldStop
    end while
    return binList
end function
```

Algorithm 2 Score maximizing splitter

```
Input
    bin - the bin object to be split with elements s, t, bounds, e, and depth
    scorer - function accepting a vector of coordinates in increasing order and the expected number of points in bin and returning a vector of scores corresponding to splits at each internal coordinate

function maxScoreSplit(bin, scorer)
    get s, t, s.bounds, t.bounds, e, depth from bin
    sort s, t in increasing order to give $sSrt, tSrt$
    $\mathbf{c}_s \longleftarrow$ append ( $sSrt$[1]-1, $sSrt$ )
    $\mathbf{c}_t \longleftarrow$ append ( $tSrt$[1]-1, $tSrt$ )        ▷ Add a split coordinate below all points
    $\mathbf{d}_s \longleftarrow$ scorer( append ( s.bounds[1], $\mathbf{c}_s$, s.bounds[2] ), e )
    $\mathbf{d}_t \longleftarrow$ scorer( append ( t.bounds[1], $\mathbf{c}_t$, t.bounds[2] ), e )
    $s_{Max}$, $t_{Max} \longleftarrow$ the indices of the maxima of $\mathbf{d}_s$, $\mathbf{d}_t$
    if all $\mathbf{d}_s = \mathbf{d}_s[s_{Max}]$ AND all $\mathbf{d}_t = \mathbf{d}_t[t_{Max}]$ then
        if $\mathbf{d}_s[s_{Max}] > \mathbf{d}_t[t_{Max}]$ then
            split bin on $s$ at ceiling(mean(s.bounds))
        else if $\mathbf{d}_s[s_{Max}] < \mathbf{d}_t[t_{Max}]$ then
            split bin on $t$ at ceiling(mean(t.bounds))
        else
            split bin randomly on $s$ or $t$ at ceiling(mean(s.bounds)) or ceiling(mean(t.bounds))
        end if
    else if $\mathbf{d}_s[s_{Max}] \geq \mathbf{d}_t[t_{Max}]$ then        ▷ Ties go to $s$
        split bin on $s$ at $\mathbf{c}_s[s_{Max}]$
    else
        split bin on $t$ at $\mathbf{c}_t[t_{Max}]$
    end if
    return upper and lower split of bin
end function
```

Algorithm 3 Marginal χ^2 scores

```
Input
    $\mathbf{x}$ - a numeric vector of potential split coordinates in increasing order
    $e$ - a numeric value giving the expected number points in a bin
    $e_{min}$ - the minimum expected number allowed for a split

function chiScoring($\mathbf{x}$, $e$, $e_{min}$)
    $n \longleftarrow$ length($\mathbf{x}$)
    $cumulative \longleftarrow \mathbf{x}[2] - \mathbf{x}[1]$              ▷ Initialize cumulative length
    $density \longleftarrow e/(\max(\mathbf{x}) - \mathbf{x}[1])$            ▷ Density under uniformity
    scores $\longleftarrow \{\}$                                       ▷ Initialize storage
    for $i = 2$ to $n - 1$ do
        $e_i \longleftarrow cumulative * density$                      ▷ Expected count below $i$
        $o_i \longleftarrow i - 2$                ▷ Observed count ignores bounds, pseudo-point
        if $e_i \geq e_{min}$ AND $e - e_i \geq e_{min}$ then
            append $\frac{(o_i-e_i)^2}{e_i} + \frac{(n-3-o_i-e+e_i)^2}{e-e_i}$ to scores      ▷ chi score of candidate split
        else
            append 0 to scores
        end if
        $cumulative \longleftarrow cumulative + \mathbf{x}[i+1] - \mathbf{x}[i]$                ▷ Update length
    end for
    return $scores$
end function
```

be replaced with any objective function, for example the mi score of Equation (3.3) with $\frac{o_i}{n-3} \log \frac{o_i}{e_i} + \frac{n-3-o_i}{n-3} \log \frac{n-3-o}{e-e_i}$ to maximize the marginal Kullback-Liebler divergence from uniformity.

This framework also supports random binning. Rather than computing a function that compares the observed distribution to a uniform one about a split, the scores can be replaced with independent and identically distributed $U[0,1]$ realizations. The coordinate of the maximum score value will then be uniformly distributed along each potential split coordinate and each margin, creating a fully random recursive binning procedure. Additionally, as these random realizations are greater than or equal to zero, adopting the indicator multiplication of Equation (3.5) controls the bin size under this form of random splitting.

3.3 The `AssocBin` package

Algorithms 1, 2, and 3 are implemented in the `R` package `AssocBin` for the mi score, chi score, and random splitting score. The core functions are:

`makeCriteria`: a function which captures its arguments and appends them into a single logical expression

`stopper`: a function which accepts a list of bins and a logical expression and evaluates the expression within each bin using `R`'s lexical scoping

`binner`: the wrapper function described in Algorithm 1 which accepts integer vectors `x` and `y`; a `stopper` function which accepts a list of bins and returns a logical vector; a `splitter` function which accepts a single bin and returns a pair of bins partitioning the input bin; an `init` function which splits the initial bin containing all points; and (optionally) additional arguments to pass to internal function calls

`chiScores`, `miScores`, `randScores`: functions which implement Algorithm 3 with line 14 replaced by the corresponding score

`maxScoreSplit`: the splitting function described in Algorithm 2 which accepts `bin` and `scorer` functions in addition to `ties` and `pickMax` which allow for custom tie and maximum choice handling

`splitX`, `splitY`: functions which accept a `bin` to be split, a numeric `bd` giving the coordinate of the split, and the indices of values `above` and `below` `bd` and return two bins resulting from a split of `bin` at `bd` along the corresponding margin

halfCutTie: the tie-breaking logic described in Algorithm 2

By writing `binner` with fully modular components, the score function, splitting logic, stop criteria, and ties can be modified to suit the preference of a user. As these are all imagined as single argument functions within `binner`, certain functions need to be defined in closures before use. To use `stopper`, for example, the stop criteria returned by `makeCriteria` must be passed as an argument to `stopper` within another function which returns a single-argument function. Similarly, `maxScoreSplit` must have its scoring function and minimum expected count set in a closure that then accepts a single argument. Though this requires some extra set up, it limits the arguments of `binner` and forces the user to consider these choices intentionally in advance. The demos and vignette included in the package provide examples.

Additional helper functions visualize and summarize the results. The `binChi` function computes the χ^2 statistic over a list of bins returned by binner and the `plotBinning` function plots a bin list and scatterplot with optional fill. Two additional functions, `depthFill` and `residualFill`, create gradient fills to communicate depth or residual magnitude based on colour range and residual function arguments. Such visualizations not only give insight into what region of the data the algorithm deems most important, but also provide a summary of the data. Rather than a scatterplot with potentially thousands of points, these visualizations display a handful of coloured regions which can be read at a glance. The full package can be found on the author's GitHub and on CRAN (Salahub, 2023a).

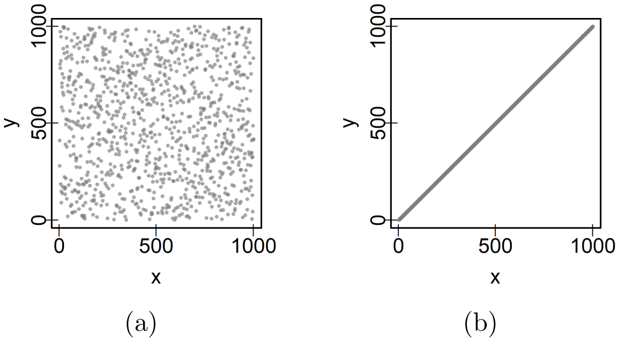
3.4 Using the algorithm

Some results gained from the exploration of this algorithm in practice are presented here. First, a series of visualizations of the algorithm on independent data and strongly associated data are given to demonstrate how it works step-by-step. Then, as both Heller et al. (2016) and Reshef et al. (2011) utilize simulated independent X and Y to generate the p-values of their methods, the null distribution of the recursive binning measure under different splitting rules is explored. This is followed by the application of the method to simulated data patterns from Newton (2009) and Liu et al. (2018). Finally, a real data set is examined.

3.4.1 Simple examples

To illustrate the behaviour of the maximum splitting algorithm from Section 3.2, consider applying it to two extremes: random independent data and data in perfect agreement.

Scatterplots of both of these simulated data sets with $n = 1,000$ are shown in Figure 3.2.

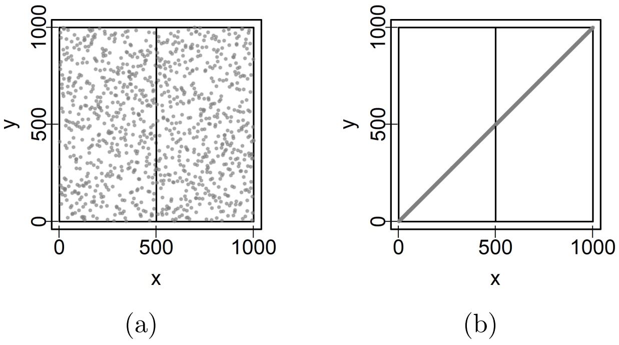


(a)

(b)

Figure 3.2: The (a) simulated random data and (b) perfect rank agreement data used to illustrate the flow of the algorithm.

As the ranks of both data sets have no gaps, the initial step of the algorithm will not find a marginal maximal split. Indeed, for any split at a point the number of points on either side will match expectation exactly. Therefore, the algorithm begins by splitting the initial bin in half by adding a vertical edge at 500 in both. This gives the bins seen in Figure 3.3.

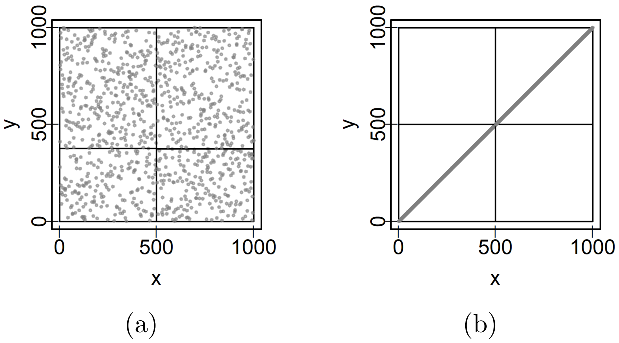


(a)

(b)

Figure 3.3: The (a) simulated random data and (b) perfect rank agreement data with the first split indicated.

Once so halved, the ranks are no longer necessarily uniform within the bins. Marginal gaps are introduced in both due to the split. Therefore, the next split is not made at the halfway point, but is instead made to optimize the deviation of counts from uniformity as measured by the score function. While identifying the location of these splits on either side in the random uniform data is difficult, the expected split in the line data would be another halving of the bins, as this produces empty corner bins which should each contain a quarter of the observations. Indeed, this is the result seen in Figure 3.4.



(a)

(b)

Figure 3.4: The (a) simulated random data and (b) perfect rank agreement data with the first two splits indicated, dividing both data sets into four bins.

Such a step-by-step demonstration could be continued, to the tedium of the reader, but an easier visualization can summarize such stepwise inspection. For each bin, the implementation described in Section 3.3 keeps track of the bin depth as well as its other features. More information about the algorithm's path can be gleaned by simply running the algorithm and shading the bins according to their depth. Setting a maximum depth of six and shading in this way produces Figure 3.5.

There are stark differences between the depth patterns for these two data sets, as might be expected. While the uniform random data continues splitting every bin somewhat haphazardly with the exception of small early splits that meet the stop criteria (the expected number of observations being 10 or fewer or no observations within the bin), the line data causes a very particular pattern of depths to emerge. It seems to chase the linear pattern by introducing many more splits along its length. In this way it assigns a greater density of bins to the regions with a greater density of points.

By chasing these regions, the algorithm produces a striking pattern in the residuals of its final bins. This pattern occurs in, for example, the Pearson residuals of bins with

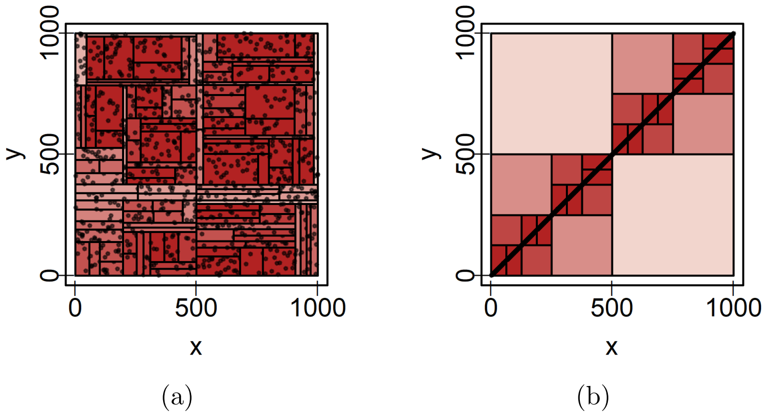


Figure 3.5: The (a) simulated random data and (b) perfect rank agreement data split to a maximum depth of six, with bins shaded darker according to their depth. The algorithm splits to the maximum depth along the line and not elsewhere while the in the random data continues splitting in every location.

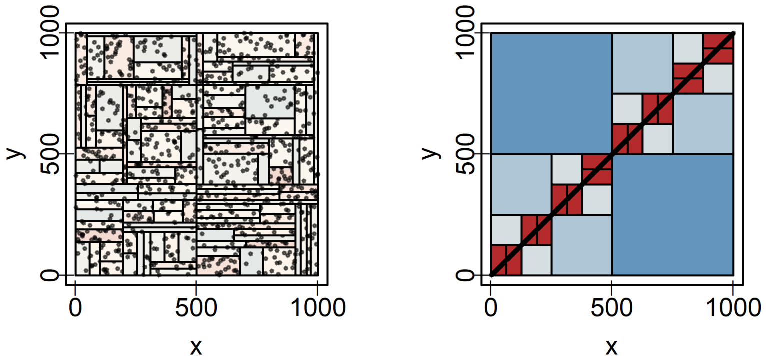
expected counts e_i and observed counts o_i,

$$r_i \text{sign}(o_i - e_i)\sqrt{\frac{(o_i - e_i)^2}{e_i}}.$$

Large positive residuals occur all along the line and large negative ones occur away from it. Figure 3.6 plots these residuals using hue to convey their sign, with the convention that blue represents negative values and red represents positive ones, and saturation to convey their magnitude, with darker saturation indicating a larger magnitude.

While bins in the random data do not display pronounced shading, indicating small residuals, the line pattern has large positive residuals all along its length and large negative residuals elsewhere. In particular, note the deep blue shading in upper left and bottom right quadrants. Even without the points, the structure of the data is easily discerned from the shading of these bins alone.

In this example, taking a sum of the squared Pearson residuals to get the χ^2 statistic gives values of 87.8 for the random data and 7000 for the line data. Of course, as the bins produced here do not follow a regular grid pattern and are generated by maximizing a score analogous to the χ^2, the distribution of the statistic warrants an investigation under independence. As these bins are produced using optimization at each step, it is not clear



(a)

(b)

Figure 3.6: (a) Simulated random data and (b) simulated data with perfect rank agreement split to a maximum depth of six, with bins shaded according to their Pearson residuals. Negative residuals are shaded blue while positive residuals are shaded red, with darker shading indicating a larger residual magnitude. Large positive residuals occur along the line and large negative ones occur elsewhere.

that n_{bin} bins would have a χ^2 distribution with $n_{bin} - 1$ degrees of freedom as is usually the case, even with the controlled minimum bin size.

3.4.2 The null distribution

To explore the null distribution of the χ^2 statistic under recursive binning, a simulation study generating many null replicates is undertaken here. Three different splitting rules are considered: random splits, chi score maximizing splits, and mi score maximizing splits. All will lead to different final bins when applied to the same data and therefore different statistic values. To satisfy the rule of thumb for the $\chi^2_{n_{bin}-1}$ distribution a better chance of fitting, all are restricted so that bins with expected counts less than 5 are not created.

10,000 independent bivariate data sets are generated $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{10,000}$ by randomly and independently shuffling the vector of ranks $\{1, 2, \ldots, 10000\}$ on each margin. Every one of these $\mathbf{x}, \mathbf{y}$ pairs is recursively binned to every maximum depth between two and ten using random binning, binning maximizing the chi score, and binning maximizing the mi score at each step. Splitting is stopped when the expected number of observations within a bin falls below ten or if a bin contains no observations. For each binning, the χ^2 statistic is computed for the final bins based on the observed and expected counts, and this statistic is recorded along with the number of final bins.

The impact of depth

To start, consider a plot of the χ^2 statistic against the number of bins for these simulated data when bin splits are selected randomly and uniformly among observed point coordinates within a parent bin. Figure 3.7 displays this plot with both margins $\log_{10}$ transformed to make the horizontal width of clusters for small depths as wide as those for large depths. A dashed line is plotted on top of the points at the 99% critical value of $\chi^2_{n_{bin}-1}$ for reference.

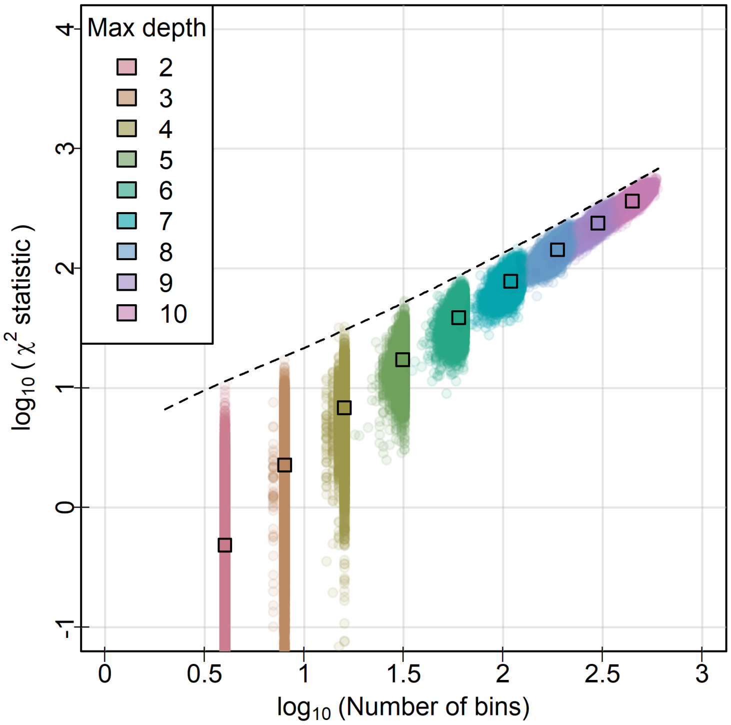


Figure 3.7: The χ^2 statistic plotted by the final number of bins for 10,000 simulated independent data sets for each of 9 depth settings with random splits. The dashed line displays the 99% critical value of the χ^2 distribution with one degree of freedom less than the number of bins, which is conservative for all depths.

The expected and minimum count stop criteria together cause the number of bins for a given maximum depth stop criterion to vary on each realization. If every bin were split at every step, a maximum depth of d would always produce 2^d bins but small bins produced by chance early in splitting are not further split. Nonetheless, a larger maximum depth will produce more bins on average than a smaller maximum depth. This produces the clusters of bin counts at each depth in Figure 3.7.

A more interesting result comes from a comparison between the 0.99 quantile of the $\chi^2_{n_{bin}-1}$ distribution and statistic values for all depths. Given that 10,000 null cases are generated for each depth, roughly 100 points are expected above this critical value if the $\chi^2_{n_{bin}-1}$ distribution is the null distribution. Instead, far fewer points are observed above the line, suggesting the $\chi^2_{n_{bin}-1}$ 99% critical value is conservative for χ^2 under random splitting. That is, choosing quantiles from this distribution to test for dependence at level 0.01 implies a test with a level of at most 0.01 for the true null distribution. Indeed, quantile regression performed later suggests the $\chi^2_{n_{bin}}$ distribution is a conservative approximation for the null distribution under random splitting.

The same result is not expected for splits chosen to maximize the chi or mi score. The χ^2 statistic for the simulated data split to optimize the chi score is shown plotted against the final number of bins in Figure 3.8 and the same plot for splits optimizing the mi score is shown in 3.9. Points are, again, coloured according to the maximal depth allowed, and the mean statistic values and number of bins for each of these depth settings are denoted with corresponding coloured squares. Additionally, the 99% critical value for the χ^2 distribution with $n_{bin} - 1$ degrees of freedom is plotted against n_{bin}.

Two changes occur in Figure 3.8 compared to random binning. First, the statistic values tend to be larger than the null case for the same value of n_{bin}. Indeed, the majority of score-maximizing binned statistics are above the $\chi^2_{n_{bin}-1}$ 99% critical line when the maximal depth is only 4, all are above the line when the depth is 6, and the gap between the critical value curve and the true distribution grows larger as the number of bins increases.

Second, the number of bins no longer displays distinct clusters by maximal depth. Whereas Figure 3.7 has clusters along the horizontal margin corresponding to each maximal depth, the maximum binning bin counts are not strongly grouped. This may be a result of the maximized method selecting smaller bins on average than random binning, resulting in fewer bins due to the minimum size and expected count restrictions that stop splitting of these small bins early. This dynamic ‘smears’ the number of bins for each maximum depth, leading to large overlap between groups by maximum depth.

Figure 3.9 shows a similar pattern for the mutual information statistic. The χ^2 statistic value rapidly increases in the number of bins, resulting in a distribution which is well above

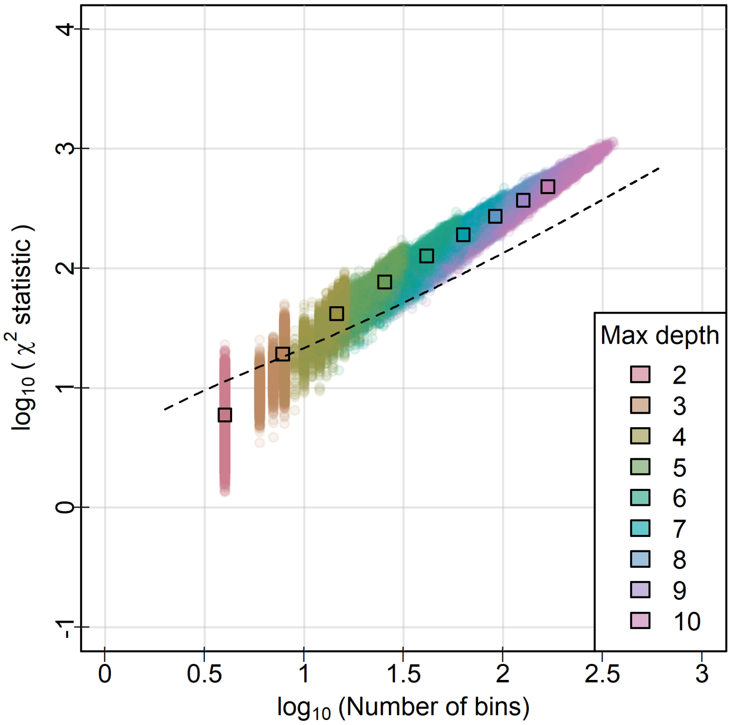


Figure 3.8: A comparison of the number of final bins and χ^2 statistic for 10,000 simulated independent data sets split to maximize the chi score over each of 9 depth settings. The dashed line displays the 99% critical value of the χ^2 distribution with one degree of freedom less than the number of bins. For all depths, many more points lie above this critical value than would be expected if the data followed a $\chi^2_{n_{bin}-1}$ distribution.

the $\chi^2_{n_{bin}-1}$ 0.99 quantile. Indeed, both seem to produce similar null distributions for the χ^2 statistic when applied to the same data.

To better compare the $\chi^2_{n_{bin}-1}$ quantiles to the null distributions across all numbers of bins, quantile regression of the statistic value on the bin depth was carried out using simulated data and the `quantreg` package in R (Koenker, 2023). Specifically, regression of the 0.95, 0.99, and 0.999 quantiles of the χ^2 statistic was carried out for binning under all three splitting methods. The results are shown in Figure 3.10. Just as indicated by the earlier plots, maximized splitting with either score leads to *inflated* statistic values, that is a statistic which is stochastically greater than a $\chi^2_{n_{bin}-1}$ random variable. Both maximized

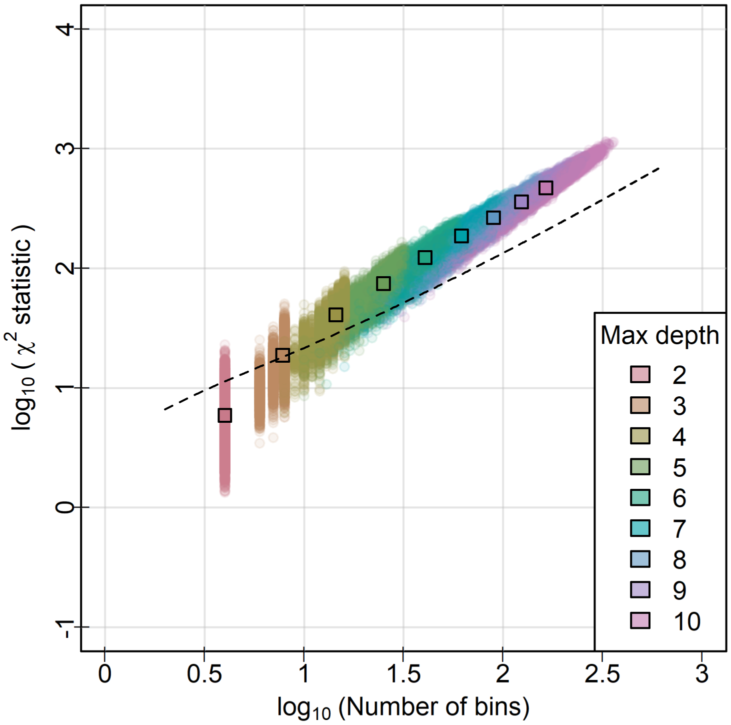


Figure 3.9: A plot of the χ^2 statistic by the final number of bins for a binning which maximized the mi score at each step. The plot looks almost identical to that generated by splitting the chi score.

methods produce almost identical upper quantiles. In contrast, random splitting produces a χ^2 statistic which is stochastically less than a $\chi^2_{n_{bin}-1}$ random variable, suggesting the $\chi^2_{n_{bin}}$ distribution could be used to generate conservative p-values (at least in the tails).

These lines not only make the difference in the distributions under the different splitting methods clear, but also could be used in practice to determine rejection or acceptance of the null hypothesis of independent data. Of course, the quantiles depend on the number of observations, so this plot is only demonstrative unless the sample size is 10,000.

In short, both depth and splitting method are highly relevant to the distribution of statistics computed on the final bins. The naive $\chi^2_{n_{bin}-1}$ distribution on the bins seems to provide a conservative distributional approximation for random splitting but not when splitting to maximize a score function. Instead, large simulation studies as performed here

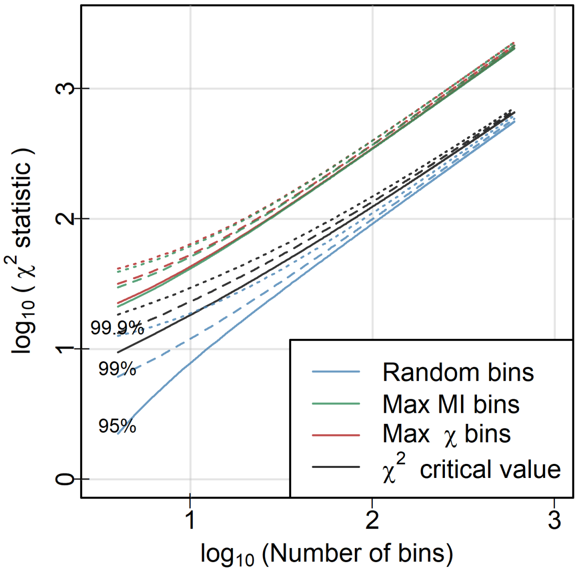


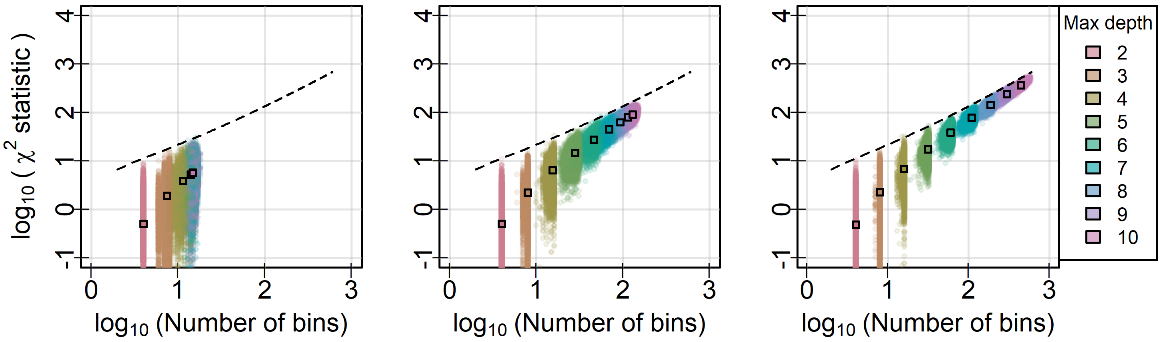
Figure 3.10: Fit quantile regression lines for several quantiles for all different splitting methods. Both maximized splitting methods lead to similar larger quantiles that grow faster with depth and are further apart, though the chi score leads to slightly larger χ^2 statistic values.

must be used as a reference to determine critical values and empirical p-values. Results could be smoothed or interpolated to cover gaps in the data due to certain bin counts occurring more and less frequently by chance.

3.4.3 Depth and sample size

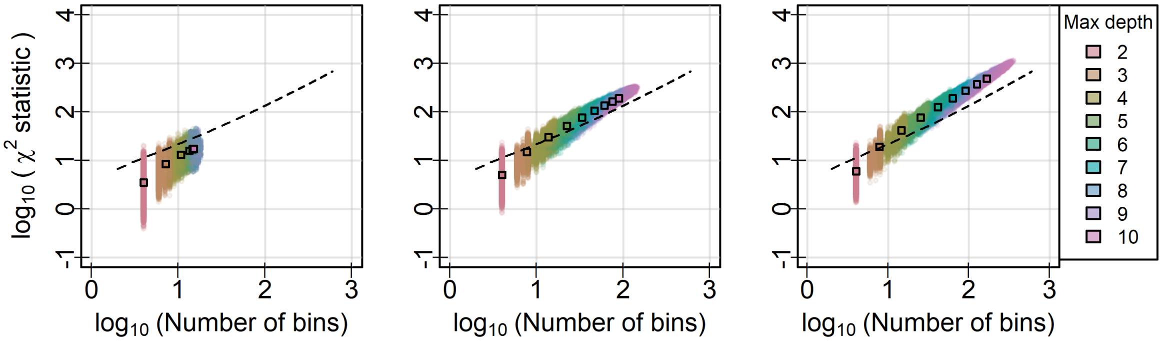
To investigate other sample sizes, this study is expanded to 10,000 samples of size $n = 100$ and $n = 1{,}000$, all split to a maximum depth of 10. The stop criteria are the same as the preceding investigation. Plots of the resulting χ^2 statistic and number of bins are compared to the plot for $n = 10{,}000$ in Figure 3.11 for the case of random splitting and in Figure 3.12 for the case of spitting which maximizes the chi score at each step. Splits maximizing the mi score were not investigated due to the very similar null distribution to those maximizing the chi score.

In both cases, the primary impact of increasing the sample size is to increase the maximum number of bins produced by the algorithm. For smaller sample sizes, the minimum



(a) (b) (c)

Figure 3.11: χ^2 statistic and number of bins for 10,000 samples of simulated random data split by random binning to a maximum depth of ten with the minimum expected bin count restricted to 5 for (a) $n = 100$, (b) $n = 1{,}000$, and (c) $n = 10{,}000$. As the sample size increases, the maximum number of bins and separation between the depths increases.



(a) (b) (c)

Figure 3.12: χ^2 statistic and number of bins for 10,000 samples of random data split maximizing the chi score with a minimum expected bin count of 5 to a maximum depth of ten for (a) $n = 100$, (b) $n = 1{,}000$, and (c) $n = 10{,}000$. As before, the sample size limits the maximum number of bins produced by the algorithm.

bin size stop criteria and bin split restrictions end splitting before many bins are created. This could have been anticipated. With a sample size of 100, for example, splitting the data into 10 bins implies an expected count of 10 per bin, at which point the stop criteria would cease further splitting. Irregular bin shapes complicate this, of course, but the basic idea stands. A smaller sample implies smaller expectations to compare to the constant stop criteria.

Importantly, this does not seem to impact the conservative approximation of the χ^2 statistic under random binning by the $\chi^2_{n_{bin}-1}$ distribution. For all three sample sizes, the observed statistic values have similar location and spread given the number of bins. As the sample size increases, the distribution merely spreads along the curve dictated by the $\chi^2_{n_{bin}-1}$ distribution.

The same is not true for splits maximizing the χ^2 score. As the sample size increases, the whole distribution of the χ^2 statistic moves upward for a given number of bins. This is particularly obvious if the mean points are compared to the line providing the 99% quantile of the $\chi^2_{n_{bin}-1}$ distribution. In Figure 3.12(a), none of the mean points lie above this line for $n_{bin} > 10$ while in Figure 3.12(c) all of them lie well above this line.

These different behaviours are probably a result of the way the maximizing algorithm chases local patterns, as demonstrated on the perfect line data. Random binning splits agnostically, but chi maximizing binning searches through all splits for the one leading to the largest χ^2 statistic. Large χ^2 values therefore occur by chance for random binning but are actively sought by maximized binning. With a larger sample, a greater variety of local patterns will be included which can inflate the χ^2 statistic. The maximized algorithm actively searches for such patterns, stochastically increasing the resulting distribution.

3.4.4 Simulated data patterns

Besides the null distribution, an important aspect of recursive binning to measure association is the ability of the algorithm to detect non-null patterns in the data. Newton (2009) provides supplemental code to generate several interesting data patterns, which have since been used by Liu et al. (2018) and Heller et al. (2016) to test their methods and also appear on the Wikipedia article for Pearson's correlation. This code is adapted and used here to repeatedly generate observations following each pattern as a test of recursive binning. Samples of 1,000 observations from each pattern are shown in Figure 3.13, and the marginal ranks of these samples are shown in Figure 3.14. Both figures have the margins dropped for brevity.

All but the final pattern contain non-linear dependence between the vertical and horizontal variables, and still display strong patterns when these variables are converted to ranks. The final sample, in contrast, consists of four clusters of equal size which can be constructed by the product of independent bimodal marginal densities. Consequently, in the ranks the sample appears uniform. The final pattern therefore acts as a control to

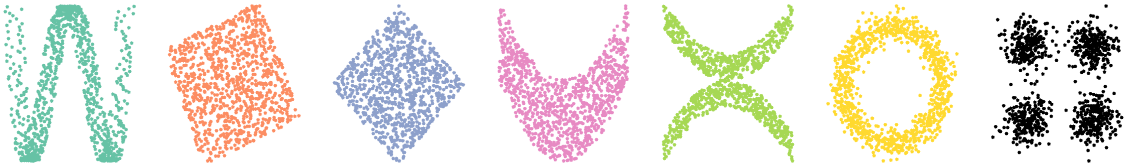


Figure 3.13: Point patterns used in Newton (2009) and Liu et al. (2018) to demonstrate different dependence structures. In the last pattern, the data are independent despite the obvious two-dimensional structure.

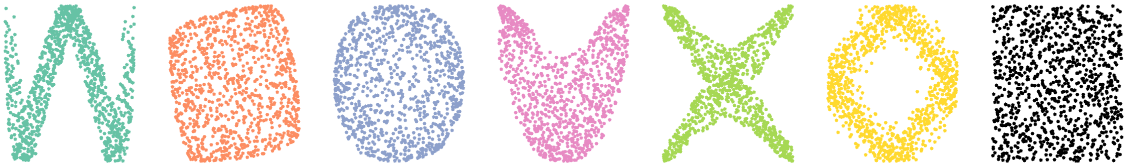


Figure 3.14: The patterns of Figure 3.13 converted to marginal ranks. Refer to these patterns as the wave, rotated square, circle, valley, cross, ring, and uniform noise respectively.

ensure that structure without dependence is not detected by recursive binning applied to the ranks.

For each pattern, both the maximized binning under the chi score and random binning algorithms were applied with maximum depths ranging from 1 to 10 and the same stop criteria as previously (an expected count ≤ 10 or an empty bin). This gives a sequence of n_{bin}, χ^2 coordinates for each pattern as the depth restriction is increased, just as in Figures 3.7 and 3.8. Rather than visualize these as points, the progression of bins and statistics with increasing maximum depth can be emphasized by joining the points for sucessive depth restrictions to give paths for each pattern in the n_{bin}, χ^2 space. The same can be done for the 10,000 simulated null samples of 1,000 points from Section 3.4.2 to give 10,000 null lines for comparison. Figure 3.15 displays paths coloured by pattern to match Figure 3.14 alongside the null paths coloured in gray. A dashed line at the 0.95 quantile of the $\chi^2_{n_{bin}}$ distribution is added to simulate a rejection boundary in Figure 3.15(b).

Figure 3.15(a) shows smooth curves of the χ^2 statistic increasing the number of bins as the maximum depth is increased while Figure 3.15(b) shows erratic paths which occasionally decrease in the statistic value as depth increases. Holding this major difference aside for a moment, all the curves for patterns with dependence lie well above the null paths,

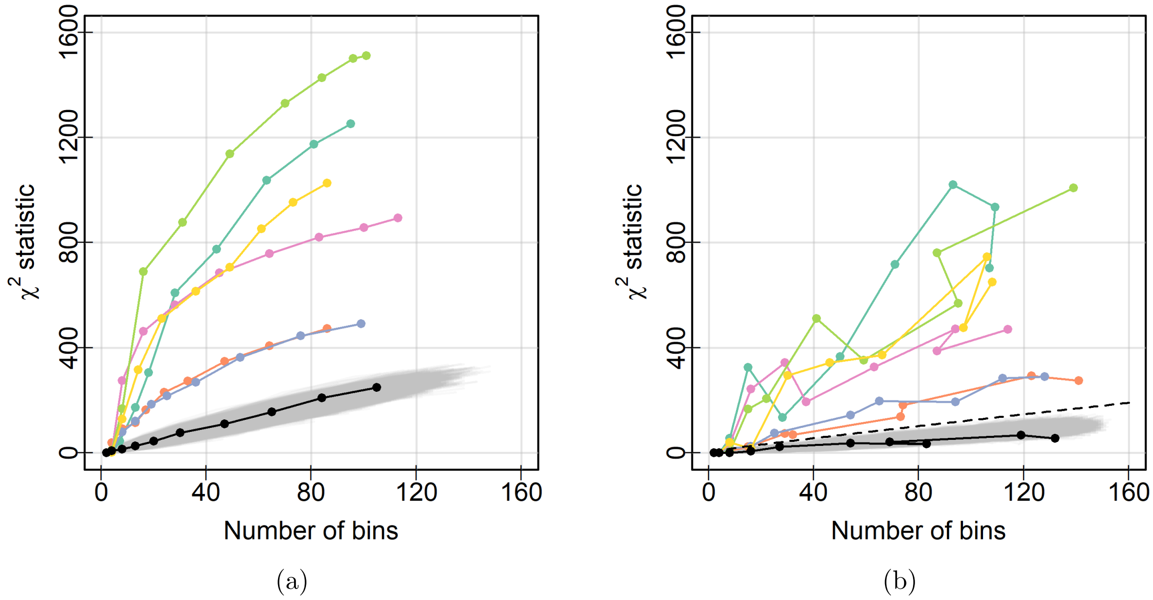


Figure 3.15: Paths for every pattern in n_{bin}, χ^2 statistic space across depth restrictions for (a) maximized chi score splitting and (b) random splitting. While the maximized splitting is deterministic for a sample, leading to smooth curves, randomized binning leads to rough and erratic paths. Despite this, both display roughly the same ordering of the patterns by colour, and all patterns have paths far above any of the 10,000 simulated null cases.

while the uniform pattern sits within the null curves for both splitting methods. This suggests empirical p-values less than 0.0001 for every pattern when the depth restriction is greater than 3. Additionally, the order of the curves is similar between the two, with the wave and the cross giving the largest χ^2 statistic values for a given number of bins and the rotated square and circle giving the smallest statistic values.

Returning to the difference in path smoothness, the erratic nature of the random binning and smoothness of the maximized binning are natural consequences of the different split methods. The maximized splitting algorithm chooses the maximum split at each step and so behaves deterministically for a given sample. In contrast, the random splitting algorithm proceeds non-deterministically and will generate different bin counts and statistic values for a given depth restriction every time it is run, even for constant data. To better see the difference this makes for the separation of the patterns and to evaluate the performance of both based on more than a single exemplar, 100 samples of 1,000 points from each pattern are generated independently and subsequently split either by chi-maximizing or random binning. The resulting paths are displayed in Figure 3.16, with the median path plotted using a thicker grey-accented line.

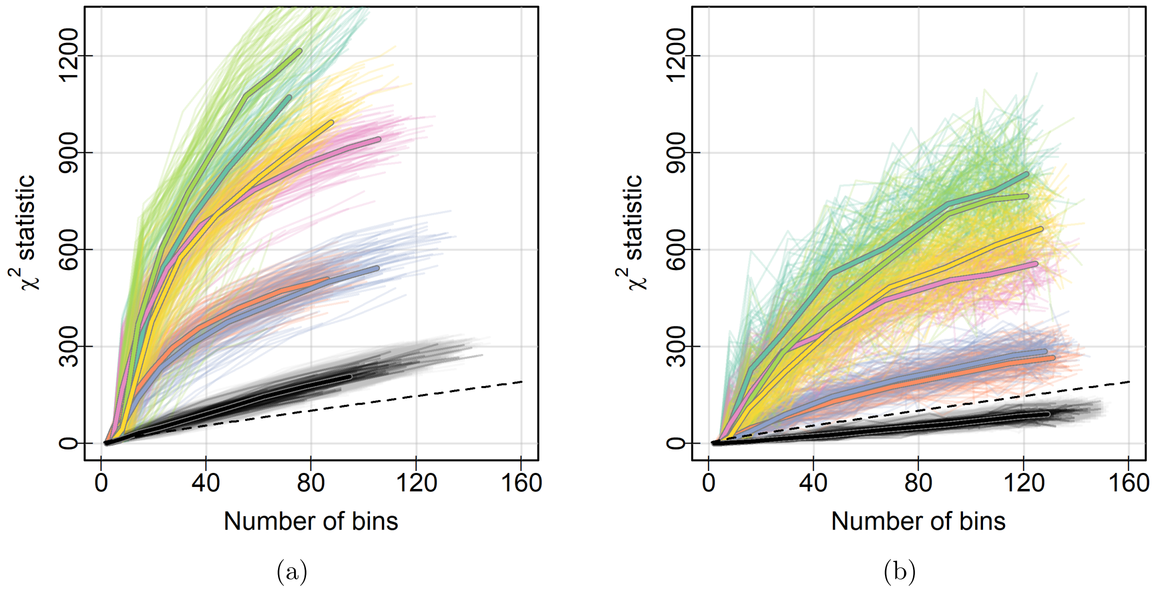


Figure 3.16: Paths of (a) maximized and (b) random binning applied to 100 independent realizations of the seven simulated data patterns compared to the null paths with median paths plotted with thicker lines accented by grey. Both spitting regimes create clear separation between the null paths and paths of patterns with dependence and both display the same ordering of patterns. The random splitting, however, creates more erratic and variable paths.

The clouds of paths show roughly the same ordering under both random binning and chi-maximized binning. Both place the wave (in dark green) and the cross (in light green) above the others, followed closely by the valley (in pink) and the ring (in yellow). Halfway between these these patterns (all with locally linear sections) and the null patterns sit the rotated square (in orange) and circle (in blue). Though these latter two result in smaller statistic values than the others, they remain easily distinguished from the null paths in both Figures 3.16(a) and 3.16(b). Indeed, under random splitting the $\chi^2_{n_{bin}-1}$ 0.95 quantile plotted with a dashed line neatly separates the null paths from these two patterns and consequently all patterns.

The close agreement of their ordering suggests that randomized binning is as powerful as maximized binning at detecting these patterns. The clear separation of the null curves from the curves of all patterns over all repetitions implies empirical p-values < 0.0001 for all patterns when the algorithm is allowed to reach its maximal depth using only bin size stop criteria. Any of these patterns would therefore be regarded as significant by both regimes for commonly used rejection levels. Moreover, as the order of curves by pattern is the same for both, the greatest difference between the maximized binning and random binning seems to be the increased the statistic value alone. If a rejection level is chosen on the χ^2 statistic proportional to the null quantile, we expect that both regimes would show the same tendency for acceptance or rejection of every pattern. Given that random binning requires less computation and has a conservative p-value provided by the $\chi^2_{n_{bin}-1}$ distribution, the similar ordering of these curves provides a strong argument for the use of random recursive binning of the ranks to detect and quantify dependence between variables.

Progression of bins in the maximizing algorithm

The median lines are somewhat more interesting, as they cross each other frequently. This suggests the ordering of these different patterns, and potentially the power of the algorithm under each splitting method to detect them, is dependent on the maximum depth. Under chi-maximized binning, the median line for the rotated square is above all others for very small numbers of bins, indicating there is a pattern in the data which can be captured quickly with few splits. In contrast, the median line for the ring pattern starts below most others for both splitting methods, indicating that a certain number of bins are required before the pattern is detected. It seems each pattern has a natural *resolution*: a number of bins required to identify the dependence.

Directly related are the different slopes of these curves for each pattern. Patterns which generate large residuals within a few splits will grow rapidly at first, and then may slow if the following residuals are relatively small. The median path of the valley pattern, for example, grows quicker than most other patterns in the number of bins for the first few depths before its rate of growth slows. In contrast, the χ^2 statistic of the ring pattern is unexceptional before a certain depth, at which point the statistic values jump abruptly. More insight into both the slope of these lines and the resolution of each pattern can be gleaned by plotting the bins at each depth for each pattern shaded by residual as in Figure 3.17.

As before, the hue of the shading is determined by the sign of the bin's Pearson residual, blue for negative and red for positive, and the saturation is determined by its magnitude. To ensure fair comparison of the residuals, all saturations are determined relative to the

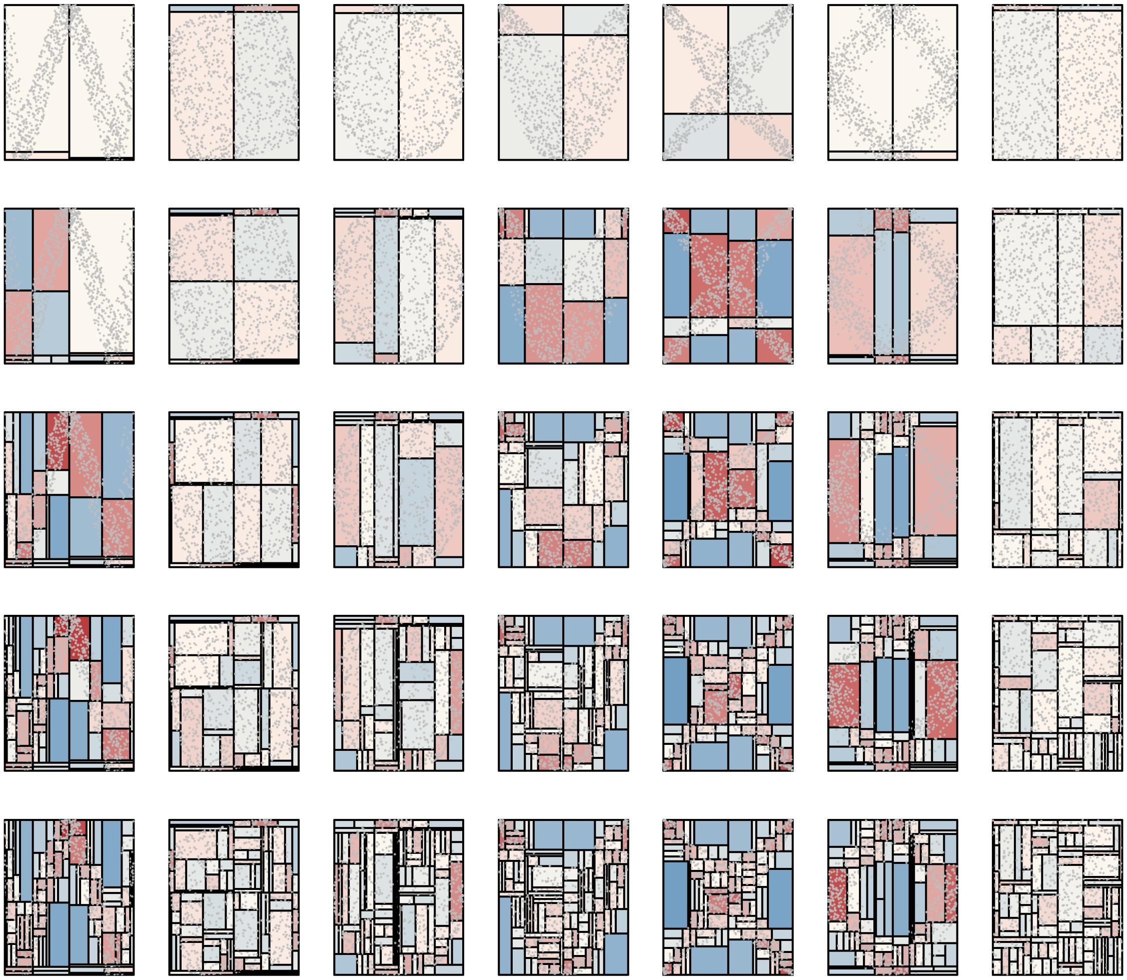


Figure 3.17: Bins for chi score maximizing splits at increasing depths. By row, depths 2, 4, 6, 8, and 10 are displayed in order. The final bins reflect and summarize the pattern of points given to the algorithm.

maximum residual observed across all depths. The rapid early growth of the rotated square can be spied immediately by its shading in the first row. After only two splits, the empty top left corner and dense top right are detected by the maximized binning algorithm and contribute large residuals, leading to a large early χ^2 statistic. Once these regions of low and high density around the margins are identified everywhere, however, the algorithm is left splitting the nearly uniform interior of the rotated square, and so grows parallel to the null paths. One interpretation of this "elbow" point is the point at which the recursive binning has adequately captured the pattern in the data. Additional splits beyond this point are ineffective at increasing the score because they occur essentially randomly based on previous splits.[5]

In contrast, the pattern observed for the ring requires a certain depth to detect. The radial symmetry of this pattern means that most early splits fail to identify the less dense regions in the corners and the centre and the more dense region along the circumference of the ring. It is only as the depth exceeds 6 that strong positive and negative residuals are found in any bins, before that point the residuals are not as large as for the other patterns.

Comparing the final row of Figure 3.17 to randomized binning at a depth of ten in Figure 3.18, a clear advantage of the maximized binning is better representation of the underlying pattern. Especially for the cross, wave, and valley, random binning produces a larger proportion of thin bins which obscure the pattern of residuals and points. In contrast, maximized binning tends to chase local patterns, leading to many small rectangular bins around high density areas that provide a sense of the pattern they summarize. Should a visual summary of the data be desired, maximized binning will give more consistent and clear results than random binning.

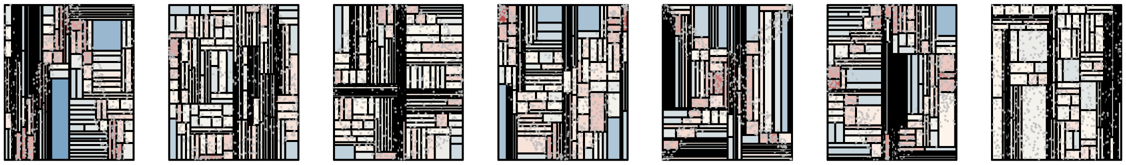


Figure 3.18: Bins for random splitting at a depth of 10

[5] Another interpretation of this behaviour is that the algorithm is effectively splitting the noise. Once large areas of relatively high and low density have been identified, the main aspect of the data determining whether further splits are productive is the level of noise. So, for example, the perfect line of Figure 3.6(b) always benefits from further splits, but the noisier pattern of the valley is adequately captured at some point.

3.5 A real data example: S&P 500 returns

For a real data example, consider the S&P 500 constituent data from Hofert and Oldford (2018). The raw data contains a time series of 505 stock prices from the first day of 2007 to the last day of 2009 for stocks included in the S&P 500 index. The goal is to evaluate the pairwise dependence present within the negative log-returns of the 461 stocks with complete records over this period using recursive binning over all $\binom{461}{2} = 106{,}030$ pairs. The negative log-return of a stock is the negative logarithm of the ratio in its end-of-day price over two consecutive days, explicitly

$$-\log \frac{S_t}{S_{t-1}}$$

for a stock with value S_t at time t. As there are 756 days recorded in the data set, there are 755 log-returns for each stock.

To remove further time dependencies between stock returns day-to-day, the negative log-returns for each stock are fit with an ARMA(1,1)-GARCH(1,1) model and the residuals are taken as a new set of independent pseudo-observations, see details in Hofert and Oldford (2018). The raw data are taken from the `qrmdata` package (Hofert et al., 2022) and processed by code adapted from the `SP500` demo from the `zenplots` package (Hofert and Oldford, 2020) in order to compute the log-returns and generate the pseudo-observations for recursive binning.

Recursive binning as in Section 3.4.4 is applied to each of the 106,030 pairs of 755 pseudo-observations to evaluate dependence. Specifically, splitting maximizes the chi score under the constraint that no bins be created with expected counts less than 5 and splitting is stopped when the expected count in a bin is less than 10, the bin contains no observations, or the depth of the bin is 6. After binning, the χ^2 statistic is computed over the bins to measure the departure of the observed distribution from that expected under uniformity. Figure 3.19 displays the resulting statistic values and n_{bin} for every pair compared to the null quantiles for uniform pairs with 1,000 points from Section 3.4.2 estimated by quantile regression[6]. Hues from blue to red encode the empirical p-value of an observed statistic for an S&P 500 pair at a given number of bins in the null data. Partial transparency (alpha-blending) is also used in this plot, but the sheer number of points makes it ineffective at

[6] This choice of null should be somewhat conservative for the case of 755 points. As shown in Figure 3.12, smaller sample sizes lead to fewer bins and smaller χ^2 statistic values for the same maximum depth. Therefore, taking the χ^2 statistic over simulated uniform samples of a slightly larger sample size provides an approximately correct, but slightly conservative, null distribution.

representing the density, so marginal histograms have been added to give a better sense of the location where values are most concentrated.

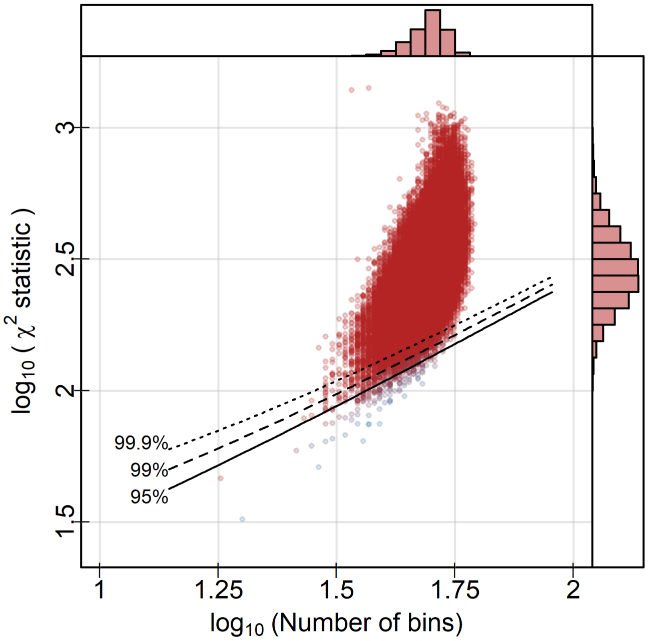


Figure 3.19: The distribution of n_{bin} and χ^2 statistics for the 106,030 S&P 500 pairs split to a maximum depth of 6 compared to some upper quantile estimates of the χ^2 statistic from n_{bin} under the null distribution of independence. Most pairs in the S&P 500 data appear to be highly significant.

Most striking in this plot is the significance of nearly every pair. Only 334 of the 106,030 pairs (0.3%) have empirical p-values less than 0.95, and the χ^2 statistic values are centred well above the fit null quantiles despite their slightly conservative nature. The pairs generally lie in a single large cluster in χ^2 statistic and n_{bin} values and the marginal histograms indicate that within this cluster most points are concentrated in a small region at its center. Only a few dozen pairs lie outside this cluster, including the two very large χ^2 statistic values for relatively small values of n_{bin} and the smallest χ^2 statistic values with the smallest values of n_{bin}. This suggests that almost every pair in this data set contains some level of dependence, inviting further exploration.

First, consider the exceptionally large statistic values. Figure 3.20 displays a matrix of the 36 pairs with the largest χ^2 statistic values in decreasing statistic order from top left

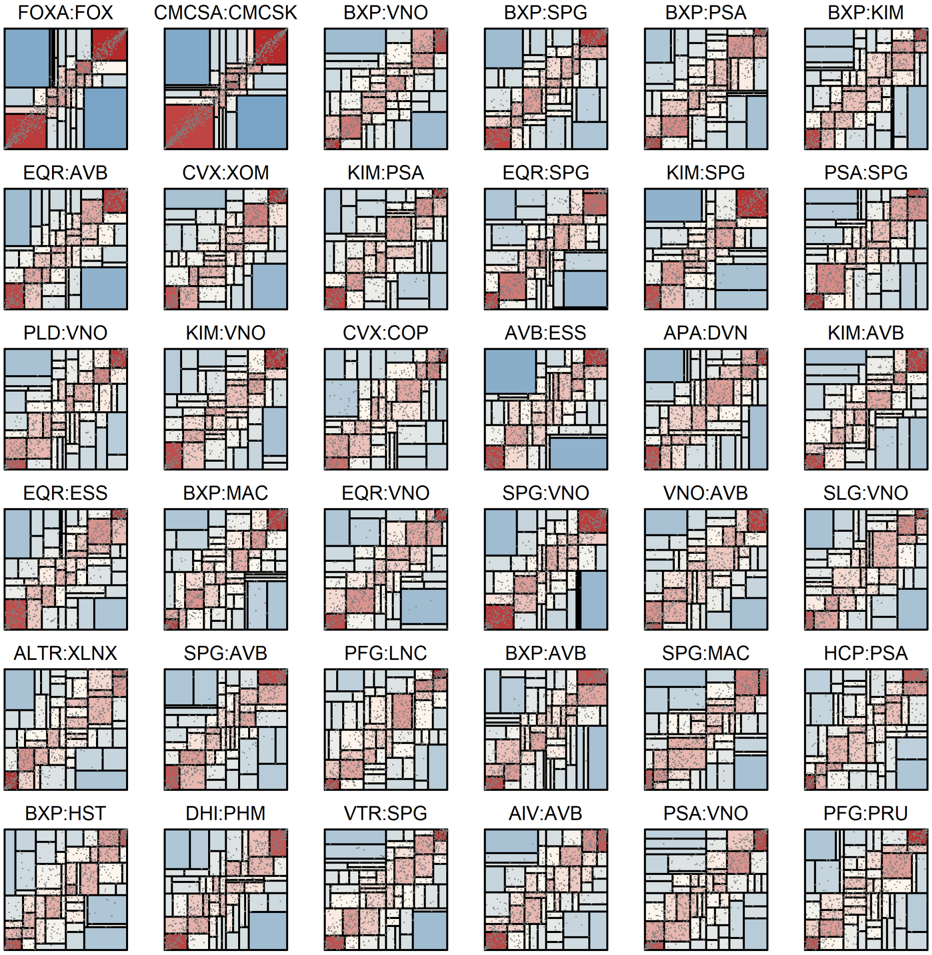


Figure 3.20: The S&P 500 pairs with the largest χ^2 statistic values after recursive binning under the splitting and stopping logic described earlier. All pairs show strong linear dependence, especially in the upper and lower tails.

to bottom right. For each pair, a scatterplots of the marginal ranks is augmented by a plot of the final binnings coloured by the sign of the Pearson residual (red for positive and blue for negative) and shaded by magnitude. The range of hues is kept constant through all subplots to support direct comparisons between any two binnings.

Immediately apparent in every plot is a strong positive linear relationship between the pairs. Bins along the diagonal line from the bottom left to the top right have more points than expected, while those in the top left and bottom right corners have fewer. Most notably, the bins in the top right and bottom left corners tend to have far more observations than expected. This suggests particularly strong *tail dependence* in these pairs, loosely the probability that large or small values of two random variables occur simultaneously. Precisely, upper tail dependence at $p \in [0, 1]$ is defined as

$$P\big(Z_1 > Q_1(p) \,\big|\, Z_2 > Q_2(p)\big)$$

and lower tail dependence as

$$P\big(Z_1 \leq Q_1(p) \,\big|\, Z_2 \leq Q_2(p)\big)$$

for random variables Z_1, Z_2 with respective quantile functions $Q_1(\cdot), Q_2(\cdot)$ (Hofert and Oldford, 2018). As the upper and lower bins approximate these conditional probabilities, albeit for separate quantile values, the shaded residuals in the tails of these plots communicate how much larger the conditional tail probabilities are than would be expected under independence.

Though detecting this tail dependence is not the goal of recursive binning, the most interesting pairs it identifies overlap considerably with the pairs that have the largest upper tail dependence. Indeed, 8 of the top 10 pairs ranked by upper tail dependence in Hofert and Oldford (2018) can be found in Figure 3.20. The top two relationships, in particular, correspond to different classes of stock in the same company. This explains not only their strong dependence (especially in the tails), but also their outlying positions in Figure 3.19 with similarly large χ^2 statistics for relatively few bins.

At the other end of the spectrum, Figure 3.21 displays the pairs with the smallest χ^2 statistic values in decreasing statistic order from top left to bottom right. Residuals are shaded according to the same hues as Figure 3.20, so a comparison of bins between the two plots is possible. In contrast to the pairs with strong association, the points, bins, and residuals show no obvious patterns for any of these pairs. The magnitude of residuals is generally small and the points show essentially random scatter. It is unsurprising that these pairs are deemed unexceptional by the recursive binning algorithm.

Finally consider the pairs with moderate statistic values. Figure 3.22 displays the 36 pairs with values closest to the median χ^2 statistic across all 106,030 pairs. Again, the shading is consistent with all previous plots to allow comparisons. These 'middling' pairs show relatively weak positive, linear relationships without the strong tail dependence of the pairs with the largest χ^2 statistic values. Nonetheless, there is a concentration of bins which are shaded faintly red about the main diagonal and others shaded faintly blue in the top left and bottom right corners distinctive of a positive linear relationship. This

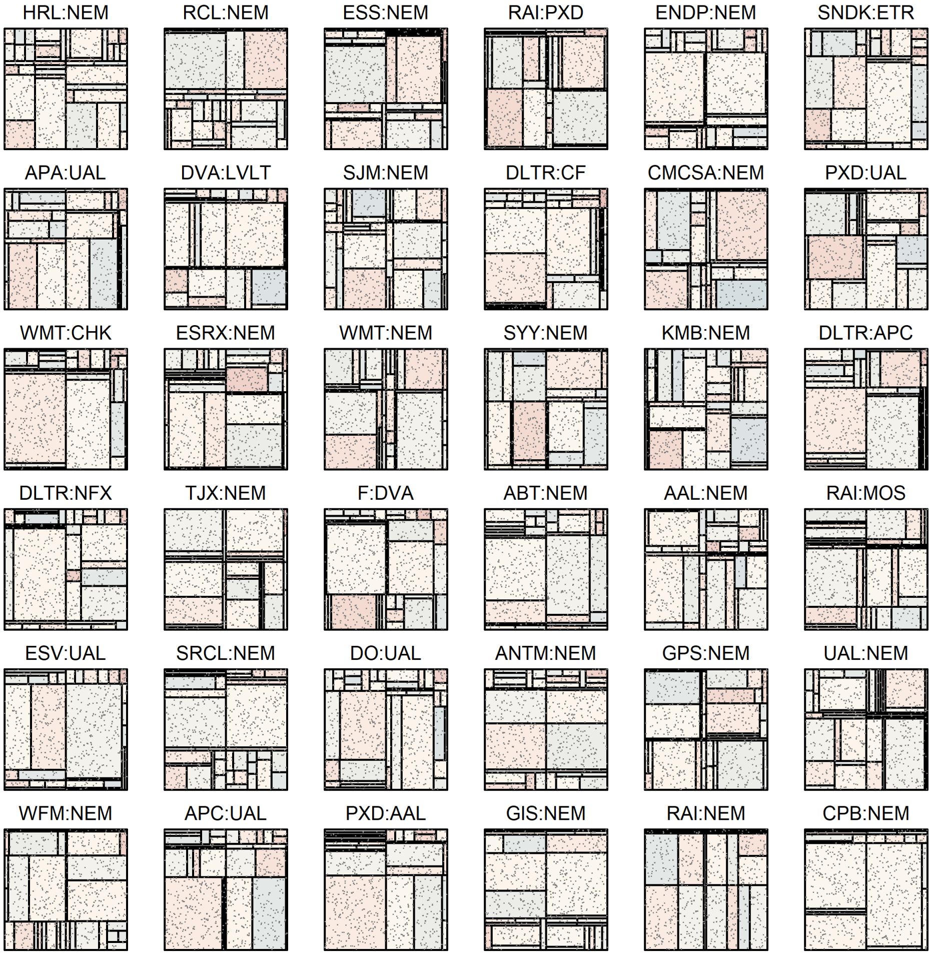


Figure 3.21: The S&P 500 pairs with the smallest χ^2 statistic values. These pairs look like realizations of random uniform data, with no strong patterns to the bins, points, or residuals.

observation, along with clear separation of statistic values for the simulated data patterns of Section 3.4.4 from the null distribution, supports the conclusion that the large statistic values across the majority of pairs in the S&P 500 data are not spurious, but instead reflect the power of recursive binning to detect even weak dependence.

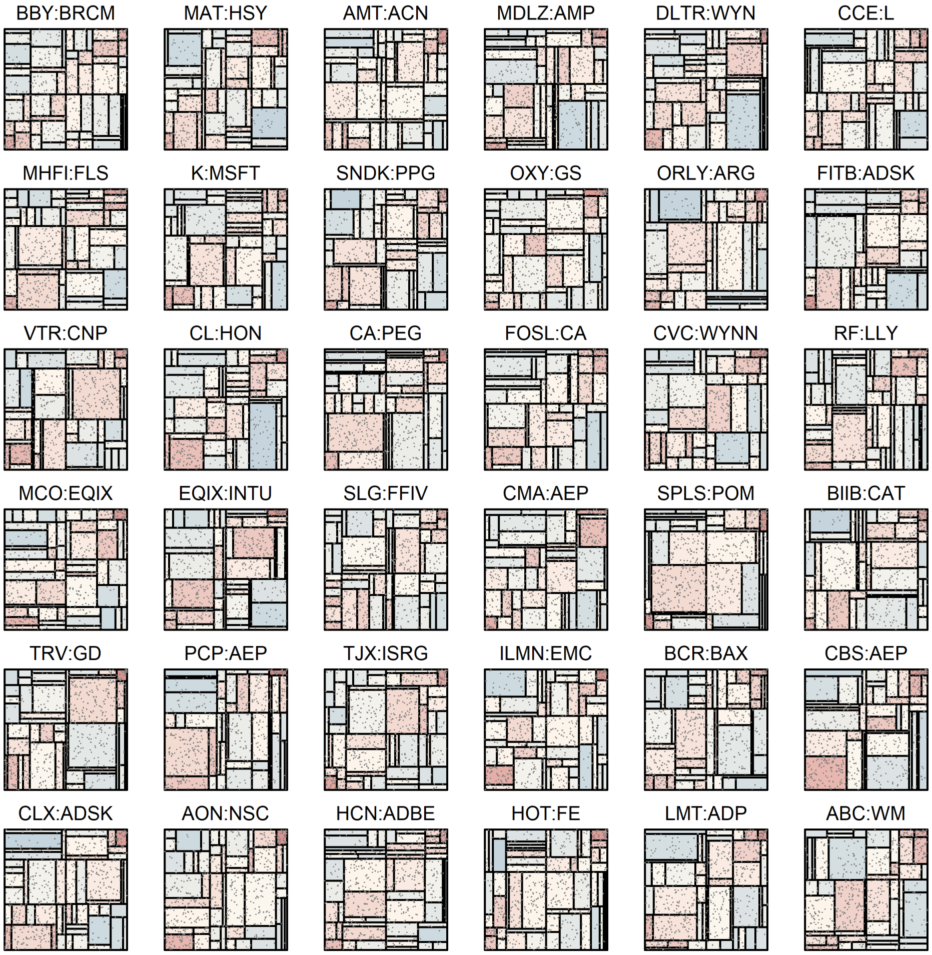


Figure 3.22: S&P 500 pairs with χ^2 statistic values nearest to the median value. Though the dependence is weak compared to Figure 3.20, there is still a concentration of bins with more points than expected along the diagonal and bins with fewer points than expected in the off-diagonal corners, suggesting weak positive linear relationships between these pairs.

3.6 Conclusions

This investigation of recursive binning as a method to measure association garners several key observations. First, and perhaps most important, the power of recursive binning to detect association does not appear to depend greatly on whether chi maximizing or randomly placed edges are used to split bins. Both the maximizing splits and the random splits showed pronounced separation of non-null patterns from the null distribution in the simulation study of Section 3.4.4. Using random splits gives a number of advantages: it is computationally faster, conceptually simpler, and produces a null distribution which is conservatively approximated by the $\chi^2_{n_{bin}-1}$ distribution if the spliting logic and stop criteria maintain the rule of thumb that the expected count of every bin be ≥ 5. If detecting dependence is the only goal of an investigation, random splits may therefore be preferred to maximizing ones despite, or perhaps because of, their simplicity.

In contrast, maximized score binning produces bins which serve as a superior visual summary of a pairwise relationship. Splits which maximize the chi score, for example, will split off empty sections of a bin and separate regions of particularly high density. The end result is a collection of bins which represent the underlying pattern of points much better than the bins which result from random splits. This comes at a cost of inflated statistic values, requiring modeling or simulation at every application to estimate the significance of an observed statistic value.

In either case, the pattern present in the data impacts the path of the algorithm. For simple linear patterns, dependence can be detected with relatively few splits and bins. More complex patterns may take many splits to detect. For many patterns tested here, it seems a natural depth or resolution is present. Splits below this natural depth may increase the χ^2 statistic drastically, but after it is reached the statistic grows more slowly at a rate comparable to the null over successive splits.

Recursive binning is a promising method to measure association. It displays high power in the detection of non-linear relationships, can be used to generate a summary visualization of data pairs, and seems to naturally highlight local dependence such as tail dependence in real data. Though the simulations here are not comprehensive, they are highly suggestive of a practical and powerful tool for sorting and summarizing pairwise relationships in large data sets.

Chapter 4

Pooling Independent Significance Tests

Supposing that M pairwise comparisons have already occurred and resulted in M p-values, a natural question is whether these p-values as a whole constitute evidence against the null hypothesis that there are no interesting patterns in the data. The multiple testing problem arises because answering this question requires different analysis than univariate p-values. A univariate threshold applied to all M p-values, for example, will no longer control the type I error at the level of the threshold. A common approach that controls the type I error (which is called the family-wise error rate in this context) is to use a function to combine the M p-values into a single value which behaves like a univariate p-value. This chapter presents a new framework to choose among these pooling functions along with a proposed pooling function designed with this framework in mind.

Specifically, Section 4.1 introduces the notation necessary to discuss this problem before Section 4.2 introduces some necessary concepts. Means of measuring the prevalence and strength of evidence in the p-values against the null hypothesis are required to understand the framework proposed later. Strength quantifies the degree to which a test favours rejection while prevalence quantifies how commonly tests which do not have the null distribution ('non-null' tests) occur. Prevalence is measured by the proportion of tests which favour rejection while strength is measured by the Kullback-Leibler divergence. Assuming that non-null tests come from a restricted beta family, the power of a UMP method for particular beta distributions is investigated for different values of the prevalence and strength in Section 4.4. A pattern of high power for either strong evidence in a few tests or weak evidence in many tests is noticed and developed into a framework for choosing pooled

p-values in Section 4.5. Following the necessary definitions to develop this framework, including the concepts of central and marginal rejection, it is proved that the significance level required to reject concentrated evidence is always less than that required to reject diffuse evidence. This supports the definition of a coefficient quantifying the degree to which a pooled p-value favours either pattern of evidence.

Section 4.6 proposes a pooling function based on the χ^2 quantile transformation which controls this preference through its degrees of freedom. It is proven that large degrees of freedom give a pooling function which prefers diffuse evidence while small degrees of freedom prefer concentrated evidence. In a simulation study, this proposal is shown to nearly match the UMP when correctly specified and is more robust to errors in specification. These conclusions are extended in Section 4.7 where a sweep of parameter values is used to identify the the most powerful choice for a given sample and suggest a region of most plausible alternative hypotheses within the framework in light of it.

4.1 Introduction

Consider a collection of M independent test statistics $\mathbf{t} = (t_1, \ldots, t_M)^\intercal$ having p-values $\mathbf{p} = (p_1, \ldots, p_M)^\intercal$ for the null hypotheses H_{01}, H_{02}, $\ldots$, H_{0M} – for example, χ^2 tests for the association of M individual genes with the presence of a disease where each H_{0i} asserts no association. Assessing the overall significance of $\mathbf{p}$ while controlling the family-wise error rate (FWER) at the outset of analysis is common practice in meta-analysis and big data applications (Heard and Rubin-Delanchy, 2018; Wilson, 2019). The FWER is the probability of rejecting one or more of $H_{01}, \ldots, H_{0M}$ when all are true, equivalent to the type I error of the joint hypothesis

$$H_0 = \cap_{i=1}^{M} H_{0i}.$$

To emphasize the null distributions, $p_i \sim U = Unif(0,1)$ for all $i \in \{1, \ldots, M\}$, this is often written

$$H_0 : p_1, p_2, \ldots, p_M \overset{\text{iid}}{\sim} U.$$

To test H_0, a statistic $l(\mathbf{p}) : [0,1]^M \mapsto \mathbb{R}$ of the p-values with a distribution that is known or easily simulated under H_0 can be computed. If $l(\mathbf{p})$ has cumulative distribution function (CDF) $F_l(l)$ under H_0, then $l(\mathbf{p})$ admits $g(\mathbf{p}) = 1 - F_l\big(l(\mathbf{p})\big) \sim Unif(0,1)$ such that rejecting H_0 when $g(\mathbf{p}) \leq \alpha$ controls the FWER at level α.[1] $g(\mathbf{p})$ therefore summarizes

[1] Note that the use of the CDF in $g(\mathbf{p}) = 1 - F_l\big(l(\mathbf{p})\big)$ implies that $g(\mathbf{p})$ is identical for any statistic that is a monotonic transformation of $l(\mathbf{p})$.

the evidence against H_0 in a statistic which behaves like a univariate p-value: its magnitude is inversely related to its significance and it is uniform when the null is true.

If we want $g(\mathbf{p})$ to additionally have convex acceptance regions like a univariate p-values, it should be continuous in each argument and monotonically non-decreasing, i.e. $g(p_1, \dots, p_M) \leq g(p_1^*, \dots, p_M^*) \leftrightarrow p_1 \leq p_1^*, \dots, p_M \leq p_M^*$. Functions failing these can behave counter-intuitively, as they may accept H_0 for small p_i only to reject as p_i increases for some margin i. Finally, if there is no reason to favour any margin, g should be symmetric in $\mathbf{p}$. The term evidential statistic refers to $g(\mathbf{p})$ meeting these criteria generally (Goutis et al., 1996), and when testing H_0 they are called pooled p-values. There is no lack of pooled p-value proposals, including the statistics of Tippett (1931), Fisher (1932), Pearson (1933), Stouffer et al. (1949), Mudholkar and George (1977), Heard and Rubin-Delanchy (2018), and Cinar and Viechtbauer (2022).

As all of these methods have convex acceptance regions and control the FWER at α under the rule $g(\mathbf{p}) \leq \alpha$, statistical power against alternative hypotheses is often used to distinguish them. Ideally, one among them would be uniformly most powerful (UMP) against a very broad alternative but this is not possible because of the generality of H_0. Indeed, Birnbaum (1954) proves that if all f_i are strictly non-increasing so that $p_i \sim f_i$ is biased to small values when H_{0i} is false, then there is no UMP test against the negation of H_0,

$$H_1 = \neg H_0 : p_1 \sim f_1, p_2 \sim f_2, \dots, p_M \sim f_M$$

where $f_i \neq U$ for at least one $i \in \{1, \dots, M\}$. As the simulation studies in Westberg (1985), Loughin (2004), and Kocak (2017) readily demonstrate, the number of false H_{0i} and the non-null distributions f_i together specify the unique most powerful test. For the particular case of testing H_0 against H_1 with $f_1 = f_2 = \dots = f_M = Beta(a, b)$ for $a \in (0, 1]$ and $b \in [1, \infty)$, the Neyman-Pearson lemma proves that the pooled p-value $HR(\mathbf{p}; w)$ induced by the statistic

$$l_{HR}(\mathbf{p}; w) = w \sum_{i=1}^{M} \ln p_i - (1 - w) \sum_{i=1}^{M} \ln(1 - p_i) \tag{4.1}$$

with $w = (1 - a)/(b - a) \in [0, 1]$ is uniformly most powerful (UMP) (Heard and Rubin-Delanchy, 2018).

Though $HR(\mathbf{p}; (1 - a)/(b - a))$ is UMP against H_1 for $f_1 = \dots = f_M = Beta(a, b)$, it is rarely assumed that $f_1 = \dots = f_M$ in the search for interesting variable pairs. Rather, some of these are assumed to be non-uniform while others are assumed null. A discussion of this setting therefore requires measures of the prevalence and strength of evidence against H_0, captured by a series of telescoping alternative hypotheses that bridge the gap between H_1 and the setting where $HR(\mathbf{p}; w)$ is UMP.

4.2 Measuring the strength and prevalence of evidence

When proving that no UMP exists for the general hypothesis $H_1 = \neg H_0$, Birnbaum (1954) provides a couple of two-dimensional examples. Though these are demonstrative, they are not instructive for the discussion of tests generally. To the same conclusions, each of Westberg (1985), Loughin (2004), and Kocak (2017) simulate a variety of populations with differing proportions of $\mathbf{p}$ generated under U or some alternative distribution. We begin by defining a telescoping series of alternative hypotheses which capture the settings explored in these empirical investigations.

4.2.1 Telescoping alternatives

Starting at H_1, assume H_{0i} is false only for $i \in J \subset \{1, \dots, M\}$ and quantify the proportion of non-null hypotheses by $\eta = |J|/M$. This implies an alternative hypothesis

$$H_2 : p_i \sim \begin{cases} f_i \neq U & \text{if } i \in J, \\ U & \text{if } i \notin J. \end{cases}$$

As no distinctions between $H_{01}, \dots, H_{0M}$ are made in H_0, η captures the prevalence of evidence against H_0 without loss of generality. If it is additionally assumed that all $i \in J$ have the same alternative distribution $f \neq U$, this gives the alternative hypothesis

$$H_3 : p_i \sim \begin{cases} f & \text{if } i \in J, \\ U & \text{if } i \notin J. \end{cases}$$

Finally, in the particular case where $\eta = 1$, $|J| = M$ and

$$H_4 : p_1, p_2, \dots, p_M \overset{\text{iid}}{\sim} f \neq U$$

is obtained. Though restricted compared to H_1, this alternative makes sense for meta-analysis or repeated experiments, where we could assume all p_i are independently and identically distributed when H_0 is false.

H_4 was distinguished from H_1 as early as Birnbaum (1954) (there called H_A and H_B respectively), but no exploration of intermediate possibilities was considered. All of Westberg (1985), Loughin (2004), and Kocak (2017) explore different combinations of η and f, and so use instances of H_3 for their investigations. When testing H_4 against H_0, Heard and Rubin-Delanchy (2018) prove $HR(\mathbf{p}; w)$ is the UMP pooled p-value if f is from a constrained beta family. By stating clearly $H_1 \supset H_2 \supset H_3 \supset H_4$, a framework for alternative hypotheses is created that contextualizes and relates these previous results. Additionally, the parameter η under H_3 naturally measures the prevalence in $\mathbf{p}$ of evidence against H_0.

4.2.2 Measuring the strength of evidence

Intuitively, if f has a density highly concentrated near zero then it provides "strong" evidence against H_0. This is because p-values generated by f will tend to be smaller than those following U and therefore will be rejected more frequently for any α. Any measure of the strength of evidence in f should therefore increase as the magnitude of f for small values increases.

This relatively simple criterion is challenging to apply to H_2. Every f_i for $i \in J$ may be distinct and a single value characterizing their multiple, potentially very different, departures from U introduces ambiguity. Taking a mean of measures, for example, conflates different instances of H_2. If $J = \{1, 2\}$, the mean strength of evidence cannot distinguish strong evidence in f_1 with weak evidence in f_2 from moderate evidence in both. This difficulty is avoided if all f_i are equal, i.e. if H_3 is chosen as the alternative hypothesis. Indeed, this choice is common in previous empirical investigations.

Westberg (1985) generates $\mathbf{p}$ by testing the difference in means of two simulated normal samples, and measures the strength of evidence by the true difference in means between the generative distributions. This is reasonable, but limits us to tests comparing population parameters and requires assumptions on $\mathbf{t}$ (the tests generating $\mathbf{p}$). Considering f directly, Loughin (2004) takes $p_i \overset{\text{iid}}{\sim} Beta(a, b)$ for all $i \in J$, restricts $a = 1 \leq b$ so that f is non-increasing, and measures the strength of evidence with one minus the median of f: $1 - 0.5^{1/b}$.[2] This measure of strength is limited to $Beta(1, b)$, though it does achieve the intuitive ordering desired. A more general measure that applies to any f is the Kullback-Leibler (KL) divergence, given by

$$D(p, q) = \int_{\mathcal{X}} p(x) \ln \left(\frac{p(x)}{q(x)} \right) dx$$

from density $q(x)$ to density $p(x)$ with mutual support on $\mathcal{X}$[3].

[2] Kocak (2017) also uses $p_i \sim Beta(a, b)$ but takes the broader $0 < a \leq b$ and does not attempt to measure the strength of f's departure from U.

[3] The KL divergence is also known as the relative entropy from $q(x)$ to $p(x)$.

Widespread application of the KL divergence in information theory and machine learning aside, one interpretation of this measure suits pooled hypothesis testing nicely. Joyce (2011) describes the KL divergence as the extra information encoded in $q(x)$ when expecting $p(x)$. The explicit assumption underlying the pooled test of H_0 is that $f_i = U$ for all $i \in \{1, \dots, M\}$, which gives a natural expected density $q(x) = U(x)$. Furthermore, this density is, in some sense, minimally informative: no region of $[0, 1]$ is distinguished from any other by U. Any additional information which discriminates particular regions of $[0, 1]$, in particular values near 0, will help inform rejection.

4.2.3 Choosing a family for the alternative distribution

The beta family of distributions is appealing as a model for the alternative distribution f under H_3 for two main reasons. First, it has the same support as U without the need for adjustment. Second, a wide variety of different density shapes can be achieved by changing its two parameters and it has a non-increasing density whenever $a \leq 1 \leq b$. This latter quality makes it ideal to model alternative p-value distributions under the assumption that p_i is biased to small values when H_0 is false. These features are likely why it is commonly used in the literature (Loughin, 2004; Kocak, 2017). More significantly, the Neyman-Pearson lemma proves that $HR\,(\mathbf{p}; w)$ is UMP for $p_1, \dots, p_M \overset{\text{iid}}{\sim} Beta\big(a, 1/w + a(1 - 1/w)\big)$ when $0 \leq w, a \leq 1$, and so a best-case benchmark for power exists to compare to other tests (Heard and Rubin-Delanchy, 2018). Were another distribution chosen, this important reference point would be absent.

When $f = Beta(a, b)$ the KL divergence also has a relatively simple expression. Letting $u(x) = 1$ be the uniform density and $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ be the beta density with parameters a and b, $D(u, f)$ is given by

$$D(u, f) = -\int_0^1 \ln f(x) dx = a + b + \ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\right) - 2.$$

For the case of the beta densities where $HR\,(\mathbf{p}; w)$ is UMP, i.e. $a \leq 1 \leq b$, we can express this in terms of a and $w = (1 - a)/(b - a)$:

$$D(u, f) := D(a, w) = 2a + \frac{1-a}{w} + \ln\left(\frac{\Gamma(a)\Gamma\left(\frac{1}{w} + a\left[1 - \frac{1}{w}\right]\right)}{\Gamma\left(2a + \frac{1-a}{w}\right)}\right) - 2. \tag{4.2}$$

This is a less convenient expression, but provides a direct link between the strength of evidence $D(a, w)$ and the UMP test against H_4 by the shared parameter w. Interestingly, though the strength depends on both a and w, the UMP test only depends on the latter.

To visualize the strength of evidence provided by different beta distributions, shaded inset densities for different choices of w and the KL divergence $D(a, w)$ are displayed in Figure 4.1 placed at $\ln(w)$, $\ln D(a, w)$.[4] When $a = w = 1$, $f = u$ so $D(a, w) = 0$. Decreasing either of a or w from 1 causes a larger magnitude of f near zero, with the limiting case $a = w = 0$ corresponding to a degenerate distribution at zero. This general trend in shapes is seen in Figure 4.1, decreasing w or increasing $D(a, w)$ increases the concentration of f near zero.

This suggests a limitation in the KL divergence. Though the ordering of beta densities by $D(a, w)$ generally conforms to the intuitive rule – larger divergences correspond to inset densities with greater magnitude near zero – the parameter w is still relevant to the shape. The KL divergence does not distinguish between departures from uniform near 1 and near 0, despite their relevance for rejection when, for example, rejecting the null hypotheses of p-values below a threshold. This is particularly obvious in the final row of inset plots in Figure 4.1. When $\ln(w) = -6$, the density for $\ln D(a, w) = -5$ is mostly flat, with a slight increase in density near zero and a large decrease near one. When $\ln(w) = 0$, however, a much larger spike in the density near zero is present.

Nonetheless, the ordering on beta densities imposed by the KL divergence is still very informative. It classifies, generally, which densities are biased to small values. Therefore, $D(a, w)$ provides a convenient measure of the strength of evidence contained in f under the alternative hypothesis, and has computationally convenient form for the case of interest where $f = Beta(a, b)$.

4.3 Pooled p-values

Having explored the possible alternatives to H_0 and some specific instances of these alternatives, we can now focus on the pooled p-values meant to test H_0 against these alternatives. Recall that any pooled p-value $g(\mathbf{p})$ is derived from the null distribution of a corresponding statistic $l(\mathbf{p})$. The statistics underlying pooled p-values are of two basic kinds, based either on the k^{th} order statistic $p_{(k)}$ of $\mathbf{p}$ (Tippett, 1931; Wilkinson, 1951) or on transformations of each p_i using some quantile function $F^{-1}(p)$ (Fisher, 1932; Pearson, 1933; Stouffer et al.,

[4] The densities of these inset plots were determined for each w, $D(a, w)$ pair by finding the corresponding a value numerically using Equation (4.2). This is the cause of the irregular plots in the upper right corner: when w is large enough the required a to obtain a set $D(a, w)$ is too small to be represented as a floating point double alongside $w \approx 1$. This is inconvenient, but these cases correspond to densities that are effectively degenerate at zero in any case.

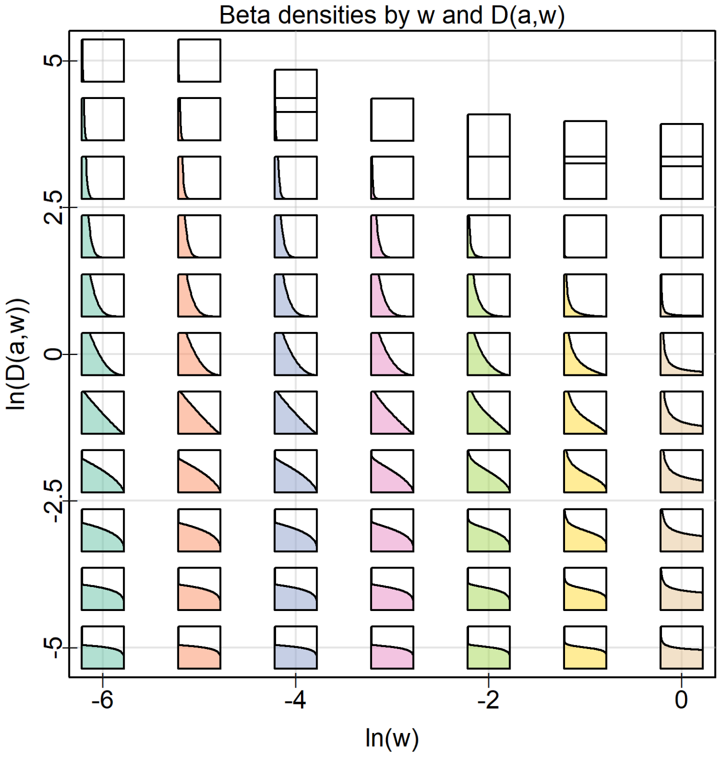


Figure 4.1: Densities and log KL divergences of $Beta\big(a, 1/w + a(1-1/w)\big)$ from U by w. Insets display the densities over $[0,1]$ horizontally and $[0,2]$ vertically and are centred at the $\ln(w)$, $\ln D(a,w)$ coordinates corresponding to the density. These densities range from nearly vertical at 0 when $D(a,w) \approx e^5$ to nearly uniform when $D(a,w) \approx e^{-5}$.

1949; Lancaster, 1961; Edgington, 1972; Mudholkar and George, 1977; Heard and Rubin-Delanchy, 2018; Wilson, 2019; Cinar and Viechtbauer, 2022). The former case takes the general form

$$ord\,(\mathbf{p}; k) = \sum_{l=k}^{M} \binom{M}{l} p_{(k)}^{l} (1 - p_{(k)})^{M-l}, \tag{4.3}$$

and the latter

$$g(\mathbf{p}) = 1 - F_M \left(\sum_{i=1}^{M} c_i F^{-1}(1 - p_i) \right) \tag{4.4}$$

where $c_1, \dots, c_M \in \mathbb{R}$ are known constants, typically $c_1 = \dots c_M = 1$. Equation (4.3) gives the pooled p-value based on $l_{ord}(\mathbf{p}; k) = p_{(k)}$, as with $Tip(\mathbf{p}) = ord\,(\mathbf{p}; 1) = 1 - (1 - p_{(1)})^M$ (Tippett, 1931), while different choices of F and F_M in Equation (4.4) give the pooled p-value based on $l(\mathbf{p}) = \sum_{i=1}^M F^{-1}(1 - p_i)$.

Obvious choices include the normal and gamma families, as these are closed under addition. For example, letting Φ be the $N(0, 1)$ CDF and choosing $F(x) = \Phi(x)$ and $F_M(x) = \Phi(x/\sqrt{M})$ gives

$$Sto(\mathbf{p}) = 1 - \Phi\left(\sum_{i=1}^M \Phi^{-1}(1 - p_i)/\sqrt{M}\right) \tag{4.5}$$

based on $l_{Sto}(\mathbf{p}) = \sum_{i=1}^M \Phi(1 - p_i)$ from Stouffer et al. (1949). Letting $G_{k,\theta}(x)$ be the CDF of the gamma distribution with shape parameter k and scale parameter θ, taking $F(x) = G_{k,\theta}(x)$ and $F_M(x) = G_{Mk,\theta}(x)$ gives

$$gam(\mathbf{p}) = 1 - G_{Mk,\theta}\left(\sum_{i=1}^M G_{k,\theta}^{-1}(1 - p_i)\right) \tag{4.6}$$

based on $l_{gam}(\mathbf{p}) = \sum_{i=1}^M G_{k,\theta}^{-1}(1 - p_i)$. The pooled p-value gam requires the choice of parameters θ and k; choosing $\theta = 1$ gives the k-parameterized gamma method from Zaykin et al. (2007) while $k = 1$ and $\theta = 2$ gives Fisher's method from Fisher (1932).[5]

R. A. Fisher's method deserves some additional consideration alongside an analogous proposal from Karl Pearson around the same time[6]. Both of $l_{Fis}(\mathbf{p}) = -2\sum_{i=1}^M \ln p_i$ (Fisher, 1932) and $l_{Pea}(\mathbf{p}) = -2\sum_{i=1}^M \ln(1 - p_i)$ (Pearson, 1933) were originally proposed only as computational tricks for the distribution of $\prod_{i=1}^M p_i$ (Wallis, 1942), but are also quantile transformations based on the χ_2^2 distribution. Let

$$F_\chi(x; \kappa) = \int_0^x \frac{1}{2^{\kappa/2}\Gamma(\kappa/2)} t^{\kappa/2-1} e^{-t/2} dt, \tag{4.7}$$

[5] Should the p-values be weighted, the gamma distribution also allows more stable weighting than the constants $c_1, \dots, c_M$ in Equation (4.4) by analogously giving each p_i an individual shape parameter k_i (or equivalently χ^2 degrees of freedom κ_i) (Lancaster, 1961).

[6] Owen (2009) notes that Pearson's proposal is actually slightly different than has been credited in the literature following Birnbaum (1954). Though the use of $\sum_{i=1}^M \ln(p_i)$ was suggested by Karl Pearson, it was in the context of comparing the value to $\sum_{i=1}^M \ln(1 - p_i)$ and taking the minimum of the two. Owen (2009) develops this idea into a series of pooling functions that perform best for concordant or discordant effect estimates in a regression setting.

be the CDF of the χ^2_κ distribution, in particular $F_\chi(x;2) = 1 - e^{-x/2}$. Therefore, $F_\chi^{-1}(1 - p; 2) = -2\ln p$ and so taking $F(x) = F_\chi(x;2)$ and $F_M(x) = F_\chi(x;2M)$ gives

$$Fis(\mathbf{p}) = 1 - F_\chi\big(l_F(\mathbf{p}); 2M\big) = 1 - F_\chi\left(-2\sum_{i=1}^{M} \ln p_i; 2M\right) = 1 - F_\chi\left(\sum_{i=1}^{M} F_\chi^{-1}(1-p_i;2); 2M\right)$$

consistent with Equation (4.4). In contrast, $l_{Pea}(\mathbf{p})$ uses lower tail probabilities by taking $F_\chi(p_i;2)$, and so

$$Pea(\mathbf{p}) = F_\chi\left(\sum_{i=1}^{M} F_\chi^{-1}(p_i;2); 2M\right)$$

departs from the general quantile transformation equation.

4.4 Benchmarking the most powerful test

Alone, $Fis(\mathbf{p})$ is preferred to $Pea(\mathbf{p})$, as Birnbaum (1954) found $Pea(\mathbf{p})$ inadmissible for the alternative hypothesis H_1 if the test statistics $t_1, \dots, t_M$ independently follow particular distributions in the exponential family. $Fis(\mathbf{p})$, in contrast, was admissible in this setting and is optimal in some sense for others (Littell and Folks, 1971; Koziol and Perlman, 1978). Together, the statistics for these two pooled p-values are combined in $l_{HR}(\mathbf{p}; w) = -\frac{w}{2} l_{Fis}(\mathbf{p}) + \frac{1-w}{2} l_{Pea}(\mathbf{p})$, the UMP statistic under H_4 when $f = Beta(a, 1/w + a(1 - 1/w))$ (Heard and Rubin-Delanchy, 2018). Intuitively, then, $HR\,(\mathbf{p}; w)$ is a test based on a linear combination of the lower and upper tail probabilities of $\mathbf{p}$ transformed to χ^2_2 quantiles. When $w = 1$ it considers the upper tail alone and when $w = 0$ the lower tail alone. Note that $l_{HR}(\mathbf{p}; w)$, the statistic, and $HR\,(\mathbf{p}; w)$, the unique corresponding pooled p-value, will be used interchangeably thoughout this paper.

Unfortunately, this imbues $HR\,(\mathbf{p}; w)$ with some practical shortcomings. While both $l_{Fis}(\mathbf{p})$ and $l_{Pea}(\mathbf{p})$ are χ^2_{2M} distributed under H_0, they are not independent and so their combination in $l_{HR}(\mathbf{p}; w)$ does not have a closed-form distribution. Approximation as in Mudholkar and George (1977) or simulation must be used to determine the α quantiles or visualize their distribution. This means, for example, the kernel density estimates of $l_{HR}(\mathbf{p}; w)$ by w in Figure 4.2 required the generation of 100,000 independent simulated samples of the case $p_1, \dots, p_{10} \overset{\text{iid}}{\sim} U$. Further, H_4 is the least general of the telescoping alternative hypotheses $H_1 \supset H_2 \supset H_3 \supset H_4$ and is a less natural choice than H_3 if only a subset of tests are thought to be significant. Empirical and theoretical investigations show the most powerful test depends on η and f under H_3, so $HR\,(\mathbf{p}; w)$ may be less exceptional under this more general hypothesis. Finally, $HR\left(\mathbf{p};w\right)$ is only UMP if w is known, which is seldom true in practice.

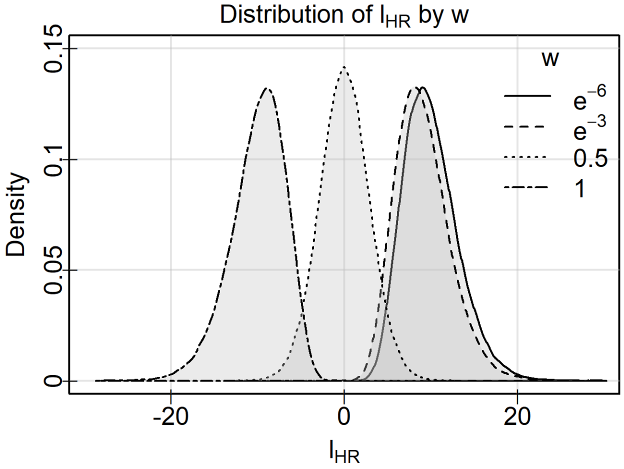


Figure 4.2: Densities of l_{HR} by w when $M = 10$. Solid lines indicate $w = e^{-6}$, dashed lines $w = e^{-3}$, dotted lines $w = 1/2$, and dot-dashed lines $w = 1$. Note how $w = 1$ and $w = e^{-6}$ are nearly mirrored distributions skewed away from zero and $w = 1/2$ is symmetric at zero.

These practical difficulties manifest in two possible errors, assuming H_4 when H_3 is true and choosing w when the true parameter is ω, and four cases of mis-specification depending on which is present. The power of $HR\left(\mathbf{p};w\right)$ under all four cases was investigated by a simulation study at level $\alpha = 0.05$. For both of H_3 and H_4 and a range of mis-specified w, $HR\left(\mathbf{p};w\right)$ was applied to factorial combinations of $D(a,w)$, w, and M covering their respective ranges. $D(a,w)$ was chosen on the log scale ranging from -5 to 5 at 0.5 increments, w was chosen on the log scale at values $-6, -5, \ldots, 0$, and the values of M were 2, 5, 10, and 20.

For each of the parameter settings, $l_{HR}(\mathbf{p};w)$'s 0.95 quantiles under H_0 given w and M are simulated by generating 100,000 independent samples $\mathbf{p}_i = p_{i1}, \ldots, p_{iM} \overset{\text{iid}}{\sim} U$, computing $l_{HR}(\mathbf{p}_i;w) = l_{HRi}$, and taking the 0.95 quantile of the sequence $l_{HR1}, \ldots, l_{HR100{,}000}$ as the 0.95 quantile of $l_{HR}(\mathbf{p};w)$ under H_0. Note that the value of a and the case do not impact this simulation, and so these quantiles are used across all a values under both H_3 and H_4.

Next a Monte Carlo estimate of the probability of rejecting H_0 using $l_{HR}(\mathbf{p};w)$ (i.e. the power of $l_{HR}(\mathbf{p};w)$) is generated for each case. The details of this estimate depend on whether H_3 or H_4 was used to generate the data and whether w was known or not when choosing l_w. To begin, consider the benchmark case when w is known and the data are generated according to H_4.

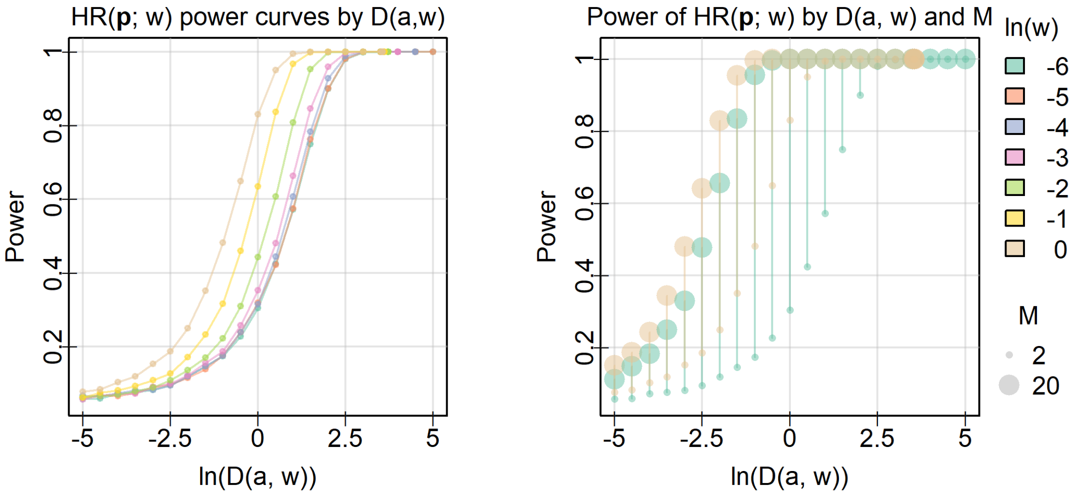
4.4.1 Case 1: correct hypothesis and w

If the data are generated under H_4 and w is chosen correctly, $HR\left(\mathbf{p}; w\right)$ is UMP and so provides the greatest power of any test. In this case, the probability of rejection for a given a, w, and M setting is estimated by generating 10,000 independent samples $\mathbf{p}_i = p_{i1}, \ldots, p_{iM} \overset{\text{iid}}{\sim} Beta\big(a, 1/w + a(1 - 1/w)\big)$. $l_{HR}(\mathbf{p}_i; w)$ is computed for each sample and compared to the simulated 0.95 quantile of $l_{HR}(\mathbf{p}; w)$ under H_0[7]. If the value is larger than the quantile H_0 is correctly rejected and if it is smaller the test incorrectly fails to reject H_0. Power is estimated as the proportion of the 10,000 generated samples which correctly lead to the rejection of H_0, giving a worst-case standard error less than 0.005 based on the binomial distribution.

This procedure is applied to all settings of M, w, and $D(a, w)$ outlined in Section 4.4, corresponding to the beta densities of Figure 4.1. The beta parameter a was determined for a given w and $D(a, w) = D$ by finding the root of $f(x) = D(x, w) - D$, while the parameter b is given by $1/w + a(1 - 1/w)$. This results in an imbalance in the settings, as $w > e^{-4}$ require a less than the typical floating point minimum value to achieve $D(a, w) = 5$. The impact on the coverage of $D(a, w)$ for each w choice as a result of this, however, was slight as shown by the small gap in plots in Figure 4.1.

Figure 4.3(b) shows a scatterplot of the power of $HR\left(\mathbf{p}; w\right)$ by $D(a, w)$ for every setting when $M = 2$ and $M = 20$, and Figure 4.3(a) shows the power curves of $HR\left(\mathbf{p}; w\right)$ by $D(a, w)$ coloured by w. Generally, power increases in both M (the number of p-values) and $D(a, w)$ (the KL divergence), which is unsurprising. Decreasing $D(a, w)$ for $f = Beta\big(a, 1/w + a(1 - 1/w)\big)$ necessarily gives a density closer to u which is therefore less likely to cause rejection, thus reducing the power. At a certain threshold on $D(a, w)$, $f \approx u$ and so the power will be approximately α for all $D(a, w)$ less than the threshold. Similarly, when $D(a, w)$ is large enough, rejection occurs almost certainly and so the power is constant at one. This suggests that most changes in the power of $HR\left(\mathbf{p}; w\right)$ occur for moderate levels of evidence; when the evidence is too weak or too strong, all pooled p-values will perform equally poorly or well. Regardless of $D(a, w)$, increasing the number of p-values makes any distributional differences between f and u more easily detectable,

[7] This corresponds to testing whether $HR\left(\mathbf{p}_i; w\right) \leq 0.05$.



(a)

(b)

Figure 4.3: Power of $HR(\mathbf{p}; w)$ by the KL divergence coloured by w and scaled by M displayed using (a) power curves when $M = 2$ joined by w and (b) lines from $M = 2$ to $M = 20$ for $w = e^{-6}$ and 0. Increasing either M or the KL divergence increases the power and the greatest rate of change in both occurs when the divergence is in the interval (e^{-2}, e^{2}).

as the whole sample KL divergence is given by $MD(a, w)$. This is why the impact of M on the power in Figure 4.3(b) increases in $D(a, w)$.

An interesting feature of Figure 4.3(a) is the ordering of the lines by decreasing w for essentially every KL divergence. This pattern holds almost everywhere with the exception of several crossings of the lowest power curves. Referring to Figure 4.1, increasing w for a given $D(a, w)$ increases the magnitude of the density near zero, thereby increasing the probability of extremely small p-values. This difference was noted in Section 4.2.2, and causes the expected increase in the power of the UMP.

4.4.2 Case 2: correct hypothesis with mis-specified w

Of course, the curves of Figure 4.3(a) are not realistic. In practice, $D(a, w)$ is set by the data generating process and we lack the perfect knowledge of w and f needed to attain them. Suppose that $\mathbf{p}$ is generated under H_4 with $f = Beta(a, \beta)$ and $a \in [0, 1]$, $\beta \in [1, \infty)$ but that $HR(\mathbf{p}; w)$ is used instead of the correct $HR(\mathbf{p}; \omega)$ with $\omega = (1-a)/(\beta - a)$. Though $HR(\mathbf{p}; w)$ is from the same family of tests, it is no longer UMP and so will not match the power achieved by $HR(\mathbf{p}; \omega)$.

The reduction of power from using $HR(\mathbf{p}; w)$ when the UMP is $HR(\mathbf{p}; \omega)$ for each a, w, ω, and M setting from Section 4.4.1 was determined by generating 10,000 independent samples $p_{i1}, \dots, p_{iM} \overset{\text{iid}}{\sim} Beta\big(a, 1/\omega + a(1 - 1/\omega)\big)$ and computing $HR(\mathbf{p}_i; w)$. The proportion of samples rejected based on the simulated null 0.95 quantiles was recorded as the power of $HR(\mathbf{p}; w)$ when data were truly generated with the parameter ω. This procedure was repeated for each of $w = e^{-6}, e^{-3}, 1/2$, and 1 for every parameter setting. Figure 4.4 displays the results when $M = 2$ for $\omega = e^{-6}$ and 1. As one setting of w matches ω in this case, the highest curve displays the power of the UMP.

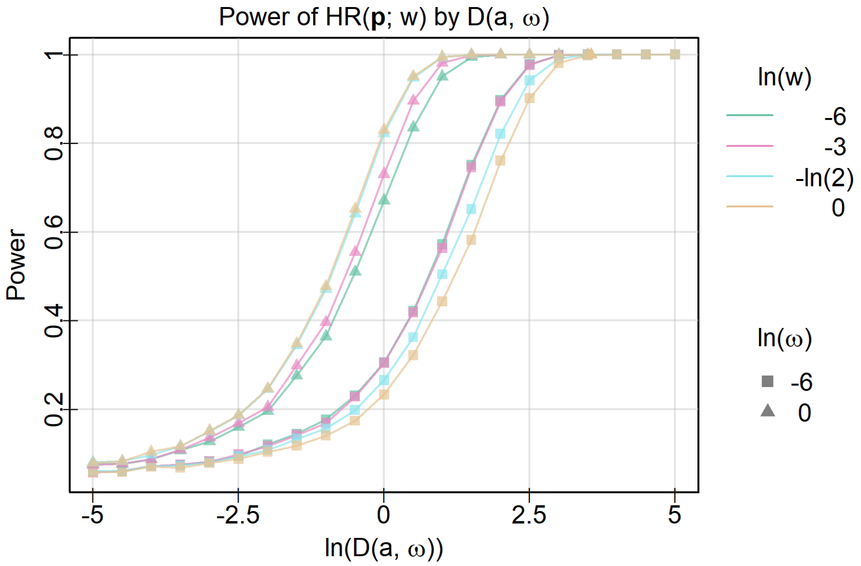


Figure 4.4: Power curves for $HR(\mathbf{p}; w)$ against $D(a, \omega)$ when $M = 2$ with colours giving the value of w and points along the curves giving ω. Mis-specification of ω has less impact on power than the non-null distribution f, but the greater the difference between w and ω, the greater the reduction in power.

Two patterns stand out in this plot. First, it is clear that mis-specification impacts the power less than the distribution of p-values under H_4. Despite the incorrect value w in $HR(; w)$, the mis-specified curves have the same shape in $D(a, \omega)$ as the UMP $HR(; \omega)$. In most cases, mis-specification results in only a slight decrease in power.

Second, larger mis-specifications lead to larger decreases in power. The curves for every w are ordered by their distance from ω for both $\omega = 1$ and e^{-6}. When $\omega = 1$, for example, the lines are ordered so $w = 1$ has the greatest power followed closely by $w = 1/2$ and more distantly by e^{-3} and e^{-6}. This pattern is reversed when $\omega = e^{-6}$. In both cases, mis-specification has the greatest impact on power for moderate $D(a, \omega)$ while mis-specification has scarcely any impact for large or small values of $D(a, w)$ where all powers converge to 1 or α, respectively. It seems prudent, therefore, to choose a middling value such as $w = 1/2$ if ω is not known but H_4 is suspected to avoid the worst impacts of mis-specification when using $HR(\mathbf{p}; w)$

4.4.3 Case 3: incorrect hypothesis, correctly specified w

Assuming all p_i are non-null is not always appropriate, however. The investigations of the previous sections are a useful benchmark and exploration of the impact of mis-specification, but do not address the natural case of H_3 when a handful of significant variables are assumed to exist in a host of insignificant ones. We therefore repeat the benchmark experiment of Section 4.4.1 under H_3 by generating 10,000 independent samples $\mathbf{p}_i$ of size $M = 10$ with the first $M\eta$ distributed according to $f = Beta\big(a, 1/w + a(1 - 1/w)\big)$ and the latter $M(1 - \eta)$ distributed according to U for $\eta \in \{0, 0.1, \ldots, 1\}$ under each of the settings explored previously. As $HR(\mathbf{p}; w)$ is symmetric in all of its arguments, this gives no loss of generality. Figure 4.5 displays contours of the power surface, the proportion of correct rejections of H_0, as a function of η and $D(a, w)$ facetted by $\ln(w)$.

Figure 4.5 places the measure of strength of evidence horizontally and the measure of prevalence vertically. Including $\eta = 0$ captures the behaviour of $HR(\mathbf{p}; w)$ under H_0 along the bottom edge of the power surface and including $\eta = 1$ captures its behaviour under H_4 along the top edge for a range of KL divergences. In particular, this means the top left part of each subplot corresponds to a generative process for $\mathbf{p}$ with relatively weak evidence in all p-values of $\mathbf{p}$ and the bottom right part corresponds to a generative process with strong evidence concentrated in a small number of p-values. Figures 4.3 and 4.4 show that rejection occurs at a rate of α once the evidence is weak enough, so the top left corner is less interesting than the top centre, which gives the power for tests of moderate strength spread throughout $\mathbf{p}$.

When w is small and $l_{HR}(\mathbf{p}; w) \approx l_{Pea}(\mathbf{p})/2$, $HR(\mathbf{p}; w)$ is relatively weak when strong evidence is concentrated in a few tests. The contours for $w = e^{-6}$ are nearly horizontal for log KL divergences between 2.5 and 5 and the power is nearly identical in the bottom right and the top left corners. On the other hand, when $w \approx 1$ and $l_w(\mathbf{p}) \approx -l_{Fis}(\mathbf{p})/2$,

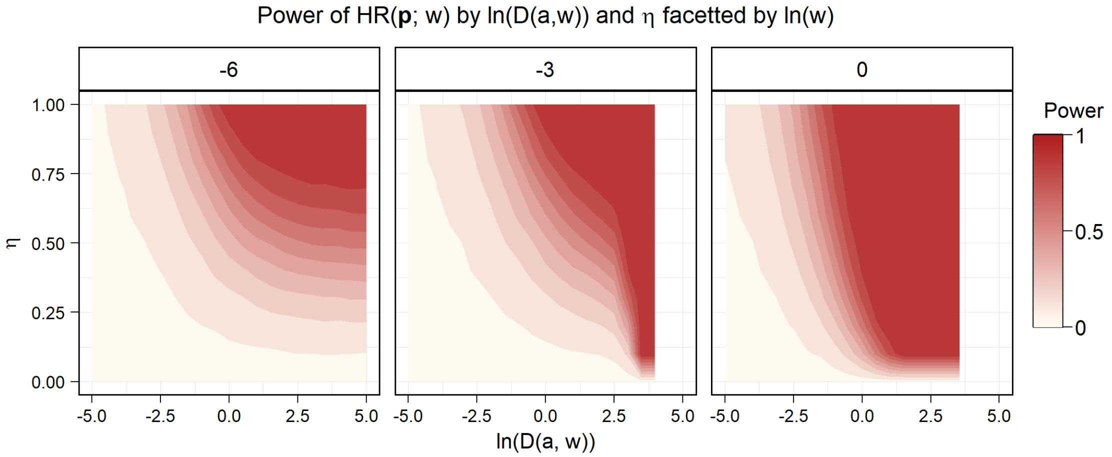


Figure 4.5: The power of $HR\,(\mathbf{p}; w)$ under H_3 by $D(a, w)$ and η split by $\ln(w)$. Note how the contour for a power of one extends nearly to the bottom of the plot when $w = 1$, but stops near 0.75 when $w = e^{-6}$.

$HR\,(\mathbf{p}; w)$ is relatively powerful when evidence is strong and concentrated in a few tests. This is starkly visible when $w = 1$, the power contours are nearly vertical, and the bottom right corner has a power of almost 1. Between these extremes, the power contours display a mix of these seemingly oppositional sensitivities to strength and prevalence.

4.4.4 Case 4: when both the hypothesis and w are incorrect

Finally, consider the most pessimistic case. In the preceding section, w at least matched the non-null distribution f under H_3, but this may not always be so. Suppose, instead, everything is mis-specified. That is, generate $M\eta$ p-values according to $f = Beta\big(a, 1/\omega + a(1 - 1/\omega)\big)$ and $M(1 - \eta)$ from the uniform distribution and compute $HR\,(\mathbf{p}; w)$ for each of $w = e^{-6}, e^{-3}, 1/2$, and 1 for every setting from Section 4.4.3 and repeat this generation 10,000 times. The power, given by proportion of rejections, follows an interesting pattern in the strength and prevalence of evidence, which is illustrated by the power contours of $HR\,(\mathbf{p}; 1)$ in Figure 4.6 and the difference in power contours in Figure 4.7.

Compared to the power contours of $HR\,(\mathbf{p}; \omega)$ in Figure 4.5, the contours of $HR\,(\mathbf{p}; 1)$ in Figure 4.6 show greater red saturation – and thus greater power – in the lower right corner of every panel. Thus, $HR\,(\mathbf{p}; 1)$ has power greater than or equal to that of $HR\,(\mathbf{p}; \omega)$

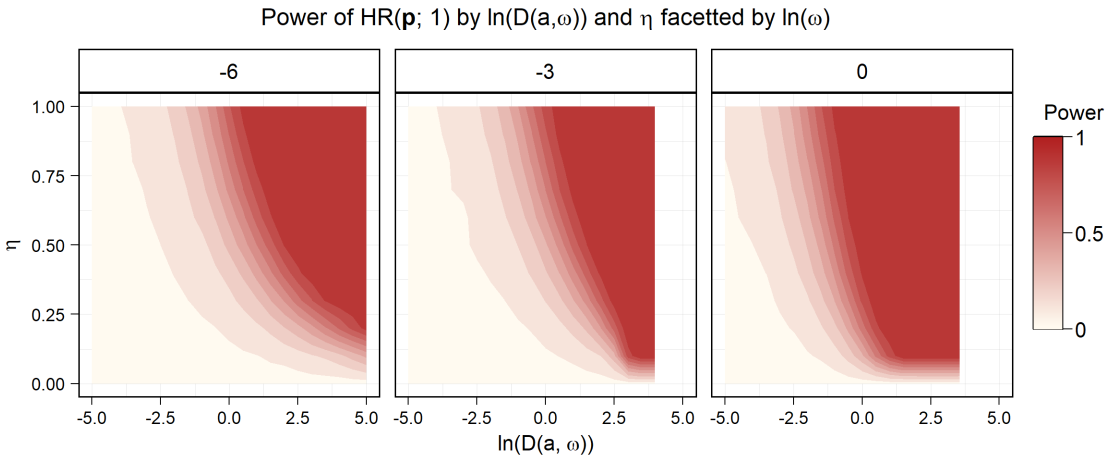


Figure 4.6: Power contours of $HR\,(\mathbf{p};1)$ under H_3 by $D(a,\omega)$ and η displayed using the saturation palette of Figure 4.5.

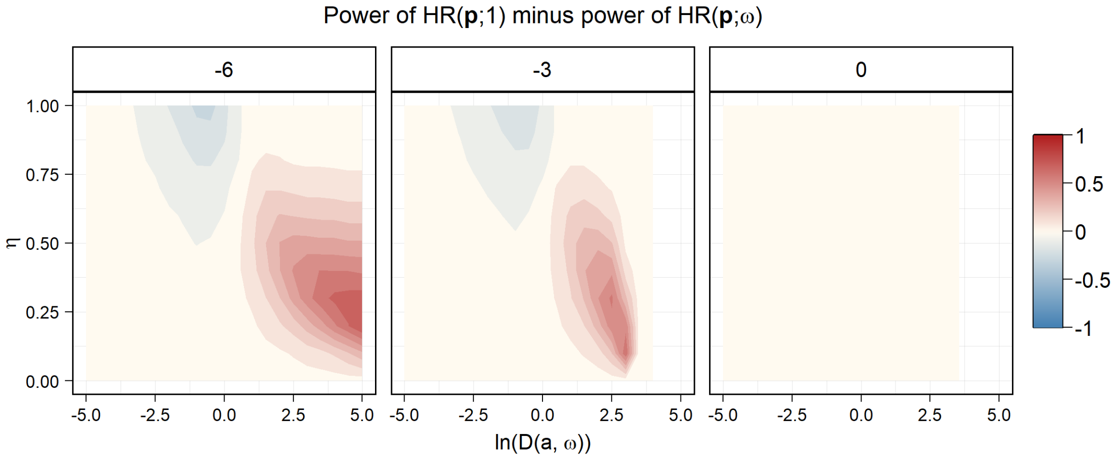


Figure 4.7: Contours of the difference in power of $HR\,(\mathbf{p};1)$ and $HR\,(\mathbf{p};\omega)$ under H_3 by $D(a,w)$ and η, displayed using a divergent palette. $HR\,(\mathbf{p};1)$ is more powerful for strong evidence concentrated in a few tests and less powerful when weak evidence is spread among most tests.

in this region for every ω. The magnitude of this improvement is unclear, however, due to the difficulty comparing the contours between plots. Figure 4.7 facilitates the comparison by plotting the difference between the power contours of $HR(\mathbf{p};1)$ and $HR(\mathbf{p};\omega)$ directly for all settings. This more precise plot demonstrates that the most powerful $HR(\mathbf{p};\omega)$ to test H_3 against H_0 depends on the the strength and prevalence of evidence alone, *not* ω.

Specifically, Figure 4.7 shows that $HR(\mathbf{p};1)$ is more powerful than the correctly-specified $HR(\mathbf{p};\omega)$ in the lower right corner of the $D(a,\omega)$, η space and does worse left of centre at the top for both $\omega = e^{-6}$ and e^{-3}. In the final panel where $\omega = 1$, $HR(\mathbf{p};1) = HR(\mathbf{p};\omega)$ and so the difference in their powers is zero everywhere. Recalling the interpretation of these regions for these facets, this indicates $HR(\mathbf{p};1)$ is more powerful than $HR(\mathbf{p};\omega)$ at testing H_3 against H_0 when strong evidence is concentrated in a few tests, but is less powerful when evidence is weaker and spread widely. This pattern holds for every ω, albeit with differences in magnitude. Additionally, it is not symmetric: when $HR(\mathbf{p};1)$ is more powerful, the magnitude of the difference tends to be greater than in regions where it is less powerful.

This parallels Loughin (2004), who found $Fis(\mathbf{p})$ more powerful than other alternatives against H_3 when strong evidence was concentrated in a few tests. As $l_{Fis}(\mathbf{p}) \propto l_{HR}(\mathbf{p};1) \implies Fis(\mathbf{p}) = HR(\mathbf{p};1)$, this means that Fisher's method is once again relatively powerful for the same setting among the UMP family of tests $HR(\mathbf{p};\omega)$. The consistency of this result in both simulation studies warrants further investigation. To aid in this, marginal and central rejection levels are introduced to capture the tendency of a test to reject weak evidence spread among all tests and strong evidence concentrated in a few, along with a meaningful quotient that combines the them.

4.5 Central and marginal rejection levels in pooled p-values

Recall that any pooled p-value $g(\mathbf{p})$ behaves like a univariate p-value, that is $g(\mathbf{p}) \in [0,1]$ and $g(\mathbf{p}) \sim U$ under H_0. It may be based on order statistics and use Equation (4.3),

$$ord(\mathbf{p};k) = \sum_{l=k}^{M} \binom{M}{l} p_{(k)}^{l} (1 - p_{(k)})^{M-l},$$

or it may be based on the quantile transformations of Equation (4.4),

$$g(\mathbf{p}) = 1 - F_M \left(\sum_{i=1}^{M} F^{-1}(1 - p_i) \right).$$

In either case, $g(\mathbf{p})$ must be non-decreasing in every argument if it is to create convex acceptance regions and therefore be admissible[8], continuous in every p_i if it is to have well-defined rejection boundaries, and symmetric in p_i if no margin is to be favoured. Given these common properties, concepts of marginal and central rejection can be defined in order to describe rejection in the cases of evidence against H_0 concentrated in one test and evidence against H_0 spread among all tests, respectively.

These correspond to the largest p-values for a given FWER α which still result in rejection in two separate cases. The first of these, the marginal level of the pooled p-value, is the largest value of the minimum p-value which still leads to rejection at level α. The second, the central level of the pooled p-value, is the largest value which all p-values can take simultaneously while still resulting in rejection.

4.5.1 Characterizing central behaviour

The simulations of Section 4.4 suggest a pooled p-value $g(\mathbf{p})$ which is powerful at rejecting weak evidence spread among all p-values will reject H_0 when all tests give relatively large p-values compared to α. Under the rejection rule $g(\mathbf{p}) \leq \alpha$, this is captured by the largest p_c such that $g(\mathbf{p}_c) = \alpha$ for $\mathbf{p}_c = (p_c, \ldots, p_c)^\mathsf{T}$. Explicitly:

Definition 1 (The central rejection level)**.** *For a pooled p-value $g(\mathbf{p})$, the central rejection level, p_c, is the largest p-value for which $g(\mathbf{p}) \leq \alpha$ when $\mathbf{p} = (p_c, \ldots, p_c)^\mathsf{T}$. That is*

$$p_c(g) = \sup \left\{ p \in [0,1] : g(p, p, \ldots, p) \leq \alpha \right\} \tag{4.8}$$

p_c quantifies the maximum p-value shared by all tests which still leads to rejection and therefore is directly related to the power of $g(\mathbf{p})$ along $\mathbf{p}_c$. As $g(\mathbf{p})$ is non-decreasing, rejection occurs for any $\mathbf{p}$ in the hypercube $[0, p_c]^M$ and so a larger p_c implies rejection for a larger volume of $[0,1]^M$. It also suggests pooled p-values smaller than p_c within this hypercube if $g(\mathbf{p})$ is continuous and monotonic.

This general definition can be applied to the pooled p-values of Equations (4.3) and (4.4) to obtain simple expressions for p_c.

Proposition 2 (The central rejection level of order statistics)**.** *$p_c\big(ord\,(\mathbf{p}; k)\,\big)$ is given by the largest $p \in [0,1]$ which satisfies*

$$\sum_{l=k}^{M} \binom{M}{l} p^l (1-p)^{M-l} \leq \alpha. \tag{4.9}$$

[8] See Birnbaum (1954) and Owen (2009) for a detailed discussion of admissibility and convexity.

Proof. The definition of p_c forces $p_{(1)} = p_{(2)} = \cdots = p_{(M)} = p_c$. Therefore $p_{(k)} = p_c$ and so p_c is the largest $p \in [0,1]$ which satisfies $ord\left((p,\ldots,p);k\right) \leq \alpha$. Expanding $ord\left((p,\ldots,p);k\right)$ gives Equation (4.9). □

While this cannot generally be solved for a closed-form p_c, the particular cases of $ord\left(\mathbf{p};1\right) = Tip(\mathbf{p})$ and $ord\left(\mathbf{p};M\right)$ admit

$$p_c(Tip) = 1 - (1-\alpha)^{\frac{1}{M}} \tag{4.10}$$

and

$$p_c\big(ord\left(;M\right)\big) = \alpha^{\frac{1}{M}} \tag{4.11}$$

respectively. Also note that $p_c\big(ord\left(;k\right)\big)$ defines a constant rejection boundary around the regions of $[0,1]^M$ where $k-1$ p-values are less than or equal to p_c, $M-k$ elements are greater, and exactly one is equal to p_c. In particular, this means that $p_c(Tip) = p_c\big(ord\left(;1\right)\big)$ is constant along each margin, as $M-1$ points are greater than $p_c(Tip)$ along each margin. Next, consider p_c for the general quantile transformation of Equation 4.4.

Proposition 3 (The central rejection level of quantile pooled p-values)**.** *Given a pooled p-value based on quantile transformations as in Equation (4.4),*

$$g(\mathbf{p}) = 1 - F_M\left(\sum_{i=1}^{M} c_i F^{-1}(1-p_i)\right),$$

the central rejection level is given by

$$p_c\big(g(\mathbf{p})\big) = 1 - F\left(\frac{1}{\sum_{i=1}^{M} c_i} F_M^{-1}(1-\alpha)\right) \tag{4.12}$$

if F and F_M are continuous CDFs.

Proof. As F and F_M are CDFs , they are monotonically non-decreasing real-valued functions over their ranges. If F and F_M are also continuous, then F^{-1} and F_M^{-1} are continuous. Therefore, we can drop the supremum from Equation 4.8 and consider the equality

$$\alpha = g(p_c, p_c, \ldots, p_c).$$

Expanding g and solving for p_c implies

$$p_c = 1 - F\left(\frac{1}{\sum_{i=1}^{M} c_i} F_M^{-1}(1-\alpha)\right).$$

□

If the p-values are unweighted, then $c_1 = c_2 = \cdots = c_M = 1$, $\sum_{i=1}^{M} c_i = M$ and the behaviour of p_c depends on the relative growth of F_M^{-1} in M. If $F_M^{-1}(1-\alpha)$ grows in M such that $\frac{1}{M}F_M^{-1}(1-\alpha)$ is unbounded, p_c will go to zero. If, on the other hand, $\lim_{M\to\infty} \frac{1}{M}F_M^{-1}(1-\alpha) = c < \infty$, p_c will go to $1 - F(c)$. This provides a general expression for the soft truncation threshold of Zaykin et al. (2007) and suggests interesting asymptotic behaviour for pooled p-values based on quantile functions along the line $p_1 = p_2 = \cdots = p_M$.

This behaviour can be demonstrated concretely for several quantile transformations. Stouffer et al. (1949) takes $F = \Phi$ and $F_M = \sqrt{M}\Phi$[9] in Equation 4.4 to give $Sto(\mathbf{p})$ and so

$$\lim_{M\to\infty} \frac{1}{M} F_M^{-1}(1-\alpha) = \lim_{M\to\infty} \frac{1}{\sqrt{M}} \Phi^{-1}(1-\alpha) = 0$$

for all $\alpha > 0$. This implies

$$\lim_{M\to\infty} p_c(Sto) = 1 - \Phi(0) = \frac{1}{2}. \tag{4.13}$$

Indeed, taking $Sto(\mathbf{p})$, substituting $p_1 = \cdots = p_M = p_c(Sto)$, and taking the limit gives

$$\lim_{M\to\infty} 1 - \Phi\left(\frac{1}{\sqrt{M}} \sum_{i=1}^{M} \Phi^{-1}\big(1 - p_c(Sto)\big)\right) = 1 - \Phi\left(\lim_{M\to\infty} \sqrt{M}\Phi^{-1}\big(1 - p_c(Sto)\big)\right).$$

Evaluating further gives

$$\lim_{M\to\infty} \sqrt{M}\Phi^{-1}(1-p_c) = \begin{cases} -\infty & \text{when } p_c > \frac{1}{2} \\ 0 & \text{when } p_c = \frac{1}{2} \\ \infty & \text{when } p_c < \frac{1}{2}, \end{cases}$$

and so $Sto(\mathbf{p})$ is either 0, $\frac{1}{2}$, or 1 for large M when $p_1 \approx p_2 \approx \cdots \approx p_M$. Furthermore, it will reject H_0 for *any* FWER level α if $p_1, p_2, \ldots, p_M$ are all less than $\frac{1}{2}$ when M is large enough.

$F_\chi(x;\kappa)$ admits similar analysis. By the central limit theorem

$$\lim_{\kappa\to\infty} \chi^2_\kappa \to Z \sim N(\kappa, 2\kappa)$$

where $N(\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2. Therefore, as $M \to \infty$

$$F_\chi(x; M\kappa) \to \Phi\left(\frac{x - M\kappa}{\sqrt{2M\kappa}}\right).$$

[9] Though it uses scaling in computation to avoid this, the distribution is the same.

This implies that the pooled p-value based on the χ^2 quantile has the limiting value

$$\lim_{M\to\infty} 1 - F_\chi\Big(MF_\chi^{-1}\big(1-p_c;\kappa\big);M\kappa\Big) = 1 - \lim_{M\to\infty}\Phi\left(\frac{MF_\chi^{-1}\big(1-p_c;\kappa\big)-M\kappa}{\sqrt{2M\kappa}}\right)$$

when $p_1 = \cdots = p_M = p_c$. As Φ is absolutely continuous the limit can be taken inside the argument to give

$$1-\Phi\left(\lim_{M\to\infty}\sqrt{\frac{M}{2\kappa}}\left[F_\chi^{-1}(1-p_c;\kappa)-\kappa\right]\right).$$

Now,

$$\lim_{M\to\infty}\sqrt{\frac{M}{2\kappa}}\left[F_\chi^{-1}(1-p_c;\kappa)-\kappa\right] = \begin{cases}-\infty & \text{when } p_c > 1-F_\chi(\kappa;\kappa)\\ 0 & \text{when } p_c = 1-F_\chi(\kappa;\kappa)\\ \infty & \text{when } p_c < 1-F_\chi(\kappa;\kappa),\end{cases}$$

and so the pooled p-value based on the χ^2 quantile transformation behaves similarly to $Sto(\mathbf{p})$. It is asymptotically either 1, $\frac{1}{2}$, or 0 when $p_1 \approx p_2 \approx \cdots \approx p_M$ depending on whether all are greater than, equal to, or less than $1-F_\chi(\kappa;\kappa)$. Additionally, this implies that

$$p_c = 1-F_\chi(\kappa;\kappa) \tag{4.14}$$

for the χ^2 quantile case. So, although p_c depends on κ, all χ^2 quantile pooled p-values have identical behaviour about their respective p_c.

The form of Equation (4.4) suggests that this result applies generally. Suppose the random variable with CDF F has a mean μ and variance σ^2. Under H_0, $F^{-1}(1-p_i)$ for $i = 1,\dots,M$ are independent and identically-distributed realizations of this random variable and so $\sum_{i=1}^M F^{-1}(1-p_i)$ is asymptotically normally distributed with mean $M\mu$ and variance $M\sigma^2$ by the central limit theorem. Therefore $F_M \xrightarrow[M\to\infty]{} \sqrt{M}\sigma\Phi + M\mu$ for the pooled p-value based on F and the harsh asymptotic boundary at p_c derived for Sto will occur for *any* evidential statistic that uses quantile functions.

4.5.2 Characterizing marginal behaviour

Besides the central behaviour of a pooled p-value $g(\mathbf{p})$, the simulations of Section 4.4 indicate large differences in power occur for the rejection rule $g(\mathbf{p}) \leq \alpha$ when strong evidence exists in a single test. This is captured by the *marginal rejection level at b*.

Definition 2 (The marginal rejection level at b)**.** *For a symmetric pooled p-value $g(\mathbf{p})$, the marginal rejection level at b, $p_r(g; b)$, is the largest individual p-value in $[0, b]$ for which $g(\mathbf{p}) \leq \alpha$ when all other p-values are $b \in [0, 1]$. Without loss of generality, define*

$$p_r(g; b) = \sup \left\{ p_1 \in [0, b] : g(p_1, b, \dots, b) \leq \alpha \right\}. \tag{4.15}$$

In particular, the marginal value when $b = 1$ is of interest, that is when there is minimal evidence against all hypotheses other than H_{01}. Therefore, also define

$$p_r(g) = \lim_{b \to 1} \sup \left\{ p_1 \in [0, b] : g(p_1, b, \dots, b) \leq \alpha \right\}. \tag{4.16}$$

Note that symmetry is only necessary to avoid defining marginal rejection levels for each index $i \in \{1, \dots, M\}$ separately and that the term *marginal rejection level* refers to Equation (4.16). If g is non-decreasing in all of its arguments, $p_r(g; b)$ gives the largest value of $p_{(1)}$ that still leads to rejection at α when the evidence in all other p-values is bounded at b. The most extreme version of this measure is given by $p_r(g)$. By taking $b = 1$, it measures the power of g for evidence in a single test when all other tests provide no evidence against H_0, and so the sensitivity of g to evidence in a single test. This leads to a key lemma for $ord\,(\mathbf{p}; 1) = Tip(\mathbf{p})$.

Lemma 1 (The marginal rejection level for the minimum statistic)**.** *The marginal rejection level for g_{Tip} has two cases:*

$$p_r(Tip; b) = \begin{cases} b & \text{for } b < 1 - (1 - \alpha)^{\frac{1}{M}} \\ 1 - (1 - \alpha)^{\frac{1}{M}} & \text{for } b \geq 1 - (1 - \alpha)^{\frac{1}{M}}. \end{cases}$$

Proof. Recall that

$$Tip(\mathbf{p}) = 1 - (1 - p_{(1)})^M$$

is a function of the minimum alone. Rejection occurs when $Tip(\mathbf{p}) \leq \alpha$, or rather when $p_{(1)} \leq 1 - (1 - \alpha)^{\frac{1}{M}}$. When $b < 1 - (1 - \alpha)^{\frac{1}{M}}$ and $p_1 \leq b$, all values are below the rejection threshold and so $p_{(1)}$ attains its upper bound. Therefore

$$p_r(Tip; b) = \sup \left\{ p_1 \in [0, b] : g(p_1, b, \dots, b) \leq \alpha \right\} = b.$$

When $b \geq 1 - (1 - \alpha)^{\frac{1}{M}}$, rejection will only occur if $p_{(1)}$ is below the rejection threshold at α, and so

$$p_r(Tip; b) = \sup \left\{ p_1 \in [0, b] : g(p_1, b, \dots, b) \leq \alpha \right\} = 1 - (1 - \alpha)^{\frac{1}{M}}.$$

□

A direct consequence of Lemma 1 is that $p_r(g_{Tip}) = p_c(g_{Tip})$, which Theorem 1 proves is uniquely true for $Tip(\mathbf{p})$.

As with p_c, a larger p_r indicates greater power in a particular region of the unit hypercube. While p_c defines the rejection cube $[0, p_c]^M$, p_r defines the rejection shell $\{\mathbf{p} \in [0,1]^M : p_{(1)} \leq p_r\}$ with a flat boundary at p_r along each margin. A larger p_r implies a larger shell and therefore a greater volume of $[0,1]^M$ where H_0 is rejected and smaller pooled p-values within this volume if $g(\mathbf{p})$ is monotonic. Again, general expressions are provided for p_r for the order statistic and quantile transformation pooled p-values.

Proposition 4 (The marginal rejection level for order statistics)**.** *For $k \geq 2$, $p_r\big(ord\,(;k)\,, b\big) = b$ when $\sum_{l=k}^{M} \binom{M}{l} b^l (1-b)^{M-1} \leq \alpha$ and does not exist otherwise.*

Proof. Recall that

$$g_{Ord}(\mathbf{p}; k) = \sum_{l=k}^{M} \binom{M}{l} p_{(k)}^l (1 - p_{(k)})^{M-l}$$

and note Equation 4.15 forces $p_{(k)} = b$ for all $k > 1$. If $\sum_{l=k}^{M} \binom{M}{l} b^l (1-b)^{M-1} \leq \alpha$, then the supremum of p_1 is b. On the other hand, if $\sum_{l=k}^{M} \binom{M}{l} b^l (1-b)^{M-1} \geq \alpha$ there is no value of p_1 which leads to rejection and so $p_r\big(ord\,(;k)\,, b\big)$ does not exist. □

In particular, this implies that $p_r(ord\,(;k))$ does not exist for $k \geq 2$, in other words the pooled p-value based on $p_{(k)}$ has a value independent of $p_{(1)}$ for $k \geq 2$. So long as k tests are less than a particular bound, $ord\,(\mathbf{p}; k)$ will reject. If fewer than k are below that bound, the values of these small p-values are irrelevant.

Proposition 5 (The marginal rejection level of quantile transformation statistics)**.** *Given an unweighted evidential statistic based on quantile transformations as in Equation 4.4,*

$$g(\mathbf{p}) = 1 - F_M \left(\sum_{i=1}^{M} F^{-1}(1 - p_i) \right),$$

if F_M and F are both continuous then

$$p_r(g; b) = 1 - F\Big(F_M^{-1}(1-\alpha) - [M-1]F^{-1}(1-b)\Big).$$

Further, if both are absolutely continuous

$$p_r(g) = 1 - F\left(F_M^{-1}(1-\alpha) - [M-1] \lim_{x \to 0+} F^{-1}(x)\right) \tag{4.17}$$

Proof. Substituting Equation 4.4 into Equation 4.15 gives

$$p_r(g;b) = \sup\left\{p : 1 - F_M\Big(F^{-1}(1-p) + [M-1]F^{-1}(1-b)\Big) \leq \alpha\right\}.$$

As both F_M and F are CDFs, they are non-decreasing. If they are also continuous, their inverses exist and the supremum can be dropped to give

$$p_r(g;b) = 1 - F\Big(F_M^{-1}(1-\alpha) - [M-1]F^{-1}(1-b)\Big).$$

If they are both absolutely continuous, then the limit

$$1 - \lim_{b\to 1} F\Big(F_M^{-1}(1-\alpha) - (M-1)F^{-1}(1-b)\Big)$$

can be taken into the argument of F to give Equation 4.17. □

Many proposals use absolutely continuous CDFs, so this can be readily applied. $Sto(\mathbf{p})$, for example, has

$$p_r(Sto) = 1 - \Phi\left(\sqrt{M}\Phi^{-1}(1-\alpha) - [M-1]\lim_{x\to 0+}\Phi^{-1}(x)\right) = 0$$

for any α and M as $\lim_{x\to 0}\Phi(x) = -\infty$. Similarly, the proposal by Mudholkar and George (1977) has $p_r = 0$, as it uses the logistic distribution which also has $\lim_{x\to 0} F(x) = -\infty$. This suggests that, for a large enough p-value on all remaining tests, no level of evidence in a single test will cause the rejection of H_0 for either of these pooled p-values; their marginal rejection levels are always 0 for large enough b.

4.5.3 The centrality quotient

Beyond providing definitions that clarify the power of a pooled p-value to detect evidence spread among all tests and evidence in a single test, p_c and p_r as defined in Equations (4.16) and (4.8) can be combined into a single value summarizing the relative preference for diffuse or concentrated evidence. First, a key relationship between p_c and p_r is proven.

Theorem 1 (Order of p_c and p_r)**.** *For a pooled p-value $g(\mathbf{p})$ that is continuous, symmetric, and monotonically non-decreasing in all arguments, $p_c \geq p_r$ if both exist. Furthermore, equality occurs iff $g(\mathbf{p})$ is constant in p_k for $p_k \neq p_{(1)}$, that is if $g(\mathbf{p}) = f(p_{(1)})$ is a function of the minimum p-value alone.*

Proof. Consider $p_c(g)$ as in Definition 4.8. Then

$$p_c(g) = \sup\big\{p \in [0,1] : g(p,\dots,p) \leq \alpha\big\}.$$

Suppose

$$p_r(g) = \lim_{b \to 1} \sup\big\{p_1 \in [0,b] : g(p_1, b, \dots, b) \leq \alpha\big\}$$

exists. If g is symmetric, $p_r(g)$ captures the marginal rejection level of g in all margins. If g is continuous, then both $p_c(g)$ and $p_r(g)$ lie on the α level surface of g. Therefore

$$g(p_c,\dots,p_c) = \alpha = g(p_r, 1, \dots, 1).$$

But g is non-decreasing in all of its arguments, so

$$g(p_c, 1, \dots, 1) \geq g(p_c, p_c, \dots, p_c) = g(p_r, 1, \dots, 1)$$

and therefore

$$p_c \geq p_r.$$

If $p_c = p_r$, then substitute

$$\alpha = g(p_c, \dots, p_c) = g(p_r, \dots, p_r).$$

As g is non-decreasing

$$g(p_r, \dots, p_r) \leq g(p_r, 1, \dots, 1) = \alpha$$

and so

$$g(p_r, 1, \dots, 1) = g(p_r, p_r, \dots, p_r).$$

This implies that the average slope of g over $[p_r, 1]$, equivalently $[p_c, 1]$, is zero for all $p_k \neq p_1$. As g is continuous and non-decreasing, this implies that the slope must be zero for every point in this interval for all $p_k \neq p_1$. As $p_k \geq p_1$ for all $p_k \in \mathbf{p}$ over this region, $p_1 = p_{(1)}$ by definition. By the symmetry of g, the same argument holds for every p_k. Therefore $g(\mathbf{p}) = f(p_{(1)})$ for some non-decreasing function f.

To prove the reverse direction note that if $g(\mathbf{p}) = f(p_{(1)})$, then

$$\alpha = g(p_c, p_c, \dots, p_c) = g(p_c, 1, \dots, 1)$$

and so $p_c = p_r$ by the definition of p_r and the continuity of g. By symmetry, this same argument holds for any margin. □

Two facts follow directly from this proof. First, Theorem 1 implies that $p_c = p_r$ only for $Tip(\mathbf{p}) = ord\,(\mathbf{p}; 1)$ among symmetric, continuous, monotonically non-decreasing p-values as $Tip(\mathbf{p})$ is the unique pooled p-value defined by $p_{(1)}$. A second corollary is the existence of a sensible *centrality quotient* to quantify the balance between central and marginal rejection levels in pooled p-values.

Definition 3 (The centrality quotient)**.** *Suppose g is a continuous, symmetric, and monotonically non-decreasing pooled p-value for which $p_r(g)$ and $p_c(g)$ defined as in Equations (4.16) and (4.8) exist, define the centrality quotient*

$$q(g) = \frac{p_c(g) - p_r(g)}{p_c(g)}. \tag{4.18}$$

Theorem 1 implies that $q(g) \in [0, 1]$ with meaningful bounds. If $q(g) = 0$, $g(\mathbf{p})$ will reject based on the smallest p-values alone, increasing the marginal rejection level as large as possible while staying non-decreasing. Moreover, $q(g) = 0$ implies $g(\mathbf{p})$ is the pooled p-value based on $p_{(1)}$ alone, $Tip(\mathbf{p})$. In contrast, when $q(g) = 1$, g cannot reject based on the evidence contained in a single test, instead it requires evidence in many or all tests, for example $Sto(\mathbf{p})$. Between these extremes, pooled p-values with larger centrality quotients will reject H_0 for a larger range of p_c values and a smaller range of p_r values, and so will be more powerful at detecting evidence spread broadly at the cost of power when evidence is concentrated in a small number of p-values.

Indeed, increasing w decreases the centrality quotient of $HR\,(\mathbf{p}; w)$. This matches the empirical results obtained in Section 4.4.4 and in particular Figure 4.4.4, where larger w values provided greater power when the prevalence of evidence was small but the strength of evidence was large and smaller w values gave greater power in the case of weak evidence with high prevalence. For $\alpha = 0.05$, the centrality quotients of a range of w values in $HR\,(\mathbf{p}; w)$ are compared to those of several quantile transformation proposals in Table 4.1 over a range of M values. As predicted by the asymptotic argument at the end of Section 4.5.1, every method tends towards a centrality of 1 as M increases and each F_M converges to a corresponding normal CDF.

Beyond $HR\,(\mathbf{p}; w)$, other methods show the same relationship between the centrality quotient and regions of relative power in the empirical explorations in Westberg (1985), Loughin (2004), and Kocak (2017). Pooled p-values with larger centrality quotients are more powerful for weak evidence spread among all tests than those with smaller centrality quotients, but are relatively weak against strong evidence concentrated in a few tests. This is suggestive of an inverse relationship between p_r and p_c over different pooled p-values, but this is not the case generally. Consider, as a counter-example, $Sto(\mathbf{p})$ and the proposal of Mudholkar and George (1977): both have $p_r = 0$ but different values of p_c.

	M			
Pooled p-value	2	5	10	20
Tippett (1931)	0	0	0	0
Cinar and Viechtbauer (2022)	0.83	0.99	1.00	1.00
Stouffer et al. (1949)	1	1	1	1
Fisher (1932)	0.91	1.00	1.00	1.00
Mudholkar and George (1977)	1	1	1	1
Wilson (2019)	0.49	0.79	0.90	0.95
$HR\,(\mathbf{p}; e^{-6})$	1.00	1.00	1.00	1.00
$HR\,(\mathbf{p}; e^{-3})$	1.00	1.00	1.00	1.00
$HR\,(\mathbf{p}; 1)$	0.91	1.00	1.00	1.00

Table 4.1: Centrality quotients for certain pooled p-values.

4.6 Controlling the centrality quotient

Table 4.1, and others which could be constructed like it, provide only a limited ability to select a centrality quotient. Most of the proposals have centrality near 1, and all proposals approach 1 as M increases. Rather than choose among these other limited proposals when power is desired in a particular region, this work proposes a family of quantile pooled p-values based on χ^2_κ which precisely controls the centrality for any M. Following Equation (4.4), define the χ^2_κ quantile pooled p-value

$$\mathit{chi}\,(\mathbf{p}; \kappa) = 1 - F_\chi\left(\sum_{i=1}^{M} F_\chi^{-1}(1 - p_i; \kappa); M\kappa\right) \tag{4.19}$$

where $\kappa \in [0, \infty)$ is the the degrees of freedom of the quantile transformation applied to the p_i and doubles as a centrality parameter that sets $q\big(\mathit{chi}\,(;\kappa)\big)$ arbitrarily.[10] This family of pooled p-values includes several widely-used previous proposals. Setting $\kappa = 2$ gives $Fis(\mathbf{p})$, $\kappa = 1$ gives the proposal from Cinar and Viechtbauer (2022), taking $\lim_{\kappa\to\infty} \mathit{chi}\,(\mathbf{p}; \kappa)$ gives $Sto(\mathbf{p})$, and taking $\lim_{\kappa\to 0} \mathit{chi}\,(\mathbf{p}; \kappa)$ gives $Tip(\mathbf{p})$. While the former two are by definition, the latter must be proven. First, we prove $\lim_{\kappa\to\infty} \mathit{chi}\,(\mathbf{p}; \kappa) = Sto(\mathbf{p})$ by applying the central limit theorem.

[10] This is similar to the gamma method of Zaykin et al. (2007), but with a different parameter choice. It is possible the same control of c may be obtained with a general gamma CDF, but sticking to the χ^2_κ simplifies the number of parameters from two to one.

Theorem 2 (Limiting value of $chi\,(\mathbf{p};\kappa)$ as $\kappa \to \infty$)**.**

$$\lim_{\kappa\to\infty} chi\,(\mathbf{p};\kappa) = Sto(\mathbf{p})$$

Proof. Note that Equation (4.19) is always a pooled p-value, i.e. has a uniform distribution for any choice of κ. By the CLT, $\lim_{\kappa\to\infty} F_\chi(x;\kappa) = \Phi(x)$, and so in the limit $chi\,(\mathbf{p};\kappa)$ becomes the pooled p-value derived from the sum of standard normal quantile transformations, $Sto(\mathbf{p})$. □

The proof for $chi\,(\mathbf{p};0)$ is slightly more involved, and relies on Theorem 1.

Theorem 3 (Limiting value of $chi\,(\mathbf{p};\kappa)$ for $\kappa = 0$)**.**

$$\lim_{\kappa\to 0} chi\,(\mathbf{p};\kappa) = Tip(\mathbf{p}) = ord\,(\mathbf{p};1)$$

Proof. Theorem 1 proves that $p_r = p_c$ for a pooled p-value if and only if that pooled p-value is $Tip = ord\,(;1)$. Therefore, the limit is proven if

$$\lim_{\kappa\to 0} p_r\big(chi\,(;\kappa)\big) = \lim_{\kappa\to 0} p_c\big(chi\,(;\kappa)\big)$$

Expressing these quantities as probability statements gives

$$p_c\big(chi\,(;\kappa)\big) = P\left(\chi^2_\kappa \geq \frac{1}{M} F_\chi^{-1}(1-\alpha; M\kappa)\right),$$

and

$$p_r\big(chi\,(;\kappa)\big) = P\left(\chi^2_\kappa \geq F_\chi^{-1}(1-\alpha; M\kappa)\right).$$

The case of χ^2_0 is a degenerate distribution at 0. That is

$$F_{\chi^2}(x;0) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0. \end{cases}$$

This can also be seen from the limit of Markov's inequality for the χ^2_κ distribution,

$$P(\chi^2_\kappa \geq a) \leq \frac{\kappa}{a},$$

which goes to zero for any $a > 0$ as $\kappa \to 0$. As $F_\chi(x;\kappa)$ is continuous and monotonically increasing for all κ, this also implies

$$F\left(\frac{\kappa}{\alpha};\kappa\right) \geq 1-\alpha$$

$$\implies 1 - F\left(\frac{\kappa}{\alpha}; \kappa\right) \leq \alpha$$
$$\implies P\left(\chi^2_\kappa \geq \frac{\kappa}{\alpha}\right) \leq \alpha$$
$$\implies F_\chi^{-1}(1-\alpha;\kappa) \leq \frac{\kappa}{\alpha}.$$

This bound is not particularly tight, for $\alpha = 0.05$ it only restricts the 0.95 quantile to be less than 20 times the mean. However, it suffices to evaluate

$$\lim_{\kappa\to 0} \left| \frac{1}{M} F_\chi^{-1}(1-\alpha; M\kappa) - F_\chi^{-1}(1-\alpha; M\kappa) \right|$$

$$= \frac{M-1}{M} \lim_{\kappa\to 0} F_\chi^{-1}(1-\alpha; M\kappa)$$

$$\leq \frac{M-1}{M} \lim_{\kappa\to 0} \frac{M\kappa}{\alpha} = 0$$

and therefore

$$\lim_{\kappa\to 0} \frac{1}{M} F_\chi^{-1}(1-\alpha; M\kappa) = \lim_{\kappa\to 0} F_\chi^{-1}(1-\alpha; M\kappa)$$

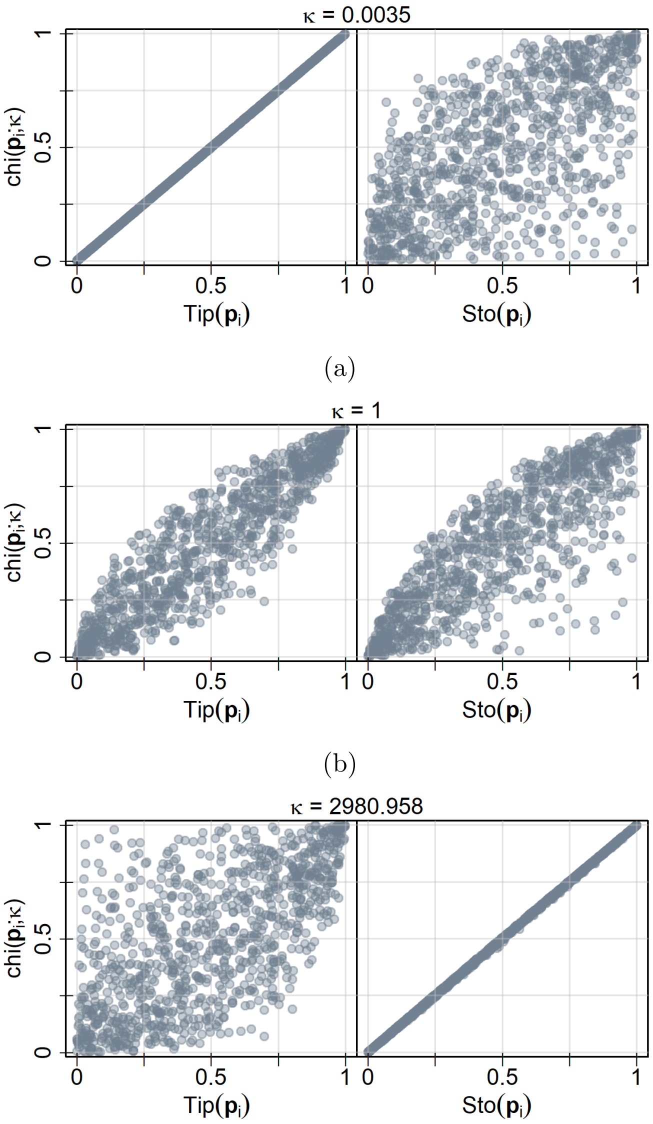
for any $\alpha > 0$. This implies that $p_c\big(chi\,(;\kappa)\big) = p_r\big(chi\,(;\kappa)\big)$ in the limit $\kappa \to 0$. □

The result of Theorem 3 can be understood intuitively using the non-central χ^2_0 of Siegel (1979) with a non-centrality parameter λ, call it $\chi^2_0(\lambda)$. When $\lambda \to 0$, $\chi^2_0(\lambda) \to \chi^2_0$ in distribution, so taking $\chi^2_0(\lambda)$ with small λ should provide some sense of how χ^2_0 behaves. Unlike χ^2_0, however, $\chi^2_0(\lambda)$ has a discrete probability mass at 0 for all $\lambda > 0$. As a result, the quantile function of $\chi^2_0(\lambda)$, F_λ^{-1}, returns zero for any input less than $e^{-\frac{\lambda}{2}}$ and so the terms in the sum

$$\sum_{i=1}^{M} F_\lambda^{-1}(1-p_i)$$

are non-zero only for those i where $p_i \leq 1 - e^{-\frac{\lambda}{2}}$. As $\lambda \to 0$, this sum becomes arbitrarily close to $chi\,(;0)$ but only the smallest p-values contribute. Eventually, only the minimum contributes to the sum, and so $chi\,(;\kappa) \approx f(p_{(1)})$ for very small κ values.

The limits $\lim_{\kappa\to 0} chi\,(\mathbf{p};\kappa) = Tip(\mathbf{p})$ and $\lim_{\kappa\to\infty} chi\,(\mathbf{p};\kappa) = Sto(\mathbf{p})$ are also demonstrated empirically by generating n_{sim} independent realizations of $\mathbf{p}$ assuming H_0 is true. For each vector $\mathbf{p}_i$, compute $chi\,(\mathbf{p}_i;\kappa)$ for a range of κ, $Sto(\mathbf{p}_i)$, and $Tip(\mathbf{p}_i)$ and compare



(a)

(b)

(c)

Figure 4.8: A comparison of $chi\left(\mathbf{p}_i;\kappa\right)$, $Sto(\mathbf{p}_i)$, and $Tip(\mathbf{p}_i)$ values for 1000 independently generated $\mathbf{p}_i \sim Unif([0,1]^5)$ in the case of (a) small κ, (b) moderate κ, and (c) large κ.

$chi\,(\mathbf{p}_i;\kappa)$ to the other two pooled p-values. Figure 4.8 shows this pattern for a few κ when $M = 5$.

As expected, the agreement between $chi\,(\mathbf{p};\kappa)$ and $Tip(\mathbf{p})$ is perfect for small enough κ, the two functions have identical outputs for $\kappa = 0.0035$ in Figure 4.8(a). Similarly, $chi\,(\mathbf{p};\kappa)$ and $Sto(\mathbf{p})$ match for large κ, as when $\kappa \approx 3000$ in Figure 4.8(c). Note that the particular values of κ where this close agreement occurs will depend on M.

Perhaps more interesting is the curved boundary of the points along the top of the plot of $chi\,(\mathbf{p}; 0.0035)$ against $Sto(\mathbf{p})$ in Figure 4.8(a), many points populate the lower right corner of this plot but there are none in the upper left. This pattern is mirrored in Figure 4.8(c) for the plot of $chi\,(\mathbf{p}; 2981)$ against $Tip(\mathbf{p})$. As $chi\,(\mathbf{p};\kappa)$ is essentially identical to one of $Tip(\mathbf{p})$ or $Sto(\mathbf{p})$ in these cases, this pattern reflects the relationship between $Tip(\mathbf{p})$ and $Sto(\mathbf{p})$. By definition, $Tip(\mathbf{p})$ considers only $p_{(1)}$, but any of the p_i can impact $Sto(\mathbf{p})$. As a result there will be many cases where a small $Tip(\mathbf{p})$ occurs despite a large $Sto(\mathbf{p})$ because a very small $p_{(1)}$ happens by chance. The reverse is impossible, if $Tip(\mathbf{p})$ is large then $p_{(1)}$ is large and therefore all values in $\mathbf{p}$ are large, suggesting a large $Sto(\mathbf{p})$.

4.6.1 Choosing a parameter

In addition to these meaningful limits, there seems to be a monotonically increasing relationship between κ and $q\big(chi\,(;\kappa)\big)$. Let $\chi^*_\kappa(\alpha)$ be the $1-\alpha$ quantile of the χ^2 distribution with κ degrees of freedom, then $chi\,(\mathbf{p};\kappa)$ has the central rejection level

$$p_c\big(chi\,(;\kappa)\big) = 1 - F_\chi\left(\frac{1}{M}F_\chi^{-1}(1-\alpha; M\kappa);\kappa\right) = P\left(\chi^2_\kappa \geq \frac{1}{M}\chi^*_{M\kappa}(\alpha)\right) \tag{4.20}$$

and the marginal rejection level

$$p_r\big(chi\,(;\kappa)\big) = 1 - F_\chi\Big(F_\chi^{-1}(1-\alpha; M\kappa);\kappa\Big) = P\Big(\chi^2_\kappa \geq \chi^*_{M\kappa}(\alpha)\Big), \tag{4.21}$$

implying

$$\begin{aligned} q\big(chi\,(;\kappa)\big) &= \frac{p_c\big(chi\,(;\kappa)\big) - p_r\big(chi\,(;\kappa)\big)}{p_c\big(chi\,(;\kappa)\big)} \\ &= P\left(\chi^2_\kappa \leq \chi^*_{M\kappa}(\alpha) \,\middle|\, \chi^2_\kappa \geq \frac{1}{M}\chi^*_{M\kappa}(\alpha)\right). \end{aligned} \tag{4.22}$$

That is, the centrality quotient of $chi\,(\mathbf{p};\kappa)$ is the conditional probability that a χ^2_κ random variable is less than $\chi^*_{M\kappa}(\alpha)$ given that it is greater than $\frac{1}{M}\chi^*_{M\kappa}(\alpha)$.

A better sense of the region corresponding to this conditional probability for $\alpha < 0.5$ is garnered by writing $\chi^*_{M\kappa}$ in terms of the mean of the $\chi^2_{M\kappa}$ distribution, $M\kappa$. Taking an arbitrary remainder function $R_{M\kappa}(\alpha) > 0$ such that $\chi^*_{M\kappa}(\alpha) := M\kappa + R_{M\kappa}(\alpha)$, substitution gives

$$q\big(chi\,(;\kappa)\big) = P\left(\chi^2_\kappa \leq M\kappa + R_{M\kappa}(\alpha)\,\middle|\,\chi^2_\kappa \geq \kappa + \frac{1}{M}R_{M\kappa}(\alpha)\right),$$

clarifying that $q\big(chi\,(;\kappa)\big)$ is a conditional probability on the right tail of the χ^2_κ distribution when $\alpha < 0.5$. Making more precise statements about $R_{M\kappa}(\alpha)$ is challenging due to the small values of κ which may be chosen for $chi\,(;\kappa)$. Most approximations of χ^2 tail probabilities and quantiles either break down when the degrees of freedom is less than one or explicitly assume more than one degrees of freedom (Hawkins and Wixley, 1986; Canal, 2005; Inglot, 2010). Nonetheless, the above probability can be computed numerically, as was done for the curves of $q\big(chi\,(;\kappa)\big)$ by $\log_{10}(\kappa)$ for M ranging from 2 to 10,000 in Figure 4.9.

The curves of $q\big(chi\,(;\kappa)\big)$ by κ have a consistent sigmoid shape for all M. Most of the change in the centrality quotient occurs for values in a three unit range in $\log_{10}(\kappa)$ for any M, though the centre of this range decreases as M grows. When $\kappa = 10^{-3}$, for example, the centrality quotient when $M = 100$ is greater than 0.8 while the same κ value corresponds with a centrality quotient of less than 0.05 when $M = 2$. Just as with any other pooled p-value, increasing M increases the centrality of $chi\,(;\kappa)$ for a given κ as the sum of independent p-values becomes more normally distributed by the central limit theorem.[11]

In practice, the inverse of the above curves may be of greater interest to control the centrality quotient under $chi\,(\mathbf{p};\kappa)$ rather than simply report it. Figure 4.9 does allow the selection of κ for a given centrality quotient by estimating the κ value where the intersection between a curve and a vertical line at κ is at the desired quotient, but a table displaying the numerically estimated inverse for evenly-spaced κ as in Table 4.2 is more precise and straightforward to use. Determining the desired $\log_{10}(\kappa)$ for a given centrality quotient and M proceeds as for a table of critical values. The user searches down the columns for the M most closely corresponding to the setting at hand, and then searches through that

[11] This can also be understood geometrically. For a pooled p-value in M dimensions, the volume of the marginal shell of width p_r is $1 - (1 - p_r)^M$, which approaches 1 for any $p_r > 0$ as $M \to \infty$. As the total volume of the rejection region is α for the rejection rule $g(\mathbf{p}) \leq \alpha$, p_r must decrease in M to hold the volume constant.

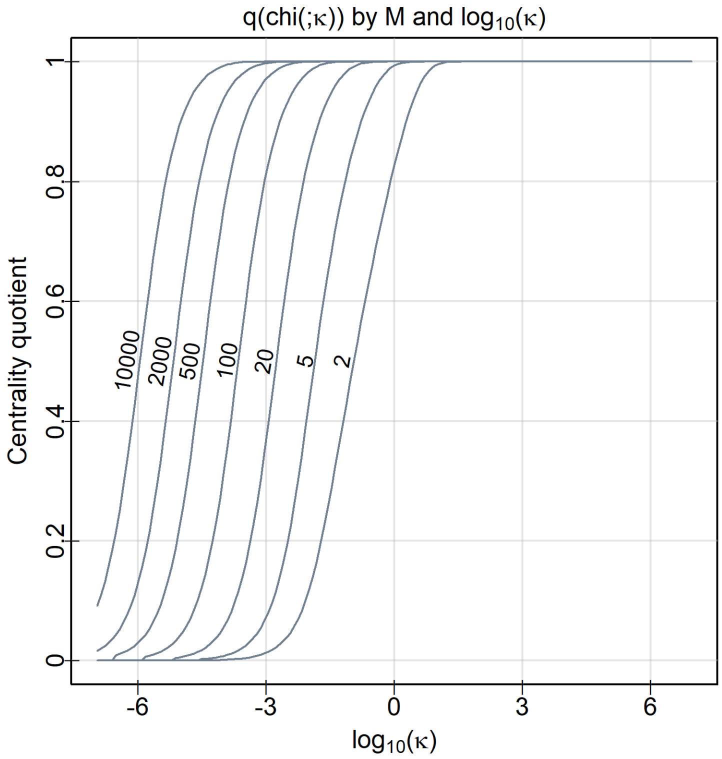


Figure 4.9: The centrality quotient of $chi\,(;\kappa)$ by $\log_{10}(\kappa)$

row for the desired column. If $q\big(chi\,(;\kappa)\big) = 1$ or 0 is desired, the table is unnecessary as $Sto(\mathbf{p})$ or $Tip(\mathbf{p})$ can be used directly. Unlike with critical value tables, there is no need to be conservative: linear interpolation between the provided $\log_{10}(\kappa)$ values is a reasonable approach to choosing κ.

Using the parameter κ of $chi\,(\mathbf{p};\kappa)$, the relative preference of $chi\,(\mathbf{p};\kappa)$ to rejection along the margins or in the centre can be directly controlled. Large κ produce a pooled p-value which is powerful at detecting evidence spread among all tests, while small κ favour the detection of concentrated evidence in a single test with extremes giving the widely-used $Tip(\mathbf{p})$ and $Sto(\mathbf{p})$. The parameter κ orders pooled p-values of the $chi\,(\mathbf{p};\kappa)$ family by relative centrality, simplifying the choice of pooled p-value and communication of results. Finally, as it is based on Equation (4.4), it is an exact quantile-based method which does not rely on asymptotic behaviour and which could, hypothetically, be computed by hand with the aid of χ^2 quantile tables.

	Centrality quotient								
M	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	-2.1	-1.7	-1.4	-1.2	-0.9	-0.7	-0.4	-0.1	0.3
5	-2.8	-2.5	-2.3	-2.1	-1.9	-1.7	-1.4	-1.2	-0.8
20	-3.7	-3.4	-3.1	-2.9	-2.8	-2.6	-2.4	-2.1	-1.8
100	-4.6	-4.3	-4	-3.8	-3.7	-3.5	-3.3	-3	-2.7
500	-5.4	-5.1	-4.8	-4.7	-4.5	-4.3	-4.1	-3.9	-3.5
2000	-6.1	-5.8	-5.5	-5.3	-5.2	-5	-4.8	-4.6	-4.2
10000	-6.9	-6.6	-6.3	-6.1	-6	-5.8	-5.6	-5.4	-5

Table 4.2: $\log_{10}(\kappa)$ values by $q\big(chi\,(;\kappa)\big)$ and M to aid in parameter selection for the desired balance of central and marginal rejection.

4.6.2 Comparing the chi-squared pooled p-value to the UMP benchmark

Recall the simulation studies that motivated the exploration of central and marginal rejection levels. After a benchmark power computation, the power of $HR\,(\mathbf{p};w)$ for $\alpha = 0.05$ was evaluated under H_4 for a range of beta alternatives $\left(f = Beta(a, 1/\omega + a(1-1/\omega))\right)$ with KL divergences from uniform $(D(a,\omega))$ spanning e^{-5} to e^5. Correct specification of w was important: the larger the magnitude of $w-\omega$, the larger the decrease in power of $HR\,(\mathbf{p};w)$ from $HR\,(\mathbf{p};\omega)$. Under H_3, mis-specification did not matter at all, the power of $HR\,(\mathbf{p};w)$ was dictated by the the proportion of false null hypotheses (η) and the strength of evidence against H_0 in each non-null hypothesis $(D(a,\omega))$. The parameter w tunes $HR\,(\mathbf{p};w)$ to favour either weak evidence spread among all tests, or strong evidence in only a few.

Finer selection of this tradeoff is achieved with the parameter κ using the $chi\,(\mathbf{p};\kappa)$ family of pooled p-values, but controlling centrality is of little use if $chi\,(\mathbf{p};\kappa)$ is not powerful under the settings that motivated their definition. The power of $chi\,(\mathbf{p};\kappa)$ at level $\alpha = 0.05$ for each $\kappa \in \{e^{-8}, e^{-4}, 1, 2, e^4, e^8\}$ was therefore determined under every setting from Section 4.4 using the same simulated samples generated under H_4. Prior to the simulation, it is expected is that large κ will be uniformly more powerful than small κ, as under H_4 evidence is spread among all tests. The results confirmed this expectation: the most powerful $chi\,(\mathbf{p};\kappa)$ for all settings under H_4 was $chi\,(\mathbf{p};e^8) \approx chi\,(\mathbf{p};2981)$. It is compared to the both the UMP and mis-specified $HR\,(\mathbf{p};w)$ in Figure 4.10 by adding a dark grey line to Figure 4.4.

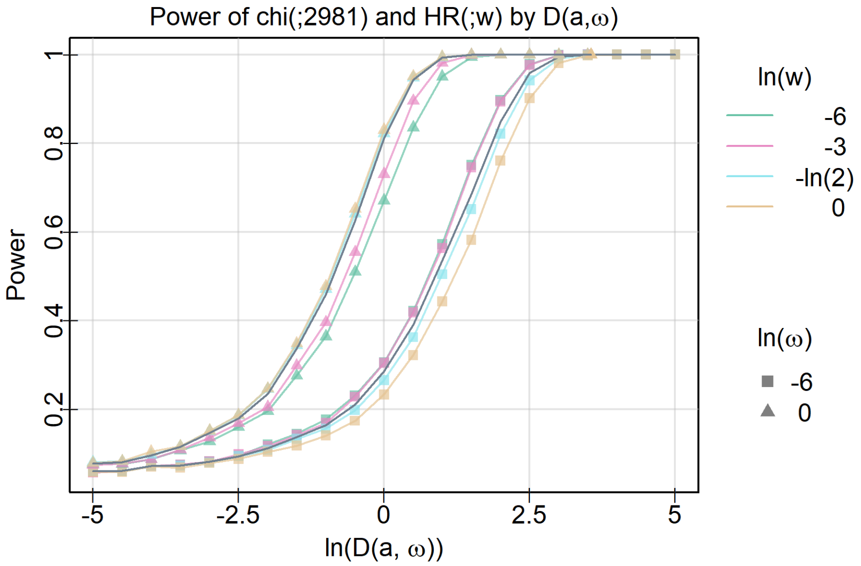
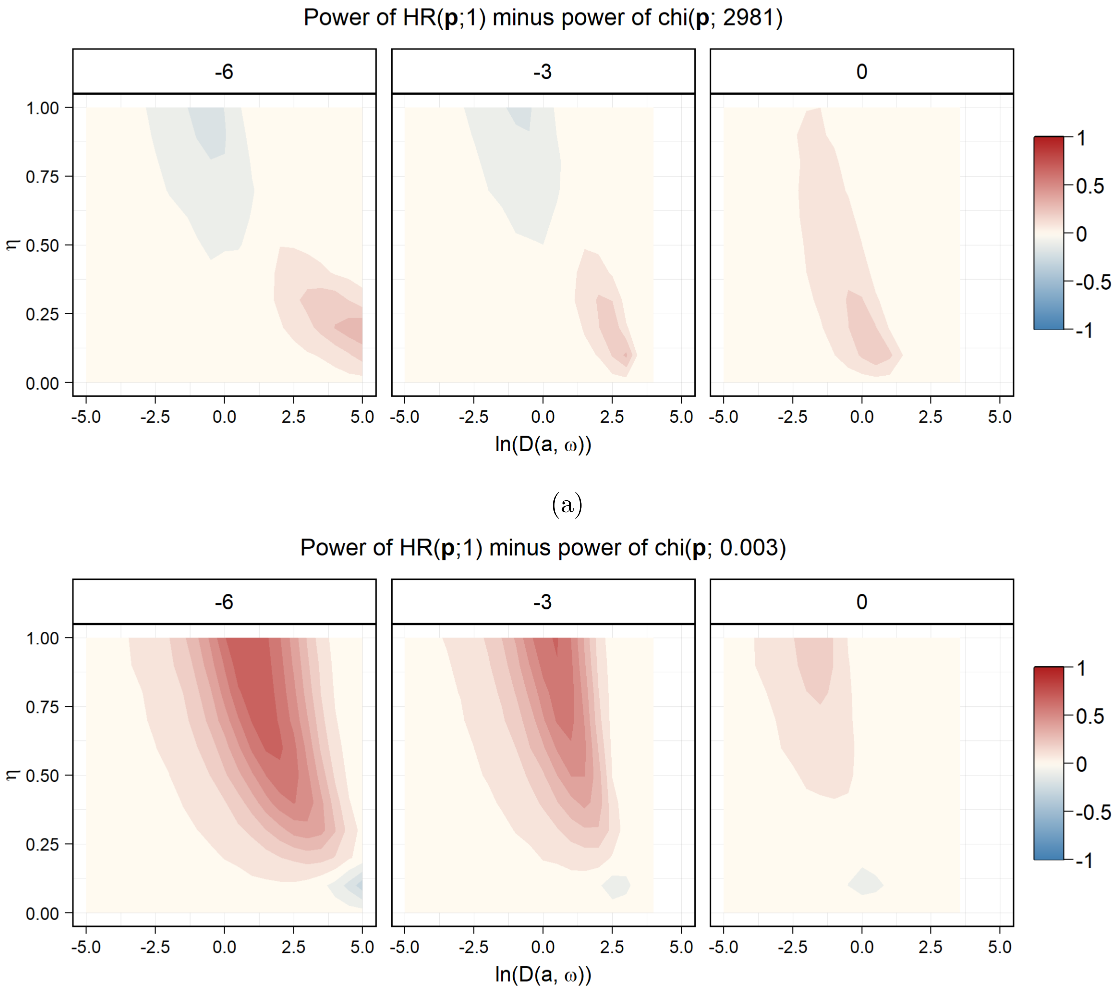


Figure 4.10: A comparison of $chi\,(\mathbf{p};2981)$ to the UMP and mis-specified $HR\,(\mathbf{p};w)$ under H_4. $chi\,(\mathbf{p};2981)$ nearly UMP power more consistently than any $HR\,(\mathbf{p};w)$, and so is more robust to f.

The pooled p-value $chi\,(\mathbf{p};2981)$ has higher power than most mis-specified $HR\,(\mathbf{p};w)$ for all settings in this case and is close to the UMP more consistently than any $HR\,(\mathbf{p};w)$. Only when $w \approx \omega$ does $HR\,(\mathbf{p};w)$ beat $chi\,(\mathbf{p};2981)$, and so it is less robust to mis-specification of f than $chi\,(\mathbf{p};2981)$. It may therefore be advisable to use $chi\,(\mathbf{p};\kappa)$ with a large κ (or simply $Sto(\mathbf{p})$) when testing H_4 with beta alternatives in the case where ω is not known, rather than risk the penalty of choosing w wrong when using $HR\,(\mathbf{p};w)$. This is despite the fact that $HR\,(\mathbf{p};\omega)$ is UMP for this setting.

For the case where $\mathbf{p}$ was generated under H_3, $chi\,(\mathbf{p};\kappa)$ was again computed for each $\kappa \in \{e^{-8}, e^{-4}, 1, 2, e^4, e^8\}$ over the 10,000 independent samples for each setting of $D(a,\omega)$, ω, and η with $M = 10$ from Section 4.4. Contour plots analogous to Figure 4.7 showing the differences in power between $chi\,(\mathbf{p};\kappa)$ and $HR\,(\mathbf{p};1) = chi\,(\mathbf{p};2)$ were generated. The reference $HR\,(\mathbf{p};1)$ was chosen because it is a test shared by both the $chi\,(\mathbf{p};\kappa)$ and $HR\,(\mathbf{p};w)$ families.

The patterns of power for $chi\,(\mathbf{p};\kappa)$ mimic those of $HR\,(\mathbf{p};w)$: large κ favour evidence spread among all tests as do small w in $HR\,(\mathbf{p};w)$. Despite this similarity, $chi\,(\mathbf{p};2981)$ has higher power when applied to the case of concentrated evidence and so is more robust



(a)

(b)

Figure 4.11: Contours for the power of $HR\,(\mathbf{p};1) = chi\,(\mathbf{p};2)$ minus (a) $chi\,(\mathbf{p};e^{8}) \approx chi\,(\mathbf{p};2981)$ and (b) $chi\,(\mathbf{p};e^{-8}) \approx chi\,(\mathbf{p};0.003)$ by η and $D(a,\omega)$ facetted by ω. Compared to Figure 4.7, (a) displays less of a penalty for the case of concentrated evidence while still outperforming $HR\,(\mathbf{p};1)$ for evidence spread among all tests.

under H_3. This is seen clearly in a comparison of the bottom right corner of Figure 4.7 to Figure 4.11(a), the former shows a much larger and darker red region than the latter.

The $chi\,(\mathbf{p};\kappa)$ family also extends the range of possible centrality parameters compared to $HR\,(\mathbf{p}; w)$. As $HR\,(\mathbf{p}; 1) = chi\,(\mathbf{p}; 2)$ is one of the boundaries of the w parameter range, no comparable pooled p-values to $chi\,(\mathbf{p}; \kappa)$ for $\kappa < 2$ exist in the $HR\,(\mathbf{p}; w)$ family. Using $chi\,(\mathbf{p}; \kappa)$ therefore gives greater control over the balance of central and marginal rejection than $HR\,(\mathbf{p}; w)$, though it seems exceptionally small κ in $chi\,(\mathbf{p}; \kappa)$, or equivalently $Tip(\mathbf{p})$, should only be used sparingly. Figure 4.11(b) shows that $chi\,(\mathbf{p}; 0.003)$ loses power almost everywhere compared to $chi\,(\mathbf{p}; 2)$ in exchange for higher power only in the case of extreme evidence in a single test. Under H_3, very small values of κ should probably only be used if such a pattern of evidence is strongly suspected.

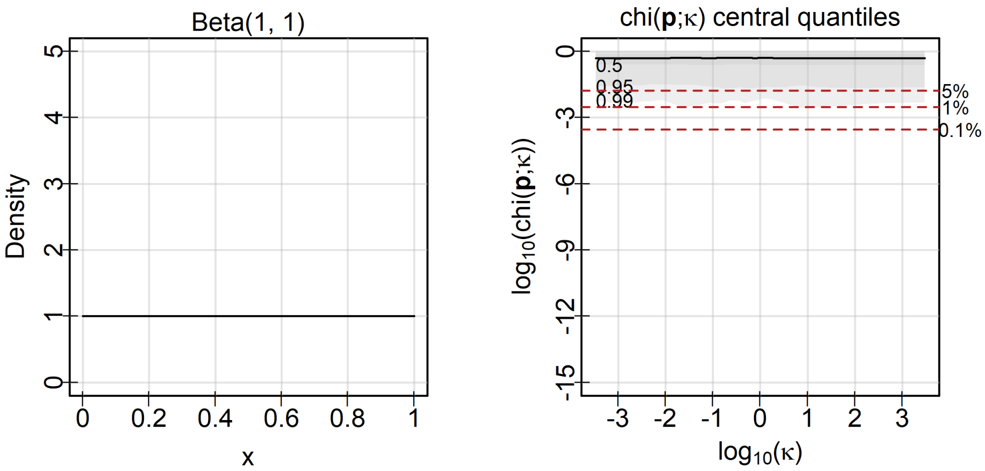
The $chi\,(\mathbf{p}; \kappa)$ family is therefore of interest both practically and theoretically. It provides control over central and marginal rejection under H_3 and robustly gives nearly UMP power for large values of κ under H_4. It has interpretable endpoints which cover a greater range of centrality quotients than $HR\,(\mathbf{p}; w)$ and gives a means of controlling the bias towards central rejection present in all quantile pooled p-values as M increases. $chi\,(\mathbf{p}; \kappa)$ is a pooled p-value with great potential as a practical tool for controlling the FWER when testing H_0.

4.7 Identifying plausible alternative hypotheses and selecting tests

The link between κ, the centrality quotient, and relative power in regions of the $D(a, w), \eta$ plane under H_3 can be exploited to identify alternatives to H_0 that could have plausibly generated $\mathbf{p}$. Rather than selecting a particular κ value, we can consider all possible κ values simultaneously, compute $chi\,(\mathbf{p}; \kappa)$ for each, and record

$$\kappa_{\min} = \underset{\kappa \in [0, \infty)}{\arg\min}\, chi\,(\mathbf{p}; \kappa) \tag{4.23}$$

As each κ value is associated with a particular centrality quotient, each κ identifies a particular region of relative power against others in the $D(a, w), \eta$ plane under H_3. At the same time, $\kappa_{\min}$ reports the value of κ which produces the smallest pooled p-value for $\mathbf{p}$ and therefore suggests the κ value where evidence against H_0 is the strongest relative to other κ values. As stronger evidence leads to more frequent rejection and higher power when H_0 is false, $\kappa_{\min}$ therefore links the evidence present in $\mathbf{p}$ to a region in $D(a, w), \eta$ if we assume H_3 is truly used to generate the data with $f = Beta\big(a, 1/w + a(1 - 1/w)\big)$.



(a)

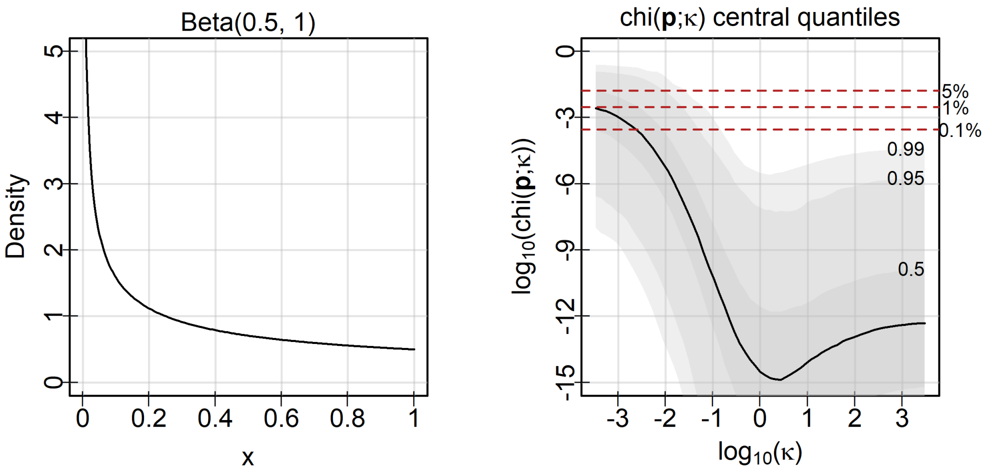
(b)

Figure 4.12: The (a) density and (b) central quantiles and median by κ for the null case. All lines are flat and above the null quantiles from the larger simulation as expected.

4.7.1 Non-increasing beta densities

Begin with a demonstration of the sweep of κ values for previously explored cases by generating curves of $chi\,(;\kappa)$ by κ for different densities under H_4. In each of the following, samples of 100 i.i.d. p-values from different beta distributions are generated independently 1,000 times and $chi\,(;\kappa)$ is computed for a sequence of κ values chosen uniformly on the log scale. Let the i^{th} sample be $\mathbf{p}_i$ and the pooled p-value computed using parameter κ_j for $\mathbf{p}_i$ be $\chi_{ij} = chi\,(\mathbf{p}_i; \kappa_j)$. To provide context to χ_{ij}, a larger simulation of 100,000 samples was generated under H_0 and the minimum of $chi\,(\mathbf{p}_i; \kappa_j)$ for the same sequence of κ_j values was recorded. Figure 4.12(b) displays the median and 0.5, 0.95, and 0.99 central quantiles for $f = Beta(1, 1)$ (equivalent to the null case) alongside the $Beta(1, 1)$ density in 4.12(a). For reference three dashed red lines at the observed 0.05, 0.01, and 0.001 quantiles of the minimum pooled p-value over the 100,000 simulated null cases have been added.

This case shows a flat median curve and flat central quantiles which are all slightly above the corresponding minimum quantiles. The null case performs as expected, κ_{min} is distributed uniformly over the range of κ values and would produce values below the null quantiles at the expected proportions. A contrasting case in shown in Figure 4.13, which uses the same layout for an identical simulation carried out when $f = Beta(0.5, 1)$.

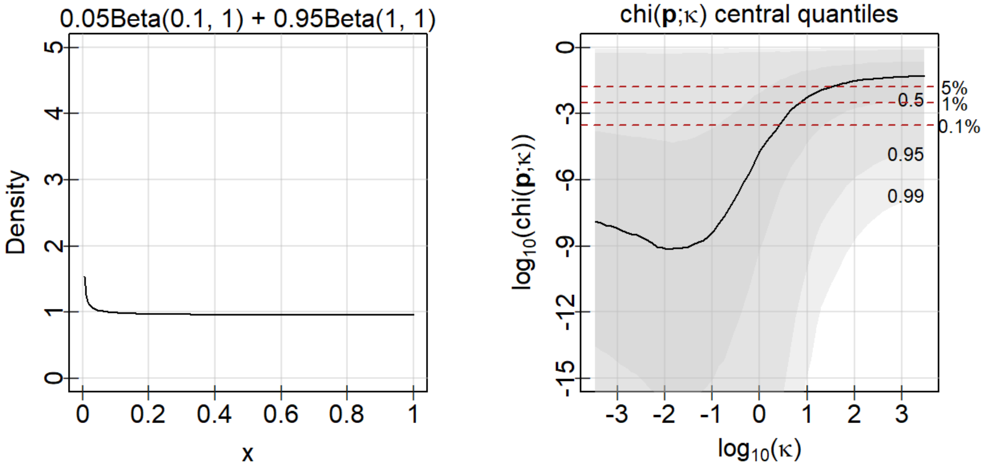


a) b)

Figure 4.13: The (a) density and (b) central quantiles and median by κ for p-values generated identically and independently from a Beta(0.5, 1) distribution. The minimum around $\kappa = 2$ corresponds to the UMP.

Displaying the median and the same central quantiles as before, there is a unique minimum at $\kappa = 2$, a lower right end to the curve than the left end, and a generally lower value across its entire length. If one of these curves was observed in practice, $\kappa_{\text{min}} \approx 2$ would be chosen and larger κ values may not be fully ruled out. This conclusion would be correct: under H_4 with $f = Beta(0.5, 1)$ the UMP pooled p-value is $HR\left(\mathbf{p}; w\right)$ with $w = (1 - a)/(b - a) = 1$ and $HR\left(\mathbf{p}; 1\right) = chi\left(\mathbf{p}; 2\right) = Fis(\mathbf{p})$. Furthermore, the power investigations in Section 4.6.2 demonstrate that $chi\left(\mathbf{p}; 2981\right) \approx Sto(\mathbf{p})$ is nearly as powerful as the UMP for all w under H_4. This confirms empirically that the level of the curve of $chi\left(\mathbf{p}; \kappa\right)$ over κ corresponds to the relative power of pooled p-values in $chi\left(\mathbf{p}; \kappa\right)$ for this beta distribution.

Of course, this conclusion should be expanded to H_3 and so mixture of $Beta(0.1, 1)$ and $Beta(1, 1)$ distributions is considered. The first distribution provides strong evidence against the null hypothesis while the second corresponds to the uniform distribution, and so contains no evidence against the null. Mixing these such that the probability of drawing from $Beta(0.1, 1)$ is 0.05 and the probability of drawing from the null is 0.95 we are placed in the $D(a, w), \eta$ space at $0.3, 0.05$. Simulating this as for the null and $Beta(0.5, 1)$ cases and displaying the central quantiles of χ_{ij} alongside the mixture density gives Figure 4.14.



(a) (b)

Figure 4.14: The (a) density and (b) central quantiles and median by κ for the mixture 0.05Beta(0.1, 1) + 0.95Beta(1, 1). Small κ values provide the smallest pooled p-values, and hence power at detecting this alternative.

The central quantiles are more variable for this case than the unmixed densities because the probability of generating any p-values from $Beta(0.1, 1)$ is small and so many samples would have included only null p-values. Nonetheless, the median has a unique minimum near $\kappa = 0.01$, and is generally lower for small κ than large κ. Considering the coordinates of this case in the $D(a, w), \eta$ plane, this is completely consistent with the earlier investigations of power where small κ values were most powerful for strong evidence concentrated in a few tests. In practice, seeing the median curve would cause us to suspect this case correctly.

4.7.2 Identifying a region of alternative hypotheses

While these one-to-one comparisons between densities and κ curves help to demonstrate the link between $\kappa_{\min}$ and the alternative densities used to generate $\mathbf{p}$, they are not incredibly informative. Given $\mathbf{p}$ and supposing we generate such a curve of $chi\,(\mathbf{p}; \kappa)$ by κ, we would need to sort through an incredible number of density-curve pairs to identify plausible alternatives corresponding to the curve obtained.

Instead, consider a more automated approach. Given a collection of p-values, this generates a curve by sweeping parameter values of κ, identifies the minimum κ values (or any below a particular threshold), and maps these back onto the plane of strength and prevalence depicted in, for example, Figure 4.11. This requires a detailed guide of where each κ value is most powerful in the $\eta, D(a,w)$ plane so that $\kappa_{\min}$ or the range of significant κ values can be placed accurately. Therefore, a simulation was carried out over 20 $\ln(w)$ values evenly spaced from -6 to 0, η values from 0 to 1 in increments of 1/80, and 80 $\ln D(a,w)$ values evenly spaced between -5 and 5. For each combination, 10,000 samples of 80 p-values were generated with 80η following the beta distribution specified by $\ln(w)$ and $\ln D(a,w)$ and $80(1-\eta)$ following the uniform distribution.[12]

Each of the 10,000 samples then had pooled p-values computed over a sweep of 65 $\ln \kappa_i$ values evenly spaced from -8 to 8 and the power was computed for the rejection rule $chi\left(\mathbf{p};\kappa_i\right) \leq 0.05$ (corresponding with a level $\alpha = 0.05$). The κ_i with the greatest power for each combination corresponds to $\kappa_{\min}$ for that combination because rejection is determined by thresholding the pooled p-value and so a higher power implies a lower distribution of the pooled p-value at a given point. Bivariate discretized Gaussian smoothing is applied to each $chi\left(\mathbf{p};\kappa_i\right)$ power surface in $\eta, D(a,w)$ for each w value in order to obtain a smoothed estimate of the power surface minimally impacted by random binomial noise. This was completed only because none of the investigations carried out indicated discontinuities in the power or distribution of p-values by κ_i.

For each w, the power surfaces of every $chi\left(\mathbf{p};\kappa_i\right)$ in $\eta, D(a,w)$ were then compared to the maximum among them point-wise. This is motivated by the simpler case shown in Figure 4.13, as several κ_i values are often equally powerful for a given setting. Specifically, the comparison was a binomial test of the difference in proportions using a normal approximation at 95% confidence. A surface was deemed equal to the maximum power at that point if the test failed to reject the null hypothesis of equal proportions.[13] For each κ and w, all of this pre-processing gave a matrix in η and $\ln D(a,w)$ indicating whether $chi\left(\mathbf{p};\kappa_i\right)$

[12] This resolution was not the only one tried, similar experiments were carried out for $M = 10$, 20, and 40 and the only impact of increasing M, the number of steps in η, and the number of steps in $\ln D(a,w)$ was increasing resolution of the same patterns. This suggests that these patterns do not depend on the sample size.

[13] For powers p_1 and p_2 computed over the same number of trials n, this computes

$$z = \frac{\sqrt{n}(p_1 - p_2)}{\sqrt{2p(1-p)}}$$

where $p = \frac{p_1+p_2}{2}$ and then compares z to normal critical values. By the CLT, $z \sim N(0,1)$ approximately for large n, and as 10,000 simulations are performed for each power estimate, this approximation should be quite accurate.

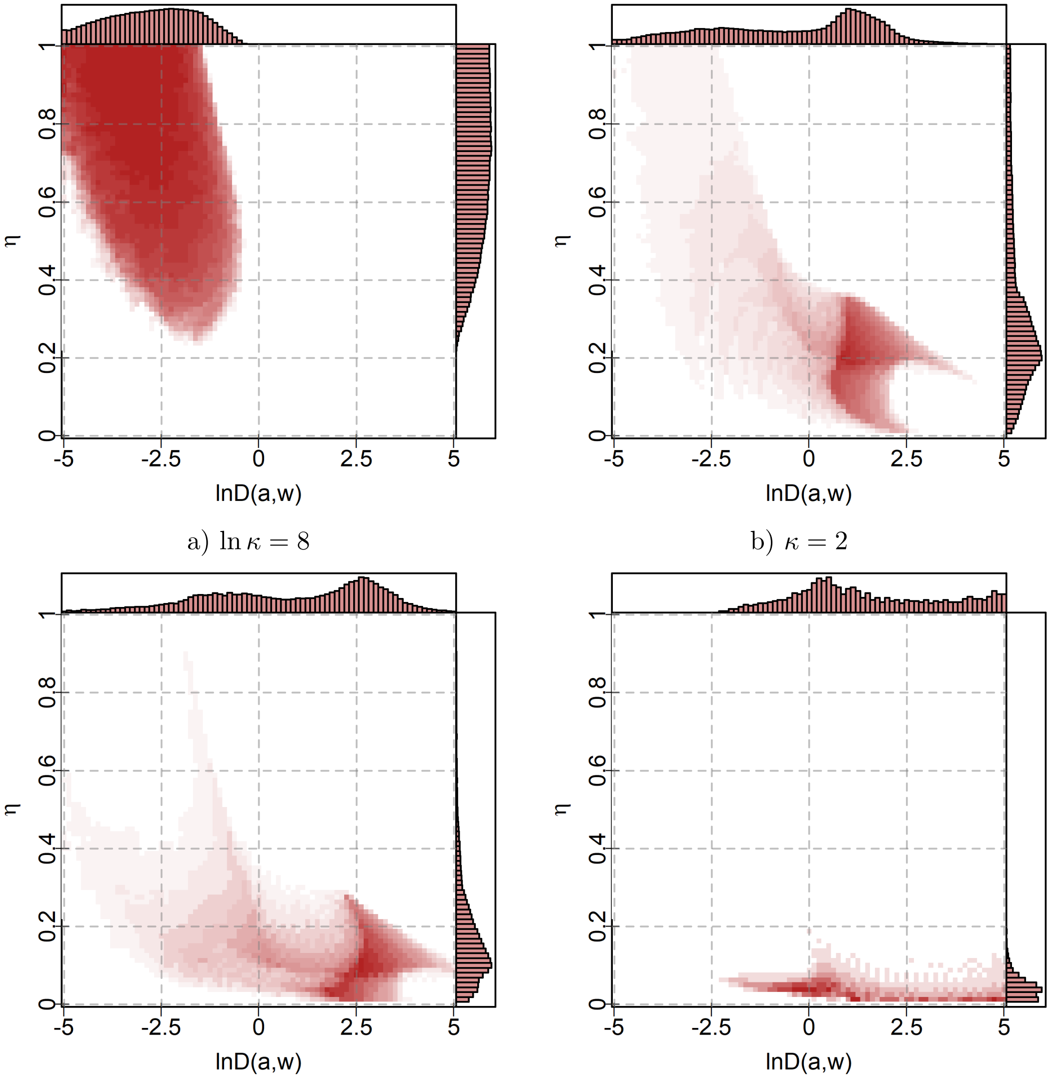
achieved the maximum power for that combination for every κ_i. To produce a final summary in η and $\ln D(a,w)$ alone, these indicators were summed over w for each κ_i. Finally, masks were added in the top right and bottom left corners where all methods are equally powerful with powers 1 and 0.05 respectively to make the meaningful patterns more visible. Figures 4.15(a)–(d) display these sums (counts of cases in w where κ_i achieved maximum power) for several κ_i in a given $\eta, \ln D(a,w)$ region, with guide histograms on each row and column added to quickly indicate the relative marginal frequencies. Each plot is therefore rich with both marginal and joint information on the regions where a particular κ_i is most powerful.

Consistent with previous investigations, this map shows that the regions where the small κ values are most powerful correspond to small η values. The mode of the histogram of η values increases steadily in κ until it is near one when $\kappa = e^8$. For $\kappa = e^{-4}$, the pooled p-value is only most powerful for settings with $\eta < 0.1$, such that a minimum of the parameter curve below $e^{-4} \approx 0.02$ indicates a small minority of tests are significant.

Given that these plots display counts of cases where a particular $\mathit{chi}\left(\mathbf{p};\kappa\right)$ is most powerful, choosing to select w evenly-spaced on the log scale inadvertently places greater weight on small values of w and under-samples large values in order to achieve more complete coverage of $D(a,w)$. Even for moderate w floating point representation limits prevent the computation of a and b for the beta distributions with large KL divergences. Choosing parameters to span strength is one of two possible perspectives, the other focuses on even exploration of the parameter space. For this parameter-based perspective, the exact same procedure was performed but with 20 w values selected at even increments from 0.05 to 1. Figures 4.16(a)–(d) present the same heatmaps as Figure 4.15 from this perspective.

There are some noteworthy differences between this and the first set of heatmaps. The bias towards smaller proportions and stronger evidence in the first is quite clear when it is compared to the second, which generally shows similar shapes but more evenly distributed saturation across this shape. This leads to changes in the regions suggested for a particular κ_{min}, but these are typically minor. The biggest difference occurs for large κ, where the bias towards small values in Figure 4.15 obscures all of the internal variation in the middle top that can be seen in Figure 4.16.

Without these plots, an analyst would be left trying to identify alternatives from a density estimate. Besides showing comparable information about the prevalence of evidence to a density estimate in the histogram along the right side of the plot, these plots of plausible alternatives give information about likely strengths of evidence and regions for the combination of both. By leveraging the links between the centrality quotient, κ, and



a) $\ln \kappa = 8$

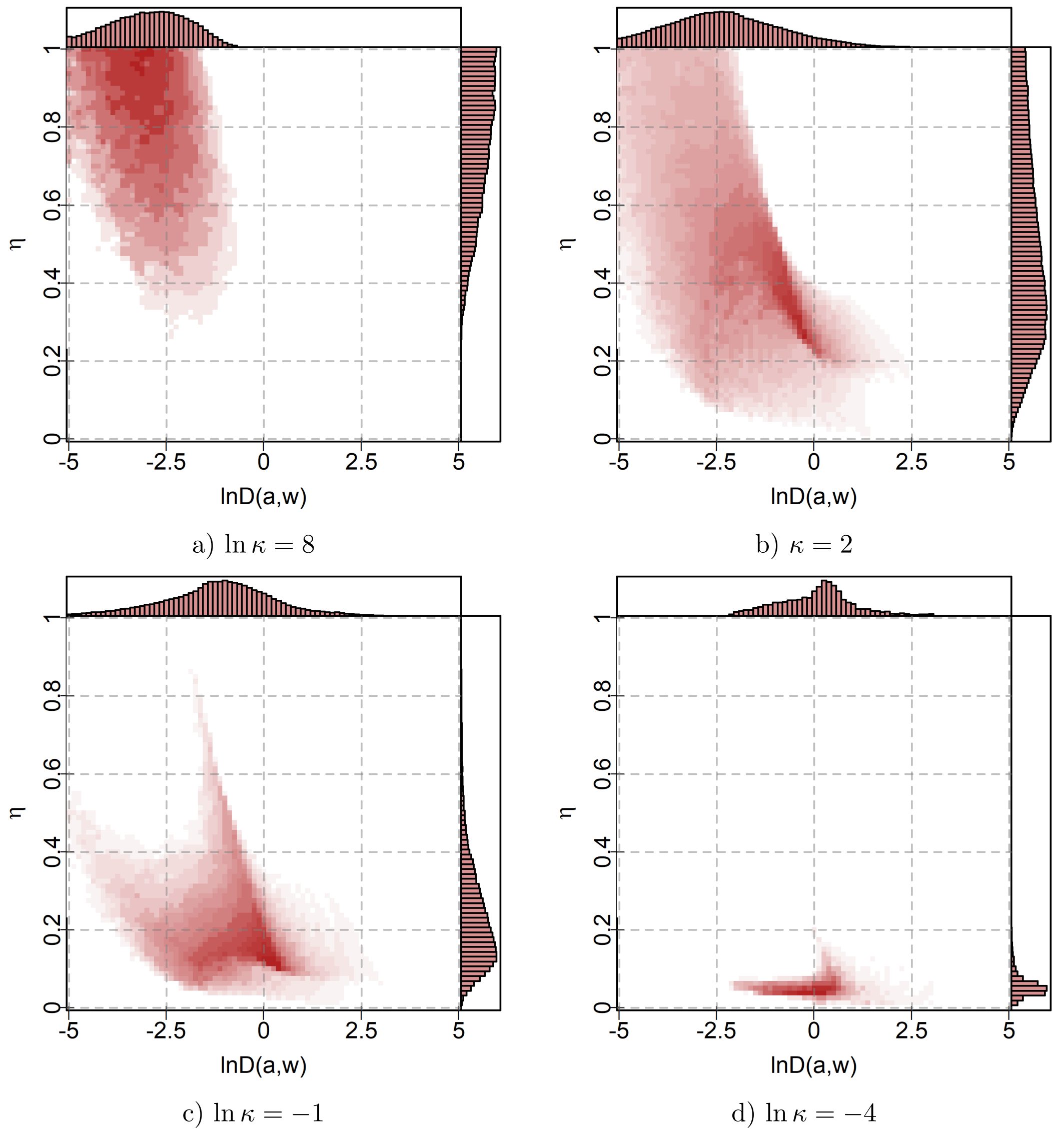
b) $\kappa = 2$

c) $\ln \kappa = -1$

d) $\ln \kappa = -4$

Figure 4.15: Likely alternatives for a range of κ values. For full coverage of $D(a, w)$, w was chosen uniformly on a log scale.

the distribution and strength of evidence in $\mathbf{p}$, these maps provide richer and clearer information.



a) $\ln \kappa = 8$

b) $\kappa = 2$

c) $\ln \kappa = -1$

d) $\ln \kappa = -4$

Figure 4.16: Likely alternatives for a range of κ values. w was chosen uniformly for these images, leading to worse coverage of $D(a, w)$ but more appropriate coverage of w.

4.7.3 Selecting a subset of tests

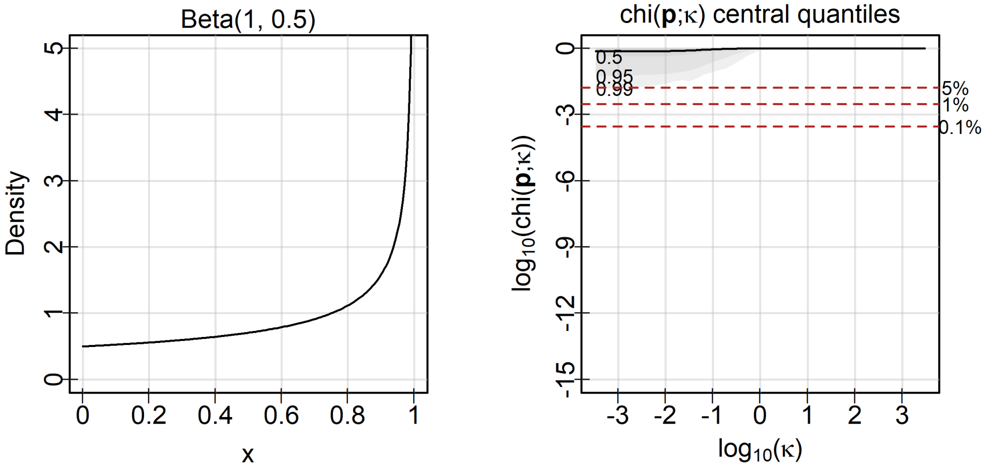
Perhaps the most important part of the alternative heatmaps presented in Figures 4.15 and 4.16 are the histograms along the right margin that indicate the likely prevalence of evidence in the data. Once $\kappa_{\min}$ has been determined using a sweep of κ values, and a plausible set of alternatives has been identified using these alternative heatmaps, the corresponding range of proportions can be used to identify a subset of tests of interest. If false positives are less problematic to analysis than false negatives, the upper bound of this range might be taken, with the other bound taken if the opposite is true. In either case, suppose the chosen proportion is η^*, then the $M\eta^*$ largest values of $F^{-1}(1 - p_i; \kappa_{\min})$ are the tests most contributing to the small value of $chi\,(\mathbf{p}; \kappa_{\min})$ and so are the tests of greatest interest that can be selected for further investigation.

4.7.4 Centrality in other beta densities

Until now, it was always assumed that p-values follow a non-increasing beta density when the null hypothesis is false. This is a reasonable assumption, many statistical tests have this property for the rejection rule thresholding the p-value at α. Relaxing this assumption, however, allows an exploration into how $chi\,(\mathbf{p}; \kappa)$ behaves for a broader variety of densities and whether centrality is still a useful concept under these other distributions of non-null p-values.

First, consider the case of a strictly increasing density under H_4. Whether or not this case is interesting is a matter of opinion. Under the convention that small p-values are evidence against H_0, such a density produces even less evidence against H_0 than the null distribution itself. It would be reasonable to expect, then, that this case produces only very large $chi\,(\mathbf{p}; \kappa)$ for all κ values. Following the same procedure as Section 4.7.1, this expectation is tested for $Beta(1, 0.5)$, resulting in the curve and density displayed in Figure 4.17.

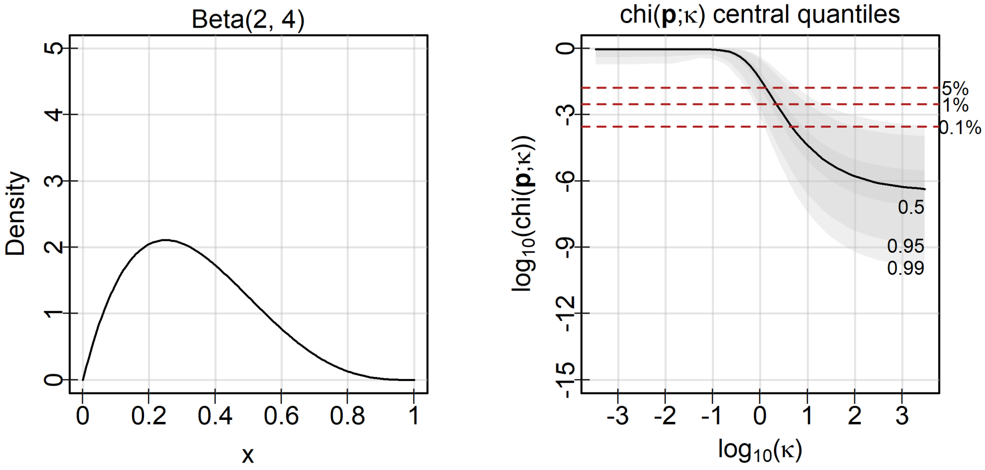
As expected, this setting gives only large $chi\,(\mathbf{p}; \kappa)$ values for every κ. There is a slight dip to small p-values for small κ, likely a result of the small p-values that still occur for this density occasionally, but it barely crosses the null reference lines. Perhaps a more realistic case is a density that rarely, if ever, produces minimum p-values small enough to warrant rejection alone, but does tend to produce far more p-values less than 0.5 than expected under the null hypothesis. For such a setting, the previous investigations into centrality suggest that $\kappa_{\min}$ should be large. An example is $f = Beta(2, 4)$ under H_4, shown in Figure 4.18.



(a) (b)

Figure 4.17: The (a) density and (b) central quantiles and median by κ for p-values from a Beta(1, 0.5) distribution. The pooled p-value is generally large compared to the empirical curve minimum quantiles.

Once again, this plot matches the expectation reasoned from centrality, despite the relaxation of the assumptions used to motivate centrality. Both of these examples suggest that the concepts of central and marginal rejection may have use beyond non-increasing densities, and provide a promising framework for future investigation.



(a)

(b)

Figure 4.18: The (a) density and (b) central quantiles and median of a p-values following a Beta(2, 4) distribution. The absence of very small p-values and bias towards smaller ones means large κ values are most powerful.

4.8 The PoolBal package

There is no lack of packages available to pool p-values in R. The most recent of these, **poolr** (Cinar and Viechtbauer, 2021), lists 8 others all providing piecemeal coverage of every pooled p-value proposal[14]. Rather than re-implement the functionality provided by these packages, the PoolBal package aims primarily to support the evaluation of the central rejection level, marginal rejection level, and centrality quotient for these and any future packages which pool p-values. As they both allow some tuning of centrality, these core functions are supported by implementations of $chi\left(\mathbf{p};\kappa\right)$ and $HR\left(\mathbf{p};\kappa\right)$ along with functions to evaluate the Kullback-Leibler divergence for general densities and compute it for the beta density in particular. This is meant to make the adoption of the framework provided in this work as simple as possible.

Briefly summarized, the functionality of PoolBal includes

klDiv, betaDiv: compute the Kullback-Leibler divergence for arbitrary densities and the uniform to Beta case, respectively

findA: invert a given Kullback-Leibler divergence and most powerful test parameter w to identify the unique Beta parameter a that corresponds to this setting

pBetaH4, pBetaH3: helpers to generate $\mathbf{p}$ under under H_4 and H_3

estimatePc, estimatePrb, estimateQ: wrappers for uniroot from base that estimate the central rejection level, marginal rejection level at b, and centrality quotient for an arbitrary function

chiPool: an implementation of $chi\left(\mathbf{p};\kappa\right)$

chiPc, chiPr, chiQ: functions to compute the central rejection level, marginal rejection level, and centrality quotient of $chi\left(\mathbf{p};\kappa\right)$ using Equations (4.20), (4.21), and (4.22)

chiKappa: a wrapper for uniroot from base that inverts a given centrality quotient to give the κ value in $chi\left(\mathbf{p};\kappa\right)$ with the corresponding quotient

hrStat, hrPool: compute $l_{HR}(\mathbf{p};w)$ and $HR\left(\mathbf{p};w\right)$ for $\mathbf{p}$, with the p-value determined empirically using simulated null data

[14] These are Dewey (2022) Zhang et al. (2020); Wilson (2019); Yi and Pachter (2018); Poole and Gibbs (2015); Dai et al. (2014); Schröder et al. (2011); Zhao (2008). Most of these packages cover a subset of pooling functions or implement adjustments for dependence rather than attempting to be the complete package for pooling p-values.

hrPc, hrPr, hrQ: functions to compute the central rejection level, marginal rejection level, and centrality quotient of $HR\,(\mathbf{p};\kappa)$ using simulation and `uniroot`

altFrequencyMat: function allowing access to a summarized version of the simulation results from Section 4.7.2

marHistHeatMap: function which generates heatmaps with marginal histograms, that is visualizations such as those in Figure 4.16.

The package can be found on the author's GitHub and CRAN (Salahub, 2023b).

4.9 Conclusion

When presented with M p-values from independent tests of hypotheses $H_{01}, \dots, H_{0M}$, a natural way to control the family-wise error rate (FWER) is by pooling these p-values using a function $g(\mathbf{p})$. If $g(\mathbf{p})$ is constructed using the sum of quantile transformations or the order statistics of the p-values, then the rejection rule $g(\mathbf{p}) \leq \alpha$ controls the FWER at α. Selecting between the many possible $g(\mathbf{p})$ requires the choice of an alternative hypothesis from the telescoping set $H_1 \supset H_2 \supset H_3 \supset H_4$ in order to determine their powers against these alternatives. H_3 and H_4, though the most restrictive, still require the choice of η, the prevalence of non-null p-values in $\mathbf{p}$, and f, the distribution of these non-null values. An obvious choice for f is the beta distribution restricted to be non-decreasing, as this biases non-null p-values lower than null p-values. By using η and the Kullback-Leibler divergence of f from the uniform distribution, both the prevalence and strength of non-null evidence can be measured.

If all the evidence is non-null, i.e. $\eta = 1$, the pooled p-value based on $l_w(\mathbf{p}) = w\sum_{i=1}^{M} \ln p_i - (1-w)\sum_{i=1}^{M} \ln(1-p_i)$ is uniformly most powerful (UMP) but is sensitive to the specification of its parameter $w \in [0,1]$. The power of $HR\,(\mathbf{p};w)$ to reject the alternative hypothesis is reduced when w does not match the true parameter ω under H_4. When $\eta \neq 1$, both the prevalence and strength of non-null evidence dictate the most powerful choice of w. Small values of w are more powerful for weak evidence spread among all tests while large values are better at detecting strong evidence in a few tests.

This reflects a more universal pattern in pooled p-values and motivates a new paradigm for selecting and analyzing them. The marginal level of rejection at α, the largest individual p-value that leads to rejection at α when all other p-values are 1, and the central rejection level at α, the largest value simultaneously taken by all elements of $\mathbf{p}$ which still leads to rejection at α, characterize this paradigm. By defining the central and marginal rejection level, a number of fundamental properties can be proven. Among them, the central rejection level of a pooled p-value satisfying some mild conditions is always greater than or equal to the marginal rejection level, with equality occurring only for $Tip(\mathbf{p})$. This order allows a centrality quotient to be defined which summarizes the preference of a pooled p-value to diffuse or concentrated evidence with a value in $[0, 1]$.

In order to control this quotient, a pooled p-value based on χ^2_κ quantile transformations was defined, $chi\,(\mathbf{p}; \kappa)$. By choosing the degrees of freedom $\kappa \in [0, \infty)$, arbitrary control over the centrality of the pooled p-value is obtained. Increasing κ raises the centrality quotient, and decreasing it drops the quotient. Furthermore, the limiting cases of $\kappa = 0$ and $\kappa \to \infty$ correspond to $Tip(\mathbf{p})$, the minimum order statistic p-value, and $Sto(\mathbf{p})$, the normal quantile transformation p-value. Both of these limiting pooled p-values are classic pooling functions which have been used and studied widely in the literature. The pooled p-value $chi\,(\mathbf{p}; \kappa)$ therefore provides a means to balance an important aspect of pooling p-values with a single parameter that has ready interpretation along its range. Comparing its power to $HR\,(\mathbf{p}; w)$ under H_3 and H_4, $chi\,(\mathbf{p}; \kappa)$ loses less power than $HR\,(\mathbf{p}; w)$ with w mis-specified. It is therefore more robust and demonstrates that the central and marginal rejection paradigm is instructive to predict which version of $chi\,(\mathbf{p}; \kappa)$ will be most powerful for a particular alternative hypothesis. In conclusion, $chi\,(\mathbf{p}; \kappa)$ and the centrality quotient are both potent tools for pooling p-values to control the FWER.

Chapter 5

A Simple Genetic Model

In this chapter, a toy genetic model is presented to provide a basic understanding of genetics. It is not meant to be comprehensive, nor to fully describe every aspect of the rich field of genomics. Instead, it has been included to contextualize this often-repeated setting for measuring association and multiple testing.

$$\mathbf{G} \xrightarrow{\textit{select}} \mathbf{S} \xrightarrow{\textit{annotate}} \mathbf{T} \xrightarrow{\textit{encode}} \mathbf{X} \xrightarrow{\textit{summarize}} \mathbf{z}$$

Figure 5.1: A model of genomic association studies.

Genetic research today routinely considers the entirety of a *genome*, defined by Doerge et al. (1997) as all heritable material potentially passed to offspring, to identify regions strongly related to physical traits. The goal is to associate measured genome sequences, the *genotype*, with physical characteristics, the *phenotype*. Traits can be *monogenic*, impacted by a single region of the genotype, or *oligogenic*, impacted by several regions together. Computational and methodological advances in the pursuit of these *quantitative trait loci* (QTLs) have distinguished *genomics* as its own field. Central to genomics is the *genome-wide association study* (GWAS), where many *markers*, sequences of nucleotides at known positions on the genome, are measured. The measurement of markers and their structure is critical in the search for QTLs.

Figure 5.1 draws on the literature to present a model outlining the conversion of raw marker measurements into data. The key steps of *selection*, *annotation*, *encoding*, and *summarization* are identified as maps between increasingly abstract representations of the genome, highlighted in plain language. While such a simple representation is no replacement for surveys such as Uffelmann et al. (2021) and Tam et al. (2019), it supplies a guiding framework suitable for the introduction of the topic to anyone with a mathematical background but very little biology experience.

The model starts with **G**, the whole genome of an organism. Genetic information is stored in DNA, a long molecule consisting of a sequence of four *nucleotide bases*: guanine, cytosine, adenine, and thymine. A *diplodic* individual inherits one version or *variant* of a complete DNA sequence from each parent, and so has two copies in all *somatic* (i.e. non-reproductive) cells. Though it can be represented as one long sequence, DNA is actually organized into *chromosomes*, separate strands of DNA which contain only a part of the sequence. As much of genetic research concerns diplodic species, this is implicitly assumed in the following.

It is generally not feasible to design a study around the measurement of all of **G**, and so the *select* step chooses regions to measure, represented in **S**. Often **S** consists of a series of *single nucleotide polymorphisms* (SNPs), single nucleotide substitutions in a known sequence at a known position. In human studies this is supported by SNP databases such as NCBI (2021) which document hundreds of millions of common SNPs in the human genome[1]. These are observed by means of arrays capable of identifying hundreds of thousands of SNPs simultaneously, see LaFramboise (2009); Tam et al. (2019). As no array simultaneously measures all SNPs, selection is necessary, motivated by previous findings and *linkage disequilibrium*, the correlation between markers at different regions of the genome that facilitates inference to regions outside of those selected in **S**. While third generation genome sequencing technologies allow for entire genomes to be sequenced, their persistent high costs and more than a decade of SNP array development leave arrays as the dominant measurement method (Heather and Chain, 2016; Hasin et al., 2017; Uffelmann et al., 2021).

After selecting SNPs to obtain **S**, the data must be *annotated*. The raw signal produced by a SNP array is fluorescence, with different degrees of fluorescence corresponding to different genotypes. Converting the fluorescent areas of an array to a genotype is a challenging problem and has developed in tandem with the arrays themselves (LaFramboise, 2009). Early models used non-parametric clustering techniques on the signal from

[1] Despite this massive database, only a fraction occur frequently enough to be used. Koboldt et al. (2013) identifies about 15 million SNPs common enough in humans to be useful

several array sections, but more complex hidden Markov and Bayesian models have also been developed. Whatever method is used, the selected regions are assigned genotypes in $\mathbf{T}$. Often these are denoted with capital or lowercase letters at each SNP, as in Siegmund and Yakir (2007) and Visscher and Goddard (2019).

Finally, relationships within $\mathbf{T}$ and between $\mathbf{T}$ and an observed trait are quantified by converting each annotated SNP to a number. We first *encode* each SNP variant with a numeric value and then *summarize* both variants at each location into a single value. Typically no distinction is made between these steps: the dominance and additive summaries move directly from an annotated genotype to a numeric value in Lander and Botstein (1989), Cheverud (2001), and Siegmund and Yakir (2007). They are separated here for clarity and full generality.

This section presents the details of this model, starting with an explanation of the model and all necessary notation in Section 5.1. The model is then used in a derivation of the Haldane *map distance*, a common measure used to locate SNPs, in Section 5.2 leading directly to a derivation of the correlation between markers under classic genetic population settings in Section 5.3. A software package in R which mirrors the model is outlined in Section 5.5 and used to simulate different recombination mechanics in the following sections before these are compared to real genetic data from mice in Section 5.7.

5.1 Denoting a genome

The model begins with

$$\mathbf{G} = [\mathbf{g}_1 | \mathbf{g}_2], \ \mathbf{g}_1, \mathbf{g}_2 \in \mathcal{B}^{N_P}$$

where $\mathcal{B} = \{\text{adenine, guanine, cytosine, thymine}\}$ is the set of nucleotide bases and N_P is the length of the genome. In humans $N_P \approx 3,234,830,000$. $\mathbf{G}$ represents the whole genome of an individual, with all chromosomes placed sequentially in two adjacent columns corresponding to the maternally and paternally inherited variants. Though both of these variants are complete double-stranded sequences of DNA, nucleotides pair uniquely. Adenine binds exclusively with thymine and guanine binds exclusively with cytosine. Therefore $\mathbf{g}_1$ and $\mathbf{g}_2$ record the pattern only for one of the two DNA strands for each column, the complementary strand is implied by this sequence and the unique binding of nucleotides.

Rather than address the whole genome, typically only a selected subset of segments are of interest. This is represented by

$$\mathbf{S} = [\mathbf{s}_1 | \mathbf{s}_2], \ \mathbf{s}_1, \ \mathbf{s}_2 \in \mathcal{B}^K$$

with $K \ll N_P$. The mapping $\mathbf{G} \to \mathbf{S}$ chooses K rows of $\mathbf{G}$ to create $\mathbf{S}$. This mapping is very seldom a random one. Previous work and databases of SNPs or other known markers motivate the choice of rows. Most commonly, then, the mapping $\mathbf{G} \to \mathbf{S}$ is a non-random selection of $M < K$ disjoint sequences from $\mathbf{G}$.

In the case where $\mathbf{S}$ contains only SNPs, the markers are *biallelic*, i.e. the population is dominated by two different sequences or *alleles* at the marker. These can be denoted using two different letters, such as A and B, or analogously the uppercase and lowercase version of the same letter, such as A and a. Converting the measured markers to letters is called annotation, a mapping $\mathbf{S} \to \mathbf{T}$ with

$$\mathbf{T} = [\mathbf{t}_1 | \mathbf{t}_2], \; \mathbf{t}_1, \; \mathbf{t}_2 \in \{A, a\}^M.$$

Denoting the i^{th} position of $\mathbf{t}_j$ as t_{ij}, $t_{lj} = A$ and $t_{mj} = A$ do not represent identical sequences at positions l and m. Instead this indicates that the sequences annotated by the capital at each position are present at their respective positions.

These annotated variants in $\mathbf{T}$ might next be converted to a numeric form. This is a mapping $\mathbf{T} \to \mathbf{X}$ such that

$$\mathbf{X} := [\mathbf{x}_1 | \mathbf{x}_2], \; \mathbf{x}_1, \; \mathbf{x}_2 \in \mathbb{R}^M.$$

Commonly this is even more restrictive, with $\mathbf{x}_j \in \{0, 1\}^M$ where

$$x_{ij} = \begin{cases} 1, & \text{if } t_{ij} = A \\ 0, & \text{if } t_{ij} = a \end{cases}, \tag{5.1}$$

is an indicator of the presence of the allele denoted by a capital.

Finally, $\mathbf{X}$ may be converted into a vector

$$\mathbf{z} \in \mathbb{R}^M$$

summarizing the individual's inherited variants using a map $\mathbf{X} \to \mathbf{z}$. Examples include *dominance*, which takes $z_i = \max\{x_{i1}, x_{i2}\}$, *homozygosity*, which takes $z_i = I_{x_{i2}}(x_{i1})$, and *additivity*, which takes $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are given according to Equation (5.1) and $I_y(x)$ is the indicator function

$$I_y(x) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases}$$

The additive summary gives $\mathbf{z} \in \{0, 1, 2\}^M$ so z_i is equal to the count of copies of A at the i^{th} marker across both of an individual's inherited variants.

Figure 5.1 displays this model, with descriptive names added to each mapping. In the first step, $\mathbf{G} \to \mathbf{S}$, segments of the genome are *selected* to obtain the marker sequences of interest. The next step, $\mathbf{S} \to \mathbf{T}$, *annotates* the chosen markers by indicating which of the common alleles is present at that marker. These annotations are then converted to numeric values, or *encoded*, in the step $\mathbf{T} \to \mathbf{X}$. Finally, the matrix $\mathbf{X}$ is *summarized* into a vector $\mathbf{z}$ with some row-wise operation.

5.2 Deriving map distance

The model presented in Section 5.1 is useful beyond providing a guide to genetic data. With only a few assumptions, it shows the Haldane map distance (Haldane, 1919) to be a corollary of the structure of DNA and mechanics of inheritance[2]. A derivation of the Haldane map using the model of Section 5.1 is completed here, starting with a simple sketch of sexual reproduction.

5.2.1 Sexual reproduction

Sexual reproduction is the recombination of the genomes of two parents to create offspring genetically distinct from both. A distinction must be made between reproductive or *sex* cells, e.g. sperm, and somatic cells. While somatic cells contain two variants of the germline, sex cells contain only one. When two sex cells combine, each provides its own variant to the resulting offspring. Inheritance is mediated by the creation of sex cells, which itself involves the random selection of variants contained within somatic cells by meiosis.

To track the parental variants which may be inherited, introduce two matrices to represent the maternal and paternal genomes of which $\mathbf{G}$ is the offspring:

$$\mathbf{M} = [\mathbf{m}_1 | \mathbf{m}_2] \text{ and } \mathbf{F} = [\mathbf{f}_1 | \mathbf{f}_2],$$

[2] In contrast to a physical distance based on the spatial position of markers, the Haldane map is a linkage distance which is based on the probability of markers being inherited together. There is a large body of literature developing linkage (or genetic map) distances, a quick overview can be found in Speed (2005). We break from the traditional probabilistic approach seen in Zhao and Speed (1996) and Lange (2002), for example, because it is not necessary to obtain the Haldane map and the derivation here makes direct use of the structure of $\mathbf{G}$ as outlined. Genetics is introduced here to contextualize later work rather than to fully review statistical methods in the field.

where $\mathbf{m}_1, \mathbf{m}_2, \mathbf{f}_1, \mathbf{f}_2 \in \mathcal{B}^{N_P}$. Crudely, sexual reproduction is the construction of $\mathbf{G}$ from one random column of $\mathbf{M}$ and one random column of $\mathbf{F}$. So, $\mathbf{G}$ could be $[\mathbf{m}_1|\mathbf{f}_2]$, for example.

The real mechanism is much more complex. During meiosis, the columns of $\mathbf{M}$ and $\mathbf{F}$ are perturbed. Rather than being inherited by $\mathbf{G}$ in the same form as in $\mathbf{M}$ and $\mathbf{F}$, regions in $\mathbf{f}_1$ may swap with regions in $\mathbf{f_2}$ and the same may occur with $\mathbf{m}_1$ and $\mathbf{m}_2$. This occurs either due to the *independent assortment of chromosomes* or due to the *crossing over* of variants within chromosomes.

Independent assortment is a direct consequence of the structure of the genome in somatic cells. Each chromosome is a separate molecule and so when sex cells are created, the variant of one chromosome inherited by offspring is independent of other chromosomes inherited from the same parent. This means that either of the paternal and maternal variants of a chromosome is equally likely to be passed on regardless of which variant is passed on for another chromosome.

Additionally, these variants may not be inherited identically as they appear in $\mathbf{M}$ or $\mathbf{F}$. There is a chance that the variants in a parent physically cross over each other while separating to form sex cells. Occasionally, this crossing results in a swap of the entire chromosome on either side of the cross, creating two completely new variants.

5.2.2 Modelling cross overs

Both crossing over and the independent assortment of chromosomes occur within each parental genome independently of the other parent, and so only one of the parents needs to be considered in modeling cross overs. Suppose it is $\mathbf{M}$.

Start with the assumption that genetic recombination is independent between chromosomes. Specifically, chromosomes not only assort independently but crossing over occurs independently on each chromosome and will affect only that chromosome's variants. This assumption can be thought of as a slightly stronger version of independent assortment. Therefore consider a vector

$$\mathbf{h} \in \{1, 2, \ldots, C\}^{N_P}$$

for $C \in \mathbb{N}$ which denotes the chromosomal membership of each row of $\mathbf{M}$. For simplicity, set $h_i \leq h_j$ for all $i \leq j$. In other words, all base pairs of a chromosome appear in adjacent rows with some specified ordering of the chromosomes. Assuming cross overs occur independently for each chromosome, a cross over in chromosome c, say, will affect only those rows of $\mathbf{M}$ where $\mathbf{h} = c$. Starting with the simplest case, where $\mathbf{h}$ is a vector of ones and so $\mathbf{M}$ contains a single chromosome, we can extend results to the entire genome by considering every other chromosome in the same way.

For this single chromosome, consider a cross over beginning at the i^{th} base pair, meaning the two variants of the chromosome physically cross at the i^{th} base pair. Assuming the variants are always perfectly aligned so that the i^{th} position on one variant will match with the i^{th} on the other, each variant is consquently separated into two parts: the part up to, but not including, the i^{th} base pair, and the part from the i^{th} base pair until the end. These two parts are then swapped between the variants, so that the first part of one variant forms a new chromosome with the second part of the other. Whenever a cross over is said to "begin at index i", it will refer to this sort of crossing: a swap of the columns for the first $i-1$ rows of $\mathbf{M}$ (or $\mathbf{F}$). Introduce an indicator vector

$$\mathbf{V} = (V_1, \ldots, V_{N_P})^{\mathsf{T}}$$

where

$$V_i = \begin{cases} 1 & \text{if a cross over at base pair } i \text{ occurs,} \\ 0 & \text{otherwise,} \end{cases} \tag{5.2}$$

and define $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{N_P})$ so that $\pi_i = P(V_i = 1)$. This can be done without loss of generality, as the order of cross overs in time does not affect the final chromosome. Any chromosome, offspring, or sex cell for which any cross overs have occurred is called *recombinant.*

As we rarely sequence the entire genome of an individual's somatic and sex cells, we will seldom see $\mathbf{M}$ and its recombinant forms. Instead, just as $\mathbf{S}$ is derived from $\mathbf{G}$, $\mathbf{M}_S$ and $\mathbf{F}_S$ are derived from $\mathbf{M}$ and $\mathbf{F}$ respectively. Swaps of the markers of $\mathbf{M}_S$ and $\mathbf{F}_S$ as inferred from $\mathbf{S}$ are then used to estimate the number of sex cells containing recombinant chromosomes. The proportion of sex cells produced with such a swap is called the *recombination rate* for the pair of markers.

However, the recombination rate for a pair of markers tells us nothing of how many cross over events occurred between them. Any odd number of events leads to a swap, while any even number will be undetectable. With this restricted view, the true count of indices i for which $V_i = 1$ cannot be known, and hence the π_i cannot be estimated individually.

5.2.3 Simplifying assumptions

Fortunately, if the recombination of two particular markers on the genome is all that is of interest, estimating individual π_i values is unnecessary. Consider two such marker positions,

j and k with $j < k$, and note that cross overs beginning at any of $j+1, j+2, \dots, k-1, k$ all result in these positions being split between variants. For identifiability assume that $\pi_{j+1} = \pi_{j+2} = \cdots = \pi_{k-1} = \pi_k = \pi_{j:k}$. Let N_c be a random variable counting the number of cross overs in the interval $\{j+1, j+2, \dots, k-1, k\}$. Then

$$P(N_c = n_c) = \binom{k-j}{n_c} \pi_{j:k}^{n_c} (1 - \pi_{j:k})^{k-j-n_c}$$

if cross overs occur independently. For brevity, let $r = k - j$ and $\pi = \pi_{j:k}$, which gives

$$P(N_c = n_c) = \binom{r}{n_c} \pi^{n_c} (1-\pi)^{r-n_c}, \tag{5.3}$$

where r is a unitless count of base pairs between positions j and k.

Recall that $N_P \approx 3{,}234{,}830{,}000$ in humans. This large number of base pairs spread over the 23 human chromosomes means that j and k will typically be separated by a great number of base pairs, and so r will be very large. Indeed, examples in Nyholt (2004), Salyakina et al. (2005), and Galwey (2009) have thousands or tens of thousands of base pairs between marker locations. Therefore, consider the limit of this expression as $r \to \infty$:

$$\lim_{r\to\infty} P(N_c = n_c) = \lim_{r\to\infty} \binom{r}{n_c} \pi^{n_c} (1-\pi)^{r-n_c}.$$

At this point, a substitution can be made:

$$\pi = \frac{\beta d(j,k)}{r} := \frac{\beta d}{r},$$

with $\beta, d(j,k) \in \mathbb{R}$. This substitution reparametrizes the probability π with a rate parameter, β, a distance measure, $d(j,k)$, and the r base pairs separating j and k. As the units of β and d will always result in a unitless product, their selection is arbitrary. Any distance d can be chosen and will invoke a corresponding β. This flexibility gives a great deal of freedom to choose a particular map distance to represent a corresponding model.

The substitution also leads to a substantial simplification, as

$$\begin{aligned} &\lim_{r\to\infty} \binom{r}{n_c} \left(\frac{\beta d}{r}\right)^{n_c} \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \\ =\ & \frac{(\beta d)^{n_c}}{n_c!} \lim_{r\to\infty} \frac{r^{n_c} + O(r^{n_c-1})}{r^{n_c}} \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \\ =\ & \frac{(\beta d)^{n_c}}{n_c!} e^{-\beta d} = \lim_{r\to\infty} P(N_c = n_c), \end{aligned} \tag{5.4}$$

the Poisson limit approximation for the binomial distribution.

Recall that if N_c is odd, it will result in a swap of markers j and k between variants, while if N_c is even, there will be no swap in the chromosome passed on. Define the recombination probability $p_r(d)$, which gives the probability of observing a swap for positions j and k with distance $d(j,k) := d$ between them. Then $p_r(d)$ is given by a sum of all odd terms from Equation (5.3). Taking the simplification of Equation (5.4) gives

$$\sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!} e^{-\beta d} = e^{-\beta d} \sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!} = \frac{1}{2}\left(1 - e^{-2\beta d}\right) = p_r(d). \tag{5.5}$$

A final substitution converts Equation (5.5) to a form familiar to researchers in genomics. Setting $\beta = \frac{1}{100}$ so that each each unit increase in d corresponds to a 0.01 increase in the expected number of cross overs gives Haldane's formula for the *map distance* in *centiMorgans* or cM.

By accounting for the structure of the genome and making a number of simplifying assumptions, our genetic model gives this classic result without any reference to the population-level differential equation used in its original derivation. Indeed, it indicates this population-level differential equation is a direct consequence of the structure of the genome. We can go a step further and compute a simple expression for genetic correlation using this distance under the additive map to summarize $\mathbf{X}$.

5.3 Genetic correlation

Recall $\mathbf{z}$ as depicted in Figure 5.1, the *correlation between markers* refers to the observed correlation matrix of the vector $\mathbf{z}$ in a particular population. This correlation has seen use in multiple test adjustment (Cheverud, 2001; Li and Ji, 2005; Galwey, 2009) and is generally important in defining linkage disequilibrium between markers. For clarity, let $\mathbf{z}$ indicate an instance of the random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_M)^{\mathsf{T}}$ which follows the distribution of the summarized values $\mathbf{z}$ in a particular population. This population may be real, as is the case when this modelling is used in practice, or purely hypothetical, as will be the case in the following analysis.

Consider two markers in the annotated matrix $\mathbf{T}$ at row indices j and k. Introduce $\mathbf{c}$, which is defined similarly to $\mathbf{h}$ from the derivation of map distance, but now indicates chromosomal membership for the markers in $\mathbf{T}$ rather than the base pairs in $\mathbf{G}$. As individual markers are not split over chromosomes, $\mathbf{c}$ is always unambiguously defined.

Either markers j and k are on the same chromosome, that is $c_j = c_k$, or they are not, and so $c_j \neq c_k$. If they are not, the assumptions of Section 5.2.2 dictate that there will be no correlation between Z_j and Z_k, as these markers will assort independently along with their respective chromosomes by assumption. If they are on the same chromosome, let $d(j,k) = d$ be the genetic distance between them measured in cM as in Equation (5.5). Denote the alleles of j with A and a respectively and use B and b analogously for k. Assume that the pairwise association of alleles at these markers in the population is of interest, i.e. that all other markers on this chromosome can be ignored for the moment. This setting creates the radically simplified

$$\mathbf{T} = \begin{bmatrix} A & a \\ b & B \end{bmatrix},$$

where the letters placed above are merely demonstrative. The simplified

$$\mathbf{X} = \begin{bmatrix} x_{j1} & x_{j2} \\ x_{k1} & x_{k2} \end{bmatrix},$$

with all entries in $\{0,1\}$ follows immediately. As was the case for $\mathbf{z}$, these lowercase entries are realizations of random variables X_{rs}, $r \in \{j,k\}, s \in \{1,2\}$. Then $\mathbf{X}$ implies

$$\mathbf{Z} = \begin{bmatrix} Z_j \\ Z_k \end{bmatrix} = \begin{bmatrix} X_{j1} + X_{j2} \\ X_{k1} + X_{k2} \end{bmatrix}$$

under the additive map. Consider $Corr(Z_j, Z_k)$ for a population resulting from the sexual reproduction of two known parents. The mechanics of sexual reproduction outlined in Section 5.2.1 and the genotype of the parents determine the distribution of Z_j and Z_k.

From the matrices $\mathbf{M}$ and $\mathbf{F}$ introduced alongside sexual reproduction, take simplified, annotated forms of these matrices to represent the paternal and maternal encodings at j and k. Explicitly,

$$\mathbf{F}_X = \begin{bmatrix} f_{j1} & f_{j2} \\ f_{k1} & f_{k2} \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} m_{j1} & m_{j2} \\ m_{k1} & m_{k2} \end{bmatrix}, \tag{5.6}$$

where all entries are once again in $\{0,1\}$. Assume that $\mathbf{F}_X$ and $\mathbf{M}_X$ are known and introduce the difference matrix

$$\mathbf{\Delta} = \begin{bmatrix} f_{j1} - f_{j2} & m_{j1} - m_{j2} \\ f_{k1} - f_{k2} & m_{k1} - m_{k2} \end{bmatrix} := \begin{bmatrix} \delta_{jF} & \delta_{jM} \\ \delta_{kF} & \delta_{kM} \end{bmatrix}. \tag{5.7}$$

This matrix will be useful in representing the correlation between Z_j and Z_k. Finally, assume that the variation in $\mathbf{Z}$ results purely from genetic recombination.

Begin with the expectation of $\mathbf{Z}$. Assuming no preferential inheritance of either variant, X_{j1} is equally likely to be either f_{j1} or f_{j2} and so takes a uniform distribution over these two possibilities. A similar logic for all other entries in $\mathbf{X}$ applies, and so

$$\begin{aligned} E[\mathbf{Z}] &= \begin{bmatrix} E[X_{j1}] + E[X_{j2}] \\ E[X_{k1}] + E[X_{k2}] \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} f_{j1} + f_{j2} + m_{j1} + m_{j2} \\ f_{k1} + f_{k2} + m_{k1} + m_{k2} \end{bmatrix}, \end{aligned}$$

from which it follows

$$\begin{aligned} Var(Z_j) &= E[(X_{j1} + X_{j2})^2] - E[Z_j]^2 \\ &= \frac{1}{4}\left[\sum_{i=1}^{2}\sum_{k=1}^{2}(f_{ji} + m_{jk})^2 - (f_{j1} + f_{j2} + m_{j1} + m_{j2})^2\right]. \\ &= \frac{1}{4}\left[(f_{j1} - f_{j2})^2 + (m_{j1} - m_{j2})^2\right] \\ &= \frac{1}{4}\left[\delta_{jF}^2 + \delta_{jM}^2\right]. \end{aligned} \tag{5.8}$$

Analogously,

$$Var(Z_k) = \frac{1}{4}\left[\delta_{kF}^2 + \delta_{kM}^2\right]. \tag{5.9}$$

While the covariance

$$Cov(Z_j, Z_k) = \sum_{l=1}^{2}\sum_{m=1}^{2} Cov(X_{jl}, X_{km}) \tag{5.10}$$

can be expressed as a sum of four terms, each of which can be considered in turn.

$Cov(X_{j1}, X_{k2})$ and $Cov(X_{j2}, X_{k1})$ can be evaluated immediately. Both of these terms measure the covariance between values on the diagonals of $\mathbf{X}$, that is the covariance between the maternally and paternally donated variants of the genome inherited from $\mathbf{F}$ and $\mathbf{M}$, respectively. These covariances therefore measure the amount of inbreeding in a population, the degree to which parents tend to have the same genotype. Crow and Kimura (1970) quantify these covariances with a coefficient r for general populations. With known parents, however, these diagonal values are independent of each other and therefore uncorrelated.[3] Explicitly, $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$ if $\mathbf{F}_X$ and $\mathbf{M}_X$ are known matrices.

[3] This can be confirmed by tedious algebra.

$Cov(X_{j1}, X_{k1})$ and $Cov(X_{j2}, X_{k2})$ measure the covariance of encodings on the same variant, and cannot be so easily dismissed. Instead, consider $Cov(X_{j1}, X_{k1})$ and expand:

$$Cov(X_{j1}, X_{k1}) = E[X_{j1}X_{k1}] - E[X_{j1}]E[X_{k1}].$$

The equal probabiliy of inheritance of variants gives $E[X_{j1}] = \frac{1}{2}(f_{j1} + f_{j2})$ and $E[X_{k1}] = \frac{1}{2}(f_{k1} + f_{k2})$. Next consider $E[X_{j1}X_{k1}]$.

There are four possible values of $X_{j1}X_{k1}$, corresponding to inheritance of either of the two parental variants with or without recombination. If no recombination occurs, an event with probability $1 - p_r(d)$, either $f_{j1}f_{k1}$ or $f_{j2}f_{k2}$ is inherited with equal probability. If a cross over between j and k leads to recombination, then either $f_{j1}f_{k2}$ or $f_{j2}f_{k1}$ is passed on with equal probability. Accounting for these four possibilities gives

$$E[X_{j1}X_{k1}] \quad = \quad \big(1 - p_r(d)\big)\left(\frac{1}{2}f_{j1}f_{k1} + \frac{1}{2}f_{j2}f_{k2}\right) + p_r(d)\left(\frac{1}{2}f_{j1}f_{k2} + \frac{1}{2}f_{j2}f_{k1}\right).$$

Combining this with the expectations of X_{j1} and X_{k1} and simplifying gives

$$\begin{aligned} Cov(X_{j1}, X_{k1}) &= E[X_{j1}X_{k1}] - E[X_{j1}]E[X_{k1}] \\ &= \frac{1}{4}\big(1 - 2p_r(d)\big)\Big(f_{j1}f_{k1} + f_{j2}f_{k2} - f_{j2}f_{k1} - f_{j1}f_{k2}\Big) \\ &= \frac{1 - 2p_r(d)}{4}\delta_{jF}\delta_{kF}. \end{aligned} \tag{5.11}$$

By the same logic

$$Cov(X_{j2}, X_{k2}) = \frac{1 - 2p_r(d)}{4}\delta_{jM}\delta_{kM}. \tag{5.12}$$

We obtain the covariance of Z_j and Z_k by substituting Equations (5.11) and (5.12) and $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$ into Equation (5.10) to get

$$Cov(Z_j, Z_k) = \frac{1 - 2p_r(d)}{4}\Big[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\Big]. \tag{5.13}$$

Finally, Equations (5.8), (5.9), and (5.13) can be combined to determine the correlation:

$$\frac{Cov(Z_j, Z_k)}{\sqrt{Var(Z_j)Var(Z_k)}} := (1 - 2p_r(d))\gamma = Corr(Z_j, Z_k), \tag{5.14}$$

where

$$\gamma = \frac{\Big[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\Big]}{\sqrt{\big(\delta_{jF}^2 + \delta_{jM}^2\big)\big(\delta_{kF}^2 + \delta_{kM}^2\big)}}. \tag{5.15}$$

So, the correlation is a product of $\big(1 - 2p_r(d)\big)$, which depends on the markers in question, and a factor γ, which depends on the known parents. An even simpler expression is obtained by substituting the Haldane recombination probability from Equation (5.5) in place of $p_r(d)$:

$$\begin{aligned} Corr(Z_j, Z_k) &= \big(1 - 2p_r(d)\big)\gamma \\ &= \left(1 - 2\left[\frac{1}{2}\left(1 - e^{-2\beta d}\right)\right]\right)\gamma \\ &= \gamma e^{-2\beta d}, \end{aligned} \tag{5.16}$$

and so using the Haldane map distance the correlation between Z_j and Z_k decays exponentially in $d(j,k)$ with an intercept γ determined by the parents' annotated matrices. As the entries in $\mathbf{M}_X$ and $\mathbf{F}_X$ are all 0 or 1, the differences in $\boldsymbol{\Delta}$ are all -1, 0, or 1. There are therefore $3^4 = 81$ potential γ values, though most of these are not unique. 17 of these are undefined, corresponding to cases where $Var(Z_j) = 0$ or $Var(Z_k) = 0$. Table 5.1 summarizes the frequency of different γ values for the remaining 64 combinations. Only five symmetrically-distributed values are possible. A number of population settings for γ are of particular interest due to their use in mouse breeding experiments outlined in Green et al. (1966).

γ	Frequency
-1	8
$-\frac{1}{\sqrt{2}}$	16
0	16
$\frac{1}{\sqrt{2}}$	16
1	8

Table 5.1: Frequency of γ values across the 64 combinations for which correlation is defined

One such setting is the the F_2 *intercross* design. *Cross* here is short for sexual reproduction, rather than crossing over. This design considers the population resulting from the cross of $\mathbf{M}_X$ and $\mathbf{F}_X$ with

$$\mathbf{F}_X = \mathbf{M}_X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

In this setting all the differences in γ are 1 and so $\gamma_{\text{inter}} = 1$.

The next is the N_2 *backcross*. Here the cross is between $\mathbf{M}_X$ and $\mathbf{F}_X$ defined as

$$\mathbf{F}_X = \begin{bmatrix} f & f \\ f & f \end{bmatrix}, \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

where $f \in \{0, 1\}$. In this setting, both differences defined on $\mathbf{F}_X$ are 0 while both of those defined on $\mathbf{M}_X$ are 1. This gives $\gamma_{\text{back}} = 1$, the same as that of the intercross population.

Other interesting cases without historical basis involve

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ or } \mathbf{M}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

as these can result in $\gamma < 0$, and so a negative correlation. For example, if

$$\mathbf{F}_X = \mathbf{M}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then $\gamma = -1$. Many other settings lead to no measured correlation. Take

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

for example. Note that these negative values are somewhat arbitrary. The encoding of 1 or 0 for particular alleles at a marker is not prescribed, but is rather an analytical choice. Therefore in any of these cases the encoding could be switched to give a positive γ of the same magnitude.

Finally, these results can be extended to the whole genome. Recalling that j and k were restricted to the same chromosome, this pairwise result can be generalized to the correlation matrix of $\mathbf{Z}$ for markers measured on different chromosomes. For markers on the same chromosome correlations will be proportional to $1 - 2p_r(d)$, where $p_r(d)$ is the probability of recombination as a function of the distance between markers. Based on the independent assortment of different chromosomes, the correlations will be zero for any pair j and k not on the same chromosome.

In other words, if $c_j = c_k$, Equation (5.14) dictates the correlation between Z_j and Z_k. On the other hand, if $c_j \neq c_k$ the correlation between Z_j and Z_k will be zero. This implies a block diagonal structure corresponding to the chromosomes with correlations dictated by the probability of recombination within each chromosome. Most generally

$$Corr(Z_j, Z_k) = I_{c_j}\left(c_k\right) \gamma\big(1 - 2p_r(d)\big), \tag{5.17}$$

and under the assumptions of the Haldane model

$$Corr(Z_j, Z_k) = I_{c_j}\left(c_k\right)\gamma e^{-2\beta d(j,k)}. \tag{5.18}$$

For a demonstration of how this expression can be helpful in practice, see how it is used to improve robustness to missing data in Appendix A.

5.4 Inbreeding

Recombination not only generates variation in the genome captured by the correlation, it can also remove all variation at a marker location in a population, called *fixation* of the genetic sequence at a marker. *Inbreeding*, the sexual reproduction of genetically similar individuals, leads to fixation if the individuals are similar enough and inbreeding continues for long enough[4]. To see how, consider all possible offspring of the most extreme case: the reproduction of two siblings with the same parents.

Let the variants of the marker at position j be x_{j1}, x_{j2} for the first sibling and y_{j1}, y_{j2} for the second sibling. Fixation is defined as the event $x_{j1} = x_{j2} = y_{j1} = y_{j2} = a$, as from that point onwards the only version of marker j which can be inherited is a. As the siblings have shared parentage, represent the two paternal copies at that marker with f_{j1} and f_{j2} and the two maternal copies with m_{j1} and m_{j2} and assume that $f_{j1} = f_{j2} = m_{j1} = m_{j2}$ is not true, i.e. that the marker is not already fixed. Next, we can simplify the problem by taking $f_{j1}, f_{j2}, m_{j1}, m_{j2} \in \{0, 1\}$ and considering the additive summary for each sibling $t_1 = x_{j1} + x_{j2} \in \{0, 1, 2\}$ and $t_2 = y_{j1} + y_{j2} \in \{0, 1, 2\}$. As each variant is passed on with equal probability, using the additive summary results in no loss of generality.

Under these assumptions we can capture the state of the siblings with the tuple $\mathbf{t} = (z_{j1}, z_{j2})$, or generally for siblings after k generations of sibling inbreeding as $\mathbf{t}^{(k)} = (z_{j1}^{(k)}, z_{j2}^{(k)})$. Fixation occurs if $\mathbf{t}^{(k)} = (2, 2)$ or $\mathbf{t}^{(k)} = (0, 0)$, as $\mathbf{t}^{(k+n)} = \mathbf{t}^{(k)}$ for any $n \in \mathbb{Z}$ in both cases, and the rules of recombination outlined prior define a transition matrix for

[4] We assume that mutations do not occur or at least that they are rare enough to be ignored.

$\mathbf{t}^{(k)}$ to $\mathbf{t}^{(k+1)}$:

$$\mathbf{R} = \begin{array}{c|ccccccccc} & (0,0) & (0,1) & (1,0) & (0,2) & (1,1) & (2,0) & (1,2) & (2,1) & (2,2) \\ \hline (0,0) & 1 & & & & & & & & \\ (0,1) & 1/4 & 1/4 & 1/4 & & 1/4 & & & & \\ (1,0) & 1/4 & 1/4 & 1/4 & & 1/4 & & & & \\ (0,2) & & & & & 1 & & & & \\ (1,1) & 1/16 & 1/8 & 1/8 & 1/16 & 1/4 & 1/16 & 1/8 & 1/8 & 1/16 \\ (2,0) & & & & & 1 & & & & \\ (1,2) & & & & & 1/4 & & 1/4 & 1/4 & 1/4 \\ (2,1) & & & & & 1/4 & & 1/4 & 1/4 & 1/4 \\ (2,2) & & & & & & & & & 1 \end{array}$$

where zeros have been left blank for readability. $\mathbf{R}$ has the stationary distribution $\pi = (1/2, 0, 0, 0, 0, 0, 0, 0, 1/2)$, is aperiodic, and has two absorbing states, both of which correspond to fixation. Therefore, as $k \to \infty$, $P(\mathbf{t}^{(k)} \notin \{(0,0), (2,2)\}) = 0$ and so fixation for any marker is inevitable under sustained brother-sister inbreeding. About 99% of genetic markers are fixed after 20 generations of inbreeding (Casellas, 2011), meaning roughly 20% of unfixed markers become fixed for every generation of brother-sister inbreeding[5].

This is not an abstract consideration, it has been used since the early 1900s to create genetically similar *strains* of mice to investigate the heritability of traits and disorders (Green et al., 1966). Some of these strains have been bred for more than a century by groups such as the Jackson Laboratory as models for disease and drug testing, and so are practically genetically uniform (Beck et al., 2000). While mutations and selection pressures favouring genetic variability can still cause genetic drift (Casellas, 2011), these strains are much less variable than natural populations and were the best instrument to reveal the associations between genetic information and physical traits for much of the twentieth century.

Modern advances in sequencing technology and computation have drastically changed this as every subject in a study can now be effectively sequenced at hundreds of thousands of markers (LaFramboise, 2009; Uffelmann et al., 2021). Model organisms still produce clearer results, as less genetic variation means less noise for models than incorporate the genome. Additionally, randomized and controlled experimental designs are possible to implement for mouse strains which cannot be conducted on humans.

The advancement of computing power has presented simulation as a third option within the last decade (Messer, 2013). Rather than fight against the dynamics of actual repro-

[5] Wright (1933), considering numerous inbreeding scenarios including half-siblings, derived a value of 19.1% per generation of brother-sister inbreeding.

duction, modern software can be used to simulate genetics under complete control to test models and gain insight into real genetic processes. Based on the simple model presented above, such a simulated genetic system was created in R.

5.5 The toyGenomeGen package

The R `toyGenomeGen` package is based on the S3 classes `genome` and `population` and functions that act on these classes to mimic genetic recombination. An instance of the `genome` class has four slots:

`encoding`: a two-column numeric matrix $\mathbf{X}$ giving the encodings of markers,

`alleles`: a list with the same length as the number of rows as $\mathbf{X}$ providing the annotations in $\mathbf{T}$,

`chromosome`: a factor giving the chromosomal membership of each row of $\mathbf{X}$, and

`location`: a list of numeric vectors providing the distance into each chromosome the markers are found.

Marker names, for example SNP identifiers from NCBI (2021), are stored as the row names of `encoding`. A simple `print` method prevents the potentially overwhelming entire genome from being printed all at once, while a `plot` method visualizes genome objects using separate lines for each chromosome on which points with shapes indicating the values in the corresponding row of `encoding` are plotted.

A `genome` object can be created either randomly using `simGenome`, based on provided slot values using `makeGenome`, or from an appropriately-structured `data.frame` using `asGenome`. Random generation is supported by helper functions which generate marker encoding matrices, locations, and chromosomes given some parameters. For flexbility, `simGenome` accepts these helpers as arguments, allowing for users to define functions that create the behaviour they would like to model.

A `population` object is a more memory-efficient representation of a list of genomes measured at the same locations which avoids redundant information by placing all $\mathbf{X}$ matrices into a list in the slot `encodings` and storing their common locations, chromosomes, and alleles in slots identical to a single `genome`. A `population` is created by calling `asPopulation` on a list of genomes, which removes row names from each encoding and fills the alleles, chromosome, and location slots using the first genome in the list after checking for consistency. A slot called marker is also added to provide the marker names. Populations can be subset on both markers and individual genomes using the function subsetPopulation, and individual genomes can be extracted using selectGenome. Dynamic selection of individual genomes based on a logical function acting as an inclusion rule is supported by the filterPopulation wrapper.

Aside from basic manipulation and display functionality, the mechanics of genetic recombination are supported by the functions meiose and sex. meiose accepts a genome object and produces a single column encoding generated by modelling the crossing over and independent assortment of its encodings. By default, it uses the Haldane map and recombination model described above, but it accepts any function which outputs a vector of indices that indicate where cross-overs occur. Given the locations of cross over events, sections of the columns of encoding are swapped accordingly. sex is a wrapper function for meiose that accepts two genome objects and then combines their independent meiosis products.

Correlation based on linkage disequilibrium within genome objects is computed using popCorrelation and theoryCorrelation. Calling popCorrelation computes Pearson's product moment correlation for a given population object and a scoring function. theoryCorrelation instead accepts a mapping function and setting which are used with Equation (5.17) to generate a theoretical correlation matrix. By default, the additive scoring $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$ and Haldane map are used. Correlation matrices generated by either can be visualized using the image wrapper corrImg with guidelines for the chromosomes added by addChromosomeLines after corrImg has been called.

In addition to these classes and functions, toyGenomeGen includes eleven real genetic data sets adapted from public data hosted on the Mouse Genome Database (MGD) (Bult et al., 2019). The MGD provides annotation data for more than a dozen mouse populations resulting from crosses of known strains of mice alongside references which allow the cM distances between markers to be determined. All of these resources are publicly provided in tab-delimited text files at the Mouse Genome Informatics website: www.informatics.jax.org, and panels with clear legends were extracted and converted to population objects for the R package.

5.5.1 Similar packages

toyGenomeGen occupies a previously unfilled niche in genetics packages for R. Flexible experiments which virtually test different models of heredity are not easily carried out in existing packages, as most provide support for analysis early in the schematic shown in Figure 5.1. The rtracklayer (Lawrence et al., 2009) and BSgenome (Pagés, 2023) packages, for example, are primarily concerned with efficiently loading and annotating full genome sequences, corresponding to steps between the representations **G**, **S**, and **T** in our schematic. valr (Riemondy et al., 2017), GenomicRanges (Lawrence et al., 2013), and a handful of less popular related packages support efficient comparison of genomic intervals, or rows within **G**.[6]

Existing simulation packages focus on mimicking existing data, often samples from the 1000 Genomes Project (Fairley et al., 2020) or NCBI's dbGaP (NCBI/NLM, 2023). This is the explicit goal of sim1000G (Dimitromanolakis et al., 2019) and TriadSim (Shi et al., 2018). As a result, both of these packages restrict the methods of recombination which can be used to move a population forward in order to replicate the patterns already observed in samples rather than explore the possible mechanisms that lead to those patterns. TriadSim, in particular, chooses an inflexible "hot-spot" model to choose recombination events.

Outside of R , there are many software tools that simulate genetics for the purpose of testing analytical methods or generating pseudo-observations under different assumptions. The United States National Institutes of Health maintains a list of reviewed genetic simulation resources which contained 227 approved simulators at the beginning of October 2023. Using the web page's "Compare by attribute" feature, these were filtered down to simulators which match the detail and functionality of toyGenomeGen, leaving only two candidate software tools which model the evolution of large populations over many generations. The first, simuPOP (Peng et al., 2012) implemented in python and C++, models recombination with its Recombinator function class that supports custom probabilities of recombination between markers or requires a distance and intensity pair for each to determine recombination. These mechanics are identical to the defaults in toyGenomeGen. The second, SLiM (Haller and Messer, 2023), is implemented in a custom scripting language that supports many different models of recombination that include and extend beyond those explored here.

Of course, the toy genomic model developed for toyGenomeGen has been created primarily to facilitate understanding and exploration. The broader scopes of simuPOP and SLiM give far more functionality to model evolution by linking genotypes to phenotypes, supporting mating pair selection, evaluating fitness, and modeling multiple species simultaneously. These complex relationships are not needed to introduce genetics in the context

[6] BSgenome, rtracklayer, and GenomicRanges all share authors and seem to be products of an organized team effort by multiple researchers at the Fred Hutchison Cancer Research Center in Seattle.

of exploring data, however. toyGenomeGen provides a light and focused alternative to these extensive software projects.

5.6 Replicating common simulation patterns

To demonstrate this use, toyGenomeGen is used to recreate simulated population settings in the literature using the default Haldane map. Cheverud (2001) investigates the correlation between markers for a single chromosome with equidistant markers for chromosome lengths of 50, 75, and 100 cM with markers equidistant at 50, 25, 12.5, and 6.25 cM were simulated for populations of 500 F_2 intercross offspring. Lander and Botstein (1989) simulates twelve chromosomes of length 100 cM with markers every 20 cM along each for a population of 250 N_2 backcross offspring.

The simulations of Cheverud (2001) and Lander and Botstein (1989) were recreated using toyGenomeGen. Specifically, these were the 100 cM chromosome with 6.25 cM separated markers of Cheverud (2001) and the twelve 100 cM chromosomes with 20 cM separated markers of Lander and Botstein (1989). The resulting simulated correlation matrices and theoretical correlation matrices are visualized side by side using corrImg in Figures 5.2 and 5.3. The same colour palette is used as earlier.

Figure 5.2(a) displays a pattern of constant off-diagonal lines of decreasing value, as expected from Equation (5.5). Roughly the same pattern is seen in Figure 5.2(b), though it is noisier. Rather than having clear constant lines along each off-diagonal, Figure 5.2(b) has regions of similar values which occur across several off-diagonal lines. This leads to the appearance of large squares of more strongly related values, a pattern absent from Figure 5.2(a).

Figure 5.3 displays the setting of Lander and Botstein (1989) with the addition of guide lines to aid in reading the plot. As suggested by Equation (5.5) and shown in Figure 5.3(a), Figure 5.3(b) has a stark block diagonal structure which agrees with these guide lines. The simulation therefore agrees very well with theory in this aspect. Within the chromosomes, there is also good agreement between Figure 5.3(a) and Figure 5.3(b). Both have decreasing correlations along the off-diagonal lines, with Figure 5.3(b) displaying similar departures from Figure 5.3(a) as Figure 5.2(b) does from Figure 5.2(a).

A more interesting noise pattern is seen between chromosomes outside the blocks in Figure 5.3(b). Unlike the strictly positive correlations seen in Figure 5.2, both negative and positive correlations are observed. Though many chromosomes show consistent patterns between their markers, with all correlations either positive or negative as between

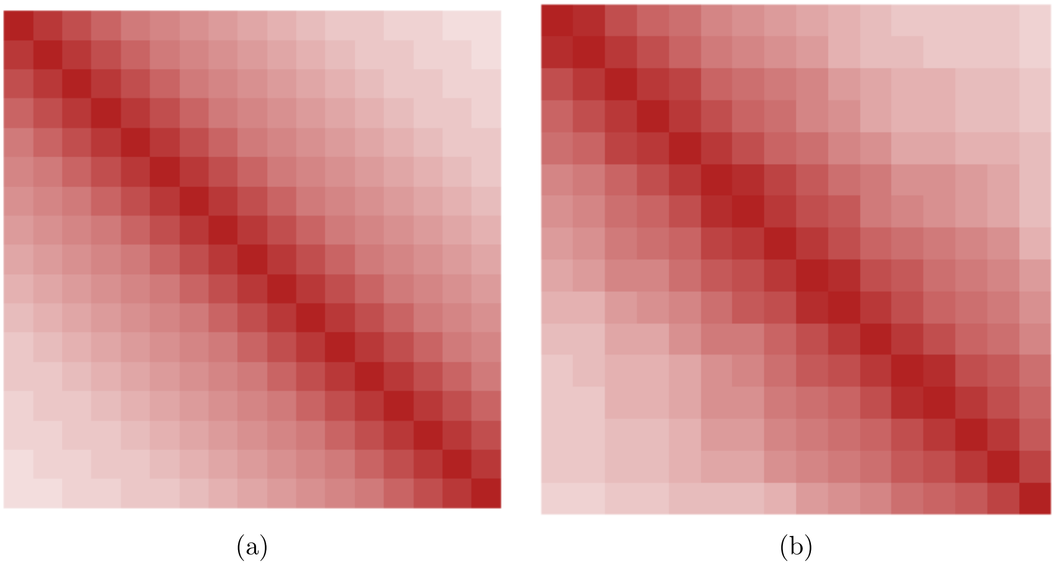


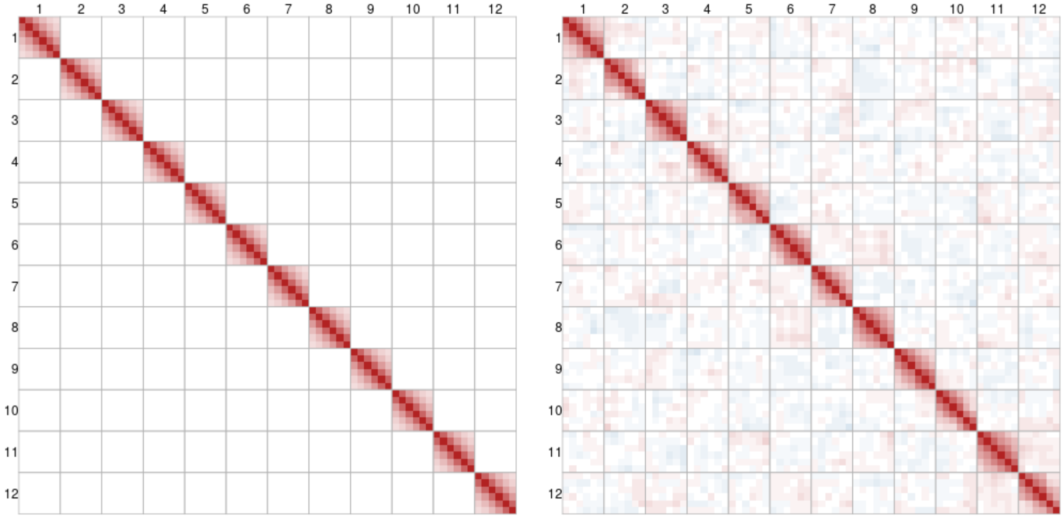
Figure 5.2: The (a) theoretical and (b) simulated correlation matrix of a population of 500 F_2 intercross offspring measured on a 100 cM chromosome with markers each 6.25 cM apart.

chromosomes 9 and 6 or 11 and 12, many have more complicated relationships. Between chromosomes 7 and 3, for example, both negative and positive correlations are observed between markers which are larger than the smallest intra-chromosomal correlations within 2. This gives a sense of what patterns we might expect in real data between chromosomes.

5.7 Comparing the model to reality

The motivation for the creation of `toyGenomeGen` and its default settings was to complement the simple genetic model derived previously. However, this simple model and the context it provides are only useful if they reflect reality. The data included in `toyGenomeGen` give a perfect opportunity to assess this.

Two of the real data sets included in `toyGenomeGen` are independent realizations of an identical population setting: the *BSB mouse cross* first outlined in Fisler et al. (1993). BSB mice are those resulting from the N_2 backcross of the C57BL/6J and *Mus Spretus* inbred mouse strains, detailed respectively in JAX (2022) and Dejager et al. (2009).



(a)

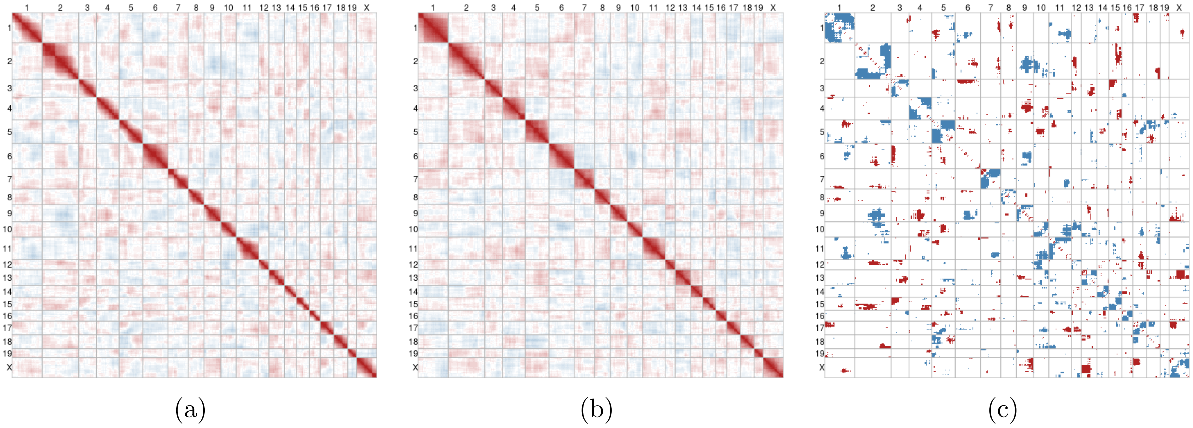
(b)

Figure 5.3: The (a) theoretical and (b) simulated correlation matrices of a population of 250 N_2 backcross offspring measured on twelve 100 cM chromosomes with markers 20 cM apart on each.

The first of these is `jax_bsb` from Rowe et al. (1994) and the second is `ucla_bsb` from Welch et al. (1996). Both the JAX and UCLA BSB cross data were downloaded from `www.informatics.jax.org/downloads/reports/index.html` before being converted to `populations` and saved.

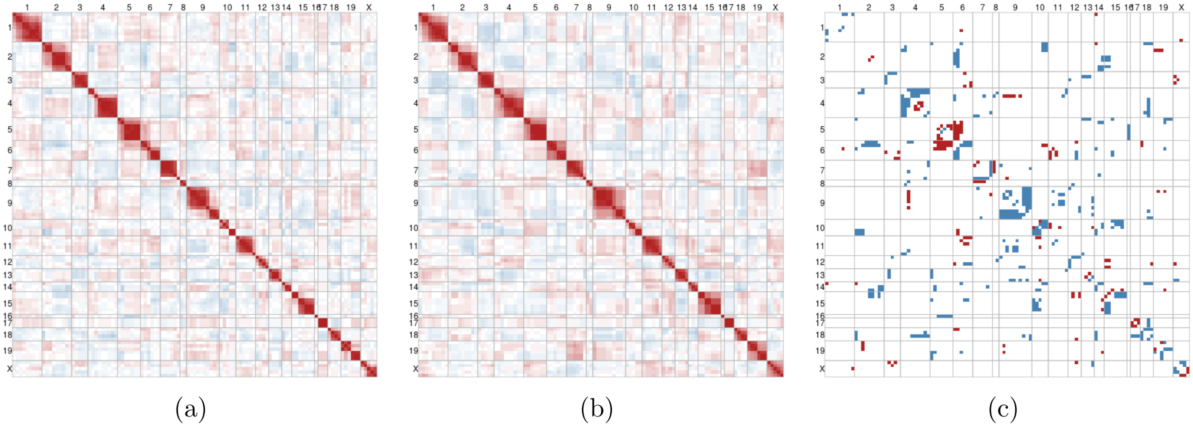
The data sets require further cleaning before being used, however. Any markers without complete observations are removed from both data sets and any individual mice with incomplete data are excluded using the `subsetPopulation` function. For the JAX BSB data this leaves 94 mice annotated at 1496 markers while the UCLA BSB data has 66 mice annotated at 111 markers. The correlation matrices for these data sets are displayed in Figures 5.4(a) and 5.5(a) respectively using the divergent palette defined earlier.

To determine the expected distribution of these correlations, the cM positions of measured markers were used to simulate 10,000 crosses under the Haldane model of independent recombination for each of the JAX BSB and UCLA BSB settings with `toyGenomeGen`. Figures 5.4(b) and 5.5(b) display example correlation matrices from one such simulated population. For both settings, the quantile of each experimental pairwise correlation was then computed using the 10,000 simulated crosses. Figures 5.4(c) and 5.5(c) shade quantiles which are less than 250 and greater than 9,750 for their respective settings. These correspond to unadjusted two-sided 95% confidence rejection regions for each correlation.



(a) (b) (c)

Figure 5.4: (a) Experimentally observed and (b) simulated correlations for markers from Rowe et al. (1994). (c) displays quantiles determined from 10,000 simulated crosses. Quantiles less than 250 or greater than 9,750 are shaded.



(a) (b) (c)

Figure 5.5: (a) Experimentally observed and (b) simulated correlations for markers from Welch et al. (1996). (c) displays quantiles determined from 10,000 simulated crosses. Quantiles less than 250 or greater than 9,750 are shaded.

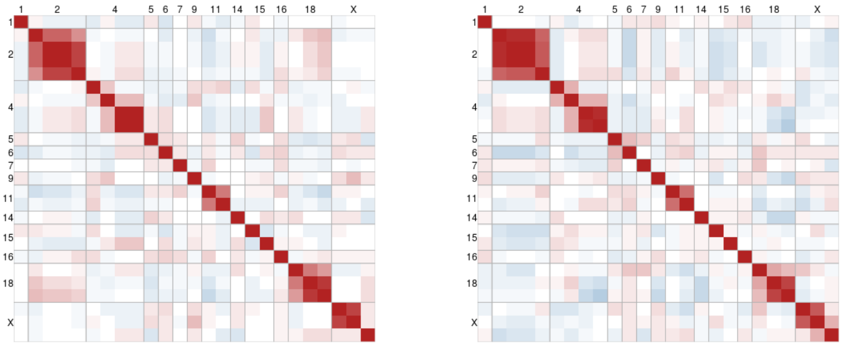
Qualitatively, the simulated examples show good agreement to experimental results. In both Figures 5.4 and 5.5 the patterns of correlation between chromosomes are similar between experiment and simulation. Figures 5.4(c) and 5.5(c) additionally suggest that patterns of departure may simply be noise. The shaded regions of unusually strong correlations do not appear to follow any clear pattern.

The similarity continues within chromosomes. Figures 5.4(c) and 5.5(c) are generally not shaded within chromosomes. In particular, very little of the region close to the diagonal is shaded. The most noteworthy pattern in either sub-plot occurs in the corners of the diagonal squares indicating chromosomes in Figure 5.4(c). Many of these corners are shaded blue, suggesting these distant intra-chromosome correlations are less than might be expected. The pattern of shading is suggestive of block structures within chromosomes where contiguous sections are fit well by the model but may have more complex dynamics between them.

A likely explanation is the non-independence, or *interference*, of cross overs. Broman et al. (2002) evaluated the pattern of cross overs in the cross of Rowe et al. (1994), the basis of the JAX BSB data, and found that cross overs were not fully independent. Most mouse chromosomes are much less than 100 cM in length, yet cross overs rarely occur within 20 cM of each other and fewer cross overs than expected occurred on the same chromosome. This interference will have little impact on the correlation between markers with short distances between them, as more than one cross over event is unlikely to occur in a short interval. Markers separated by longer distances are impacted by this observed interference to a greater extent, as the observed number of double cross overs will be less than expected. This increases the chance that distant markers will be separated in meiosis by a single crossover, leading to a weaker correlation than predicted by the model.

That said, this pattern is not repeated in Figure 5.5 and the shading of quantiles has not been adjusted to account for the many multiple tests performed in each plot. In order to get a greater sense of this experimental departure from our simple model, the common markers measured between the UCLA BSB and JAX BSB data were identified and the correlation matrices computed for these common locations in order to view the behaviour of two experimental replicates rather than two cases with one. These correlations are displayed in Figure 5.6. Most chromosomes have only one marker measured in common between these experiments, but chromosomes 2, 4, and 18 have several.

These common markers were again used to simulate 10,000 independent replicates of each of the JAX and UCLA crosses which were paired and the average of the correlation matrices computed for each pair. Independence was assumed because the experiments of Rowe et al. (1994) and Welch et al. (1996) were carried out years apart in different labs. The results of this simulation are displayed in the novel *correlation test plot* of Figure 5.7.



(a) JAX BSB data

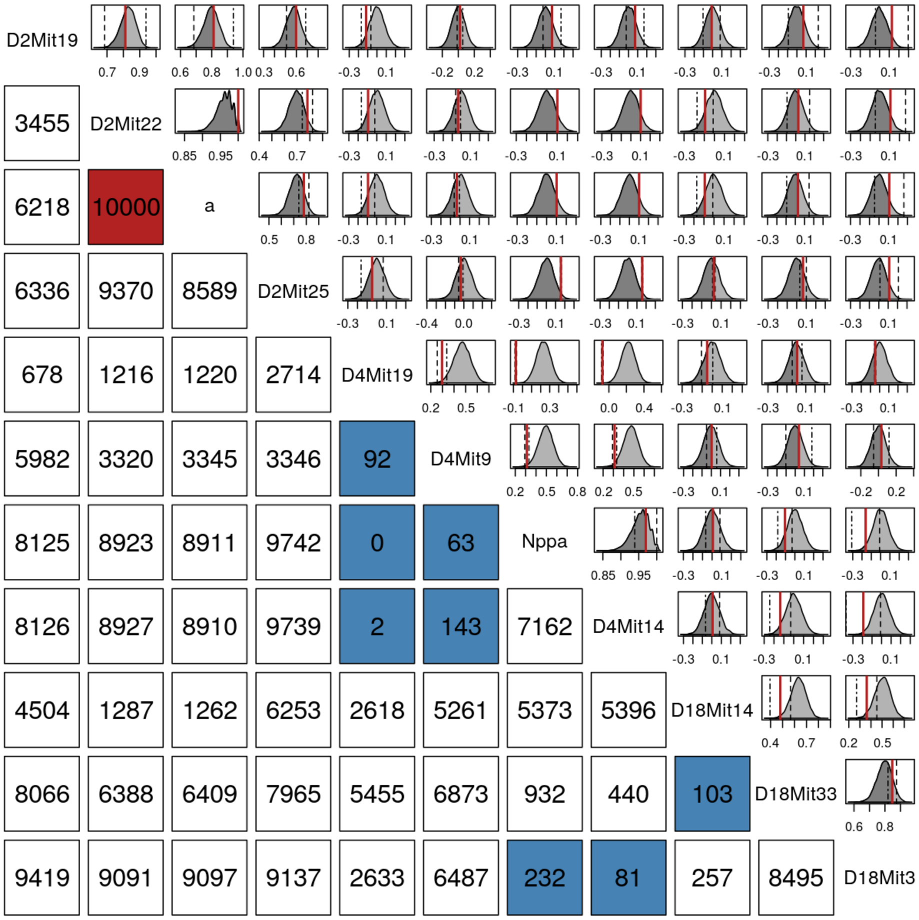
(b) UCLA BSB data

Figure 5.6: Pairwise correlations for the common marker positions of the JAX and UCLA BSB data.

5.7.1 The correlation test plot

The heatmaps displayed so far have some shortcomings. As they are symmetric about the diagonal, identical information is encoded by the cells on either side, meaning effectively half of the plot is simply a copy. Further, since all the plot is occupied by hue information, effectively communicating the distribution of repeated samples is impossible. These limitations are necessary when displaying hundreds or thousands of cells, but the common BSB markers number in the tens. This gives an opportunity to display much more information.

Figure 5.7 simultaneously displays the observed quantile and simulated distribution of the mean correlation across the 10,000 pairs using a matrix of panels. Along the diagonal, each panel displays the name of a marker. Above the diagonal, the panels display a kernel density estimate (KDE) of the distribution of mean correlations across all 10,000 simulated pairs. Added to this plot are a dashed line to indicate the JAX correlation, a dot-dashed lined to indicate the UCLA correlation, and a thick line to indicate their mean. The area of the KDE below this mean is shaded, and the number of simulated pairs in this shaded region is displayed in the corresponding panel below the diagonal. The panels below the diagonal therefore report the quantile of the mean in the simulated data, with shading added when the value is less than 250 or above 9,750. This plot allows us to see not only the quantiles of the observed means, but also the distributions of those means across the 10,000 simulated populations.



D2Mit19										
3455	D2Mit22									
6218	10000	a								
6336	9370	8589	D2Mit25							
678	1216	1220	2714	D4Mit19						
5982	3320	3345	3346	92	D4Mit9					
8125	8923	8911	9742	0	63	Nppa				
8126	8927	8910	9739	2	143	7162	D4Mit14			
4504	1287	1262	6253	2618	5261	5373	5396	D18Mit14		
8066	6388	6409	7965	5455	6873	932	440	103	D18Mit33	
9419	9091	9097	9137	2633	6487	232	81	257	8495	D18Mit3

Figure 5.7: The *correlation test plot* for the JAX and UCLA BSB crosses. The upper cells show the distribution of 10,000 simulated averaged correlations between the JAX and UCLA BSB crosses. The experimental results are marked by broken lines and their mean marked by a thick line. The bottom cells give the quantile of the corresponding mean.

In Figure 5.7, the distributions of simulated mean correlations are generally symmetric and unimodal. The shape and spread of the distribution of correlations seems highly dependent on the proximity of a pair of markers. Markers which are close together in cM and have a high correlation display very little variation across the simulations relative to markers which are further apart on the same chromosome or are on different chromosomes. Generally, the observed mean correlations in the real data are not extreme relative to the simulated distributions. This can be seen in both the KDEs above the diagonal and the quantiles shown below it.

Of the fifty five lower cells, nine are shaded. The first of these, between markers D2Mit22 and a, is misleading. The observed quantiles are computed by counting the values less than or equal to that observed, but this pair has an observed mean correlation of 1. It is therefore necessarily larger than or equal to all other mean correlations, despite having an identical value to 291 simulated means. This shading should therefore be ignored, as the value is not so unusual.

All but one of the remaining shaded cells involve chromosome 4. Within chromosome 4, the lines for the independent realizations of correlation from the JAX and UCLA data are much closer together on the kernel density estimate than in other cells. They are almost identical between the experiments. This consistent departure of markers on chromosome 4 from the expectations of the model therefore suggests that chromosome 4 may experience stronger cross over interference than chromosome 1 and 18. Chromosome 4 is therefore noteworthy for the poorer fit of the model to its correlations and the consistency of these correlations over independent experiments, observations supported entirely by the package `toyGenomeGen`.

5.8 Conclusion

This chapter is not meant to make the reader an expert in genetics. The model of genetic measurement and derivation of the Haldane map distance were simple and limited. Primarily, they have been included because genetics is a frequent setting for the application of pairwise measures of association to detect interesting relationships. Creating a clear and simple model that (at least somewhat) reflects real data and implementing a toy version of that in `R` facilitates easier discussion later.

Such explorations are not without new tools, however. The `toyGenomeGen` package provides a simple way to test different feed-forward models of genetics, and so to participate in recent work moving away from classic map distances (Veller et al., 2020; Kivikoski et al., 2023). `toyGenomeGen` forces no particular choice of location structure of cross function, and so could be adapted to generate distributions under settings using other measures of distance on a genome.

As well, the straightforward context motivated a novel take on the scatterplot matrix that takes advantage of the mirrored cells to display both test and distributional information. These plots are well-suited to visualize how an observed matrix compares to a simulated distribution of matrices, without the need to reduce either to a univariate summary.

Chapter 6

Example Application: Monogenic and Oligogenic Traits

The χ^2 pooling method of Section 4 can be applied to the genetic model from Section 5 to relate a trait to genetic information on N individuals. Recall the last step of the schematic in Figure 5.1, where annotated genetic information is encoded numerically and summarized into a single value at each marker. The problem is associating a matrix of summarized encodings

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M],$$

where $\mathbf{z}_m = (z_{m1}, z_{m2}, \ldots, z_{mN})^\mathsf{T}$ gives the summarized encoding of marker m for each individual, to a trait vector $\mathbf{y} = (y_1, \ldots, y_N)^\mathsf{T}$ providing an experimental measurement for each individual. Let the theoretical correlation matrix of $\mathbf{Z}$ be $\mathbf{\Sigma} = [\rho_{ij}]$ where $\rho_{ii} = 1$ and ρ_{ij} is given by Equation 5.18, that is

$$\rho_{ij} = I_{c_j}(c_i)\,\gamma e^{-2\beta d(i,j)}$$

where γ is a constant taking values outlined in Section 5.3. Commonly, $\gamma = 1$.

When $\mathbf{y}$ is realized from $\mathbf{Z}$, two patterns are possible. The trait can either be oligogenic, arising from genetic information encoded in numerous genes or marker locations, or monogenic, arising from the genetic code in a single gene or even at a single marker. A complete list of identified monogenic traits and conditions in humans can be found at Online Mendelian Inheritance in Man (OMIM), and include some kinds of albinism and cystic fibrosis. Common examples of oligogenic traits in humans include eye colour, hair colour, and height.

These two types of association between genotype and phenotype can be simply conceptualized as linear models. That is, we assume

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\intercal$ is a constant vector and $\boldsymbol{\epsilon}$ is a random vector accounting for unmeasured sources of variation such as environmental factors. If $\mathbf{y}$ is monogenic, then $\beta_m = 0$ for all but one index $m \in \{1, \ldots, M\}$. If $\mathbf{y}$ is oligogenic, then multiple entries in $\boldsymbol{\beta}$ are non-zero. Fitting linear models of this type for genomic studies is computationally challenging due to the large size and high collinearity of $\mathbf{Z}$, and so a pairwise significance test t can be applied to every $m \in \{1, \ldots, M\}$ to give M statistics $t_m = t(\mathbf{y}, \mathbf{x}_m)$ and M p-values $p_m = P(t(X, Y) \geq t_m | Cov(X, Y) = 0)$. By taking only those t_m (p_m) greater than (less than) a threshold, the columns of $\mathbf{Z}$ can be filtered in advance to reduce the computational burden by identifying the most promising markers. The null hypothesis is assumed to be

$$H_0 : p_1, \ldots, p_M \overset{\text{iid}}{\sim} U(0, 1)$$

as before.

Central and marginal rejection and the centrality quotient are useful concepts to guide our approach to this problem. If $\boldsymbol{\Sigma} = \mathbf{I}$, a monogenic trait will give uniform p-values for all of $\mathbf{p} = (p_1, \ldots, p_M)^\intercal$ except for one single p-value which is biased towards zero. In contrast, many of the p-values in $\mathbf{p}$ should be biased to small values for a oligogenic trait. Therefore, pooled p-values with small centrality quotients will be powerful in the identification of monogenic traits and pooled p-values with large centrality quotients will powerfully identify oligogenic traits. Performing a sweep of κ values in $chi(\mathbf{p}; \kappa)$ provides distinctive curves that distinguish these cases under independence, as in Section 4.7.1 for mixtures of beta distributions.

Of course, we cannot assume independence in genetics. Markers follow the correlation structure outlined in Chapter 5, that of a block diagonal matrix with decreasing correlations away from the diagonal within each block. Applying the rejection rule $chi(\mathbf{p}; \kappa) \leq \alpha$ only controls the family-wise error rate (FWER) at α under independence, for dependent data this control is not guaranteed. Addressing this problem is critical to controlling the FWER in genetic data.

6.1 Independent genetic data

To demonstrate the different curves of $chi(\mathbf{p}; \kappa)$ in κ for monogenic and oligogenic traits under independence, the JAX BSB data set from the `toyGenomeGenR` package (explored in detail in Section 5.7) is used as a model genome from which traits are generated. Briefly, it records the annotations of genetic markers at 1496 locations in 94 mice of the same inbred strain. For simplicity, traits are generated using the linear model of Equation 6.1 and $\boldsymbol{\beta} \in \{0,1\}^M$ with $\epsilon_1, \ldots, \epsilon_M \overset{\text{iid}}{\sim} N(0, 0.3)$. That is, the trait either depends on a marker or does not, there is no difference in the coefficients between marker sites that are related to the trait. To obtain independent markers within the data, the markers nearest to the midpoint of each chromosome of each genome are selected, giving 20 marker encodings for each of the 94 individuals. Figure 6.1 displays an image of the sample correlation matrix with no patterns present beyond noise, confirming the selected sites are uncorrelated in this sample.

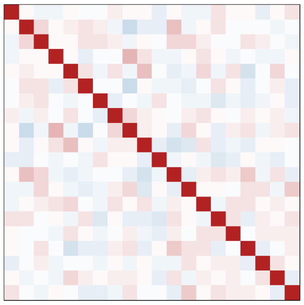
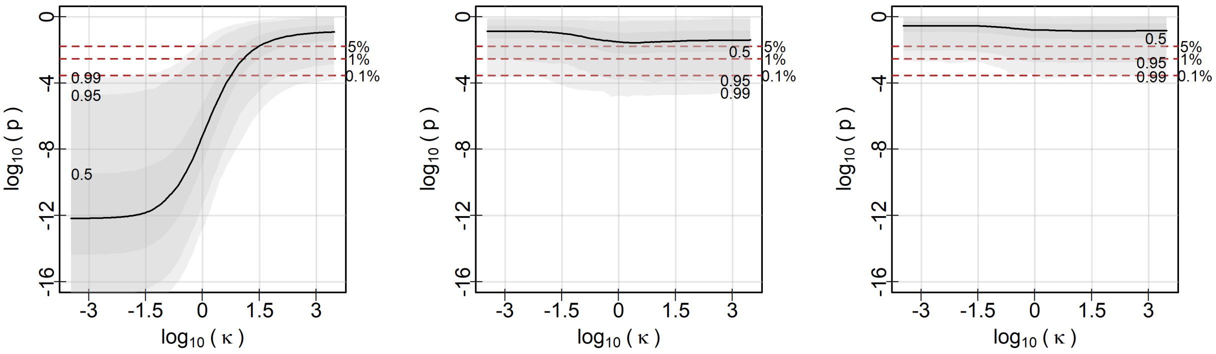


Figure 6.1: The sample correlation matrix of the selected midpoint markers on each chromosome of the JAX BSB data. These are consistent with underlying model correlations of zero.

For $\eta = 0.05, 0.25, 0.5, 0.75$, and 1, the first $M\eta$ markers are used to generate $\mathbf{y}$ independently 1000 times. Each marker vector $\mathbf{z}_i$ is then tested against $\mathbf{y}$ to generate 20 p-values for each repetition using random recursive binning with a stop criterion limiting the depth to two as in Section 3. Values of κ were selected at increments of 0.1 on the natural log scale over the range -8 to 8, and a sweep of all κ values was performed for each sample of 20 p-values. The resulting curves for $\eta = 0.05, 0.5$, and 1 are shown in Figure 6.2.

Consistent with the known alternative densities in Section 4.7.1, increasing η increases the κ that minimizes the curve. Unlike the previous explorations, however, the KL divergence of the distribution of p-values is no longer controlled and instead depends on the distribution of p-values under the model. An unfortunate consequence of this is that the value of $chi\left(\mathbf{p}; \kappa\right)$ increases for all κ as η increases. As the test of H_0 is based on thresholding $chi\left(\mathbf{p}; \kappa\right)$, this suggests the power of H_0 decreases as η increases for this linear model.



(a)

(b)

(c)

Figure 6.2: Pooled p-values for pairwise tests of independent markers against traits by κ when (a) $\eta = 0.05$, (b) $\eta = 0.5$, and (c) $\eta = 1$. Red horizontal lines indicate the corresponding null quantiles of the minimum of the κ curves. The regions where $chi\,(\mathbf{p};\kappa)$ is smallest are as expected: increasing η increases the prevalence of evidence across all tests and therefore the κ which minimizes the curve.

This makes sense intuitively, as any individual marker will be a poorer predictor of the trait when the trait is oligogenic and so all will produce weaker evidence, equivalently p-values less biased towards zero, when tested pairwise. Detecting such a subtle bias from a relatively small sample of $M = 20$ is challenging.

That said, the shapes of the $chi\,(\mathbf{p};\kappa)$ curve still differ in η and so provide information about the alternative hypothesis generating the data. This is especially true when combined with Figure 4.16 to link the minimum κ with alternatives in the η, KL divergence plane. Figure 6.2(a) shows a majority of $chi\,(\mathbf{p};\kappa)$ curves have a clear minimum at small κ, suggesting the plausible alternatives can be seen in Figure 4.16(d) and correctly implying the alternative has a small η value. Figure 6.2(b) shows a curve of $chi\,(\mathbf{p};\kappa)$ which tends to have has its minimum near moderate κ values, suggesting η near 0.5 are most likely by using Figure 4.16(b). Finally, Figure 6.2(c) tends to have a minimum for large κ, suggesting correctly that η is larger than 0.6 using Figure 4.16(a). Even in this final, weakest, case $chi\,(\mathbf{p};\kappa)$ changes noticeably in κ. The distribution of $chi\,(\mathbf{p};10^{-3})$ effectively matches the null quantiles while $chi\,(\mathbf{p};10^{3})$ is biased to smaller values. Therefore, the shape of the curve still provides useful information and taking the minimal κ gives an indication of the region of plausible alternatives for (b) and (c), though it is less compelling than for (a).

6.2 Dependent genetic data

In practice, markers of interest are commonly on the same chromosome and so dependence cannot be ignored. The impact of this dependence was investigated in the JAX BSB data by applying the same linear model to generate traits using 10 marker annotations evenly spaced along each of chromosomes one and two. This keeps the total number of markers fixed at $M = 20$ but gives the block correlation structure seen in Figure 6.3 that is typical of genetic marker data.

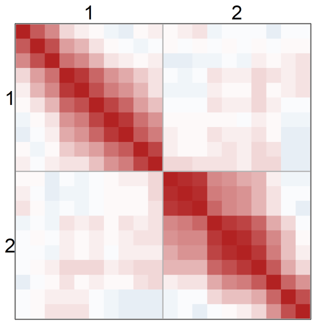
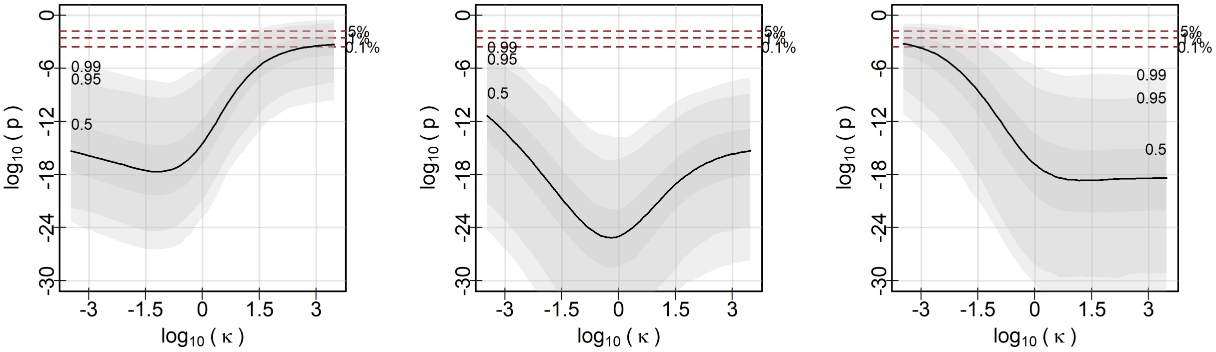


Figure 6.3: Correlations for 20 markers sampled evenly across chromosomes one and two in the JAX BSB data, grouped and labelled by chromosome.

Repeating the generation method from the independent case, curves of $chi\,(\mathbf{p}; \kappa)$ by κ can be constructed which naively make no adjustment to account for the dependence known in advance. Figure 6.4 displays these curves and the impact of dependence is clear: $chi\,(\mathbf{p}; \kappa)$ is drastically smaller than in the independent case at every η for every κ. Despite this, the minimizing κ for the curves in Figure 6.4 are similar to those of Figure 6.2. Indeed, the conclusions of these plots change very little with the introduction of dependence and the curves still correctly identify η. The shape of these curves and their minima seem to be accentuated by the dependence if anything, giving an even stronger signal of the η generating the p-values.

Though the relative shape of these curves still provides useful information, their much smaller values compared to the independent case suggest a problem. Thresholds developed in the independent case (such as the quantiles plotted with dashed red lines) no longer provide a meaningful indication of the significance of the curve shapes. Some adjustment therefore has to be made so these thresholds remain meaningful under dependence or to account for dependence when generating the curves.



(a)

(b)

(c)

Figure 6.4: Pooled p-values for pairwise tests of dependent markers aginst by κ without adjustment when (a) $\rho = 0.05$, (b) $\rho = 0.5$, and (c) $\rho = 1$. Despite the introduction of dependence, the curves give the same conclusions as the independent case.

6.3 Adjusting for dependence

A direct approach estimates the appropriate threshold using permutations, the method of choice to obtain p-values under dependence in Conneely and Boehnke (2007), Han et al. (2009), and Cinar and Viechtbauer (2022). Just as was done to generate the null quantiles for the minimum of the $chi\,(\mathbf{p};\kappa)$ curve by κ, many repeated cases of dependent $\mathbf{Z}$ which are unrelated to $\mathbf{y}$ could be generated and used to construct $chi\,(\mathbf{p};\kappa)$ curves. There are two ways to generate these examples. The first simulates H_0 by shuffling the phenotype measurements in $\mathbf{y}$ and computing $\mathbf{t}$ against the unshuffled columns of $\mathbf{Z}$ many times. This breaks any relationships present between $\mathbf{Z}$ and $\mathbf{y}$ without changing the characteristics of either or assuming any parametric distribution, but requires observed data. A second method would generate many representative $\mathbf{Z}$ using, for example, the machinery of `toyGenomeGenR` and $\mathbf{y}$ independently from some presumed underlying distribution. Unlike the first, this could be done before any data is observed.

In either case, the number of possible permutations grows rapidly in N, requiring the generation of many $\mathbf{Z}$ and $\mathbf{y}$ pairs to explore the space of possible permutations. Implementation therefore imposes a considerable computational burden (Han et al., 2009; Cinar and Viechtbauer, 2022), and proposed remedies still require many thousands of permutations[1]. Rather than spend such computational effort to determine the correct threshold for a preliminary aspect of analysis that only indicates whether H_0 is false, other approaches choose to approximate the distribution of pooled p-values under dependence.

One group of methods does this by computing m_{eff}, the “effective number of tests” present in M correlated tests of significance, using functions of the eigenvalues of $\mathbf{\Sigma}$. The statistic associated with a pooled p-value is then scaled by $\frac{m_{eff}}{M}$ before using the independent distribution function to compute its p-value (Cheverud, 2001; Nyholt, 2004; Li and Ji, 2005; Galwey, 2009). While they make no distributional assumptions, these adjustments are ad hoc and often provide anti-conservative adjustments that fail to control the FWER at the nominal level (Salyakina et al., 2005; Cinar and Viechtbauer, 2022). In light of this, estimating the effective number of tests is an adjustment best avoided for handling dependence.

Assuming that $\mathbf{t}$ are from a multivariate normal distribution with mean zero and correlation matrix $\mathbf{\Sigma}$, Conneely and Boehnke (2007), Han et al. (2009), and Cinar and Viechtbauer (2022) provide adjustments based on approximate normal integrals. This is motivated, in part, by the asymptotic normality of many test statistics in genetics. Under this assumption, a p-value can be computed by an appropriate normal integral scaled for better agreement with the empirical distribution. Though these methods are accurate, they give a joint p-value for the vector $\mathbf{t}$, making their application to the curves of a pooled p-value that is a function of $\mathbf{t}$ less straightforward. The integrals could still be computed for $chi\,(\mathbf{p};\kappa)$, but a simpler and faster option exists, the method of Brown (1975) along with its refinements in Yang et al. (2016), Poole et al. (2016), and Cinar and Viechtbauer (2022).

Developed specifically for the pooled p-value $Fis(\mathbf{p}) = chi\,(\mathbf{p}; 2)$, this first computes the covariance between $\ln p_i$ and $\ln p_j$ given ρ_{ij} and the distribution of the underlying statistics $\mathbf{t}$. By assuming $\mathbf{t} \sim MVN(0, \mathbf{\Sigma})$, this can be simulated and approximated by a polynomial (Brown, 1975; Yang et al., 2016), computed directly by the appropriate bivariate normal integral (Cinar and Viechtbauer, 2022), or approximated by the empirical CDF (Poole et al., 2016). Next, the approximation of Satterthwaite (1946) is applied to match the first two moments of the sum of correlated $\ln p_i$ values to $c\chi^2_k$ where c is positive scaling constant. This method is faster and simpler than permutation tests while still giving approximately correct p-values when applied to a linear model in simulation tests performed by Yang et al. (2016).

[1] Knijnenburg et al. (2009), for example, model the tail of permuted statistics with a generalized Pareto distribution estimated by maximum likelihood and still require thousands of simulations to obtain a good approximation.

6.3.1 Modifying Brown's method

To generalize the method of Brown (1975) for any κ in $chi\,(\mathbf{p};\kappa)$ and p-values resulting from non-normal statistics, apply Satterthwaite's moment-matching approximation to

$$l_\chi(\mathbf{p};\kappa) = \sum_{m=1}^{M} F_\chi^{-1}(1-p_m;\kappa)$$

with a dependent p_i and p_j. Explicitly, this assumes $l_\chi(\mathbf{p};\kappa) \sim c\chi_d^2$ and so

$$E\big[l_\chi(\mathbf{p};\kappa)\big] = E\left[c\chi_d^2\right]$$

and

$$Var\big(l_\chi(\mathbf{p};\kappa)\big) = Var\left(c\chi_d^2\right)$$

to give

$$c = \frac{Var\big(l_\chi(\mathbf{p};\kappa)\big)}{2E\big[l_\chi(\mathbf{p};\kappa)\big]}$$

and

$$d = \frac{2\left(E\big[l_\chi(\mathbf{p};\kappa)\big]\right)^2}{Var\big(l_\chi(\mathbf{p};\kappa)\big)}.$$

Using these constants, the adjusted pooled p-value is computed as

$$1 - F_\chi\left(\frac{1}{c}l_\chi(\mathbf{p};\kappa);d\right).$$

The expectation, $E\big[l_\chi(\mathbf{p};\kappa)\big] = M\kappa$, is unchanged by dependence, but the variance becomes

$$Var\big(l_\chi(\mathbf{p};\kappa)\big) = \sum_{i=1}^{M} Var\big(F_\chi^{-1}(1-p_i;\kappa)\big) + \sum_{i=1}^{M}\sum_{j\neq i} Cov\big(F_\chi^{-1}(1-p_i;\kappa), F_\chi^{-1}(1-p_j;\kappa)\big).$$

Regardless of the dependence present between variables, $Var\big(F_\chi^{-1}(1-p_i;\kappa)\big) = 2\kappa$ while the cross terms must be estimated. Noting that the variance is not affected by dependence, define

$$r_{ij} = Cor\big(F_\chi^{-1}(1-p_i;\kappa), F_\chi^{-1}(1-p_j;\kappa)\big)$$

and then adjustment for dependence requires estimating

$$2\kappa r_{ij} = Cov\big(F_\chi^{-1}(1-p_i;\kappa), F_\chi^{-1}(1-p_j;\kappa)\big)$$

based on ρ_{ij} for the markers generating test statistics t_i and t_j from which p_i and p_j are computed. Expanding the covariance expression gives

$$E\left[F_\chi^{-1}(1 - F_t(t_i); \kappa) F_\chi^{-1}(1 - F_t(t_j); \kappa)\right] - E\left[F_\chi^{-1}(1 - F_t(t_i); \kappa)\right] E\left[F_\chi^{-1}(1 - F_t(t_j); \kappa)\right]$$

where F_t is the CDF of t. The univariate expectations are straightforward to evaluate, as $1 - F_t(t_i)$ is uniformly distributed so

$$E\left[F_\chi^{-1}(1 - F_t(t_i); \kappa)\right] = \kappa.$$

Substituting this and the joint distribution of t_i and t_j into the expression for the covariance gives

$$2\kappa r_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_\chi^{-1}(1 - F_t(t_i); \kappa) F_\chi^{-1}(1 - F_t(t_j); \kappa) f(t_i, t_j) dt_i dt_j - \kappa^2$$

where $f(t_i, t_j)$ is the joint density of t_i and t_j.

At this point, the methods of all of Yang et al. (2016), Poole et al. (2016), and Cinar and Viechtbauer (2022) would compute the integral by assuming

$$\begin{bmatrix} t_i \\ t_j \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{bmatrix}\right)$$

and evaluating it numerically. Both Yang et al. (2016) and Poole et al. (2016) then use polynomials to summarize the relationship between r_{ij} and ρ_{ij}, while Cinar and Viechtbauer (2022) instead provide a look-up table in their R software package.

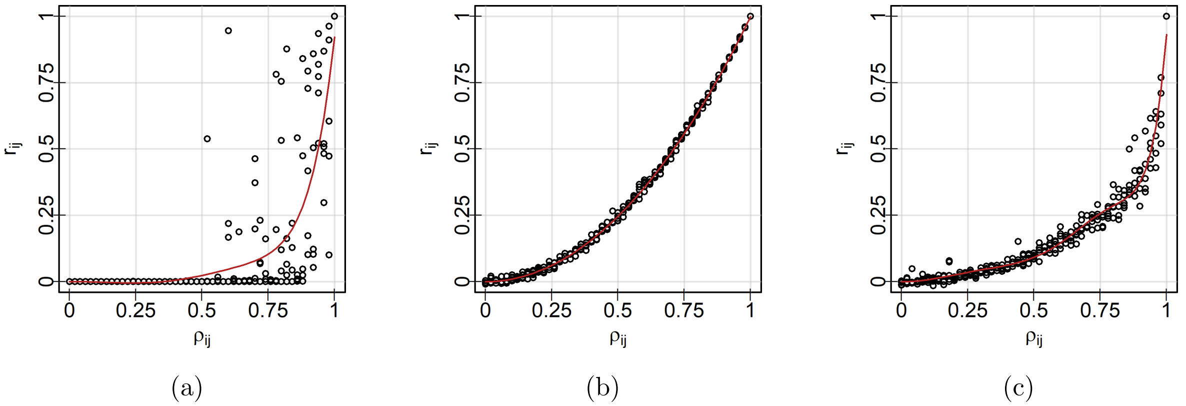
To address the non-normality of the statistic presented in Section 3, Monte Carlo integration of simulated genetic data is used to estimate r_{ij} as a function of ρ_{ij} rather than assuming the distribution of t_i, t_j. Following Yang et al. (2016), the relationship between the two uncovered by Monte Carlo integration is summarized by a tenth-order polynomial for each κ value. This uses Monte Carlo integration to generalize the method of Brown (1975) to non-normal statistics $\mathbf{t}$ pooled with arbitrary $chi\left(\mathbf{p}; \kappa\right)$.

Specifically, for each $\rho_{ij} \in \{0, 0.02, 0.04, \ldots, 1\}$ and $\ln \kappa \in \{-8, -7.9, -7.8, \ldots, 8\}$, matrices of correlated marker encodings $\mathbf{Z} \in \{0, 1\}^{94 \times 2}$ are generated 10,000 times using simplified code from the `toyGenomeGenR` package[2]. As the goal of the simulation is to determine the null relationship of ρ_{ij} and r_{ij}, both of the simulated markers $\mathbf{z}_i, \mathbf{z}_j$ are compared to an independently generated trait $\mathbf{y} = (y_1, \ldots, y_{94})^\intercal$ with $y_1, y_2, \ldots, y_{94} \overset{\text{iid}}{\sim} N(0, 0.3)$ using random recursive binning with a depth limit of 2 as described in Chapter 3

[2] This code removed structures, names, and formatting for the sake of increased efficiency.

to keep the computation time of the simulation relatively short. The obtained p-values p_i and p_j are transformed to $F_\chi^{-1}(1-p_i;\kappa)$ and $F_\chi^{-1}(1-p_j;\kappa)$ and the correlation between the transformed values r_{ij} is computed for each κ and ρ_{ij}. To obtain repeated observations at each point, this entire procedure was replicated 5 times and tenth-order even polynomials predicting r_{ij} from ρ_{ij} using these simulated observations are fit using least squares for each κ.

The exclusion of odd polynomial terms follows Yang et al. (2016) and makes sense in the case of simulated genetic data. Negative correlation between $\mathbf{z}_i$ and $\mathbf{z}_j$ does not reflect a different pattern of measuresments, but is instead a consequence of the arbitrary choice of encoding. Simply flipping the encoding for each marker from 0 to 1 and 1 to 0 gives the same magnitude of correlation with opposite sign (see Section 5.3). Recursive binning as a measure of association will give identical p-values for both the positive and negative cases as the marginal pattern of values is the same, and this symmetry implies only even functions need to be considered. Plots of simulated r_{ij} and ρ_{ij} for several κ are shown in Figure 6.5.



(a)

(b)

(c)

Figure 6.5: Correlation between $F_\chi^{-1}(1-p_i;\kappa)$ and $F_\chi^{-1}(1-p_j;\kappa)$ by ρ_{ij} when (a) $\log_{10}(\kappa) = -3.5$, (b) $\kappa = 1$, and (c) $\log_{10}(\kappa) = 3.5$ with tenth-order even polynomials fit by least squares plotted in red over top. More variation is observed for small and large κ values than for moderate ones.

The fitted polynomials corresponding to the red lines drawn on Figure 6.5 are summarized in Table 6.1. As described above, each is a tenth-order even polynomial and so takes the form

$$r_{ij} = c_1\rho_{ij}^2 + c_2\rho_{ij}^4 + c_3\rho_{ij}^6 + c_4\rho_{ij}^8 + c_5\rho_{ij}^{10},$$

the table reports these coefficients for $\kappa = 10^{-3.5}, 1, 10^{0.3} \approx 2$, and $10^{3.5}$. For moderate κ, the small coefficients for larger exponents suggest that $r_{ij} \approx \rho_{ij}^2$. Large and small κ have much more complicated polynomials.

κ	c_1	c_2	c_3	c_4	c_5
$10^{-3.5}$	-0.18	1.92	-4.05	3.14	0.09
1	0.98	0.05	-0.06	0.03	0.00
2.0	1.00	-0.02	0.14	-0.25	0.13
$10^{3.5}$	0.79	-4.78	18.03	-26.18	13.07

Table 6.1: Polynomial model coefficients relating r_{ij} to ρ_{ij} for several κ values

To determine where these polynomials provide a reasonable approximation, a plot of the coefficient of determination R^2 of the models by $\log_{10}(\kappa)$ was produced in Figure 6.6. As demonstrated in the example plots of Figure 6.5, the large variation for small κ values leads to large residuals and poor performance of the polynomial model while moderate κ values are predicted essentially perfectly. The poor model fit due to the increased variation for extreme κ values suggests that the estimation of r_{ij} from ρ_{ij} is less accurate in these cases, and so the adjusted p-value will be as well.

The same pattern is repeated in the findings of Section 7.4, where the evidential estimate is much less variable in simulated data sets when generated using $chi\,(\mathbf{p}; \kappa)$ with a moderate κ value. Both observations are related to the behaviour of $F_\chi^{-1}(p; \kappa)$ for extreme κ.

Observe that Figure 6.5(a) has $r_{12} = 0$ for all 5 repeated measurements whenever $\rho_{12} < 0.5$ and recall from Section 4.6 that when $\kappa \approx 0$ the CDF $F_\chi(x; \kappa)$ approximates a step function at $x = 0$. This means the quantile function can be thought of as

$$F_\chi^{-1}(1 - p; \kappa) \approx \begin{cases} \infty & \text{when } p \geq 1 - \epsilon \\ 0 & \text{when } p < 1 - \epsilon, \end{cases}$$

for $\epsilon > 0$ which can be made arbitrarily small by taking κ arbitrarily close to zero. Very small differences in p_i and p_j are therefore magnified when they are converted by the χ_κ^2 quantile function, weakening their linear dependence. Only for the strongest relationships where $p_j \approx p_j$ is this effect negligible.

Though still more variable than moderate κ, the increased variation is less severe for large κ, compare Figure 6.5(a) to Figure 6.5(c). This is explained by a quirk of the p-values computed. Occasionally, data are simulated which, by chance, do not differ from

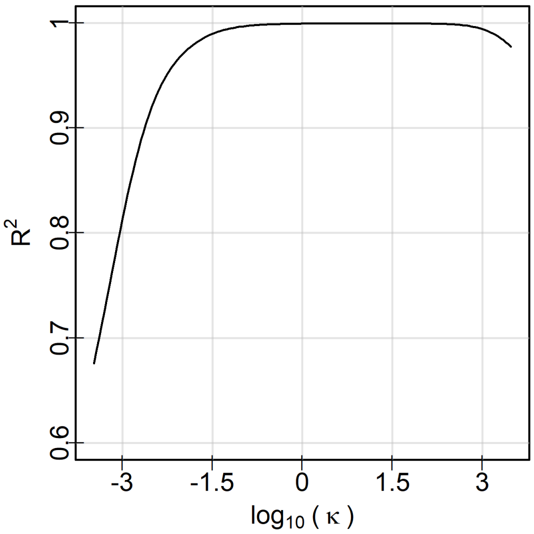


Figure 6.6: The polynomial model coefficients of variation by $\log_{10}(\kappa)$. For κ values between $10^{-1.5}$ and 10^{3}, the models achieve excellent fits.

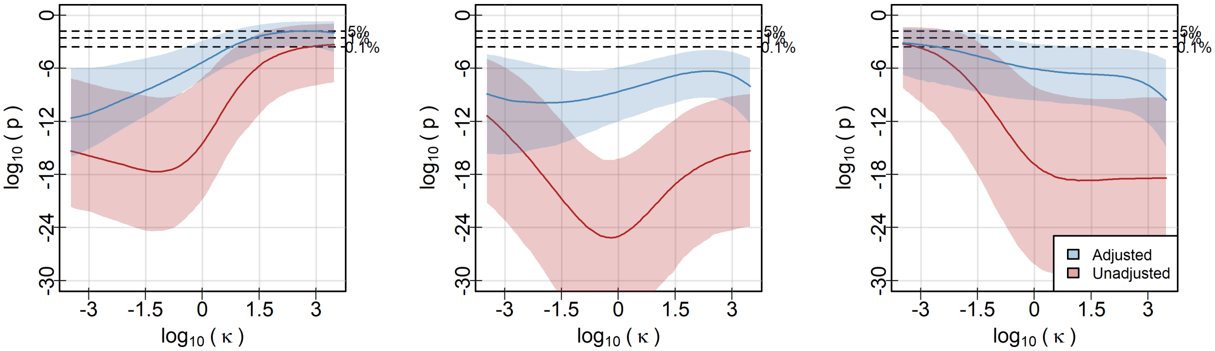
expectation at all when split using the recursive binning algorithm. In this case, $t_i = 0$ or $t_j = 0$ and so $p_i = 1$ or $p_j = 1$ which means $F_\chi^{-1}(1 - p_i; \kappa) = 0$ or $F_\chi^{-1}(1 - p_j; \kappa) = 0$ for all κ. If this occurs for only one of $\mathbf{z}_i$ or $\mathbf{z}_j$, the other statistic will have a very large p-value that is not quite 1, call it $1 - \delta$, and so its transformed quantile will be strictly greater than zero. As κ increases, the difference between $F_\chi^{-1}(0; \kappa) = 0$ and $F_\chi^{-1}(\delta; \kappa)$ grows as the expected value of χ^2_κ grows, pulling the δ quantile away from zero for any $\delta > 0$. This increasing difference in the case of the random agreement of expected and observed counts in the recursive binning algorithm is thereby responsible for the poorer fit as κ increases.

Nonetheless, the coefficient of determination in Figure 6.6 is never less than 0.65 and is greater than 0.95 for most of the range of κ values. This means a majority of the variation is explained in every case and only a small proportion remains unexplained in the best cases. With these polynomials, the correlation matrix between markers $\mathbf{\Sigma}$ can be converted into approximate covariances between $F_\chi(1 - p_i; \kappa)$ and $F_\chi(1 - p_j; \kappa)$ for $\log_{10} \kappa \in [-1.5, 3]$, supporting the Satterthwaite approximation of the distribution of $l_\chi(\mathbf{p}; \kappa)$ to adjust for dependent p-values.

6.3.2 Applying the adjustment

Recall the curves of *chi* $(\mathbf{p}; \kappa)$ by κ for dependent p-values in Figure 6.7 and for independent p-values in Figure 6.2. Under dependence, the curves had nearly universally smaller values than in the independent case, suggesting exaggerated significance to the results. The preceding method can now be used to adjust the dependent curves to bring their significance levels closer to the independent curves.

First, the observed correlation matrix of the genetic data that generated these p-values is used to estimate the covariance of $F_\chi^{-1}(1-p_i; \kappa)$ and $F_\chi^{-1}(1-p_j; \kappa)$ based on the polynomials describing the results of Monte Carlo integrations spanning the range of correlations. Given the covariance between $F_\chi^{-1}(1-p_i; \kappa)$ and $F_\chi^{-1}(1-p_j; \kappa)$, a Satterthwaite approximation of the distribution of $l_\chi(\mathbf{p}; \kappa) = \sum_{m=1}^{M} F_\chi^{-1}(1-p_m; \kappa)$ is used to compute adjusted p-values. Repeating this process for each κ value leads to an adjusted curve. The central 95% quantiles and medians of the adjusted curves and original curves are plotted in Figure 6.7.



(a)

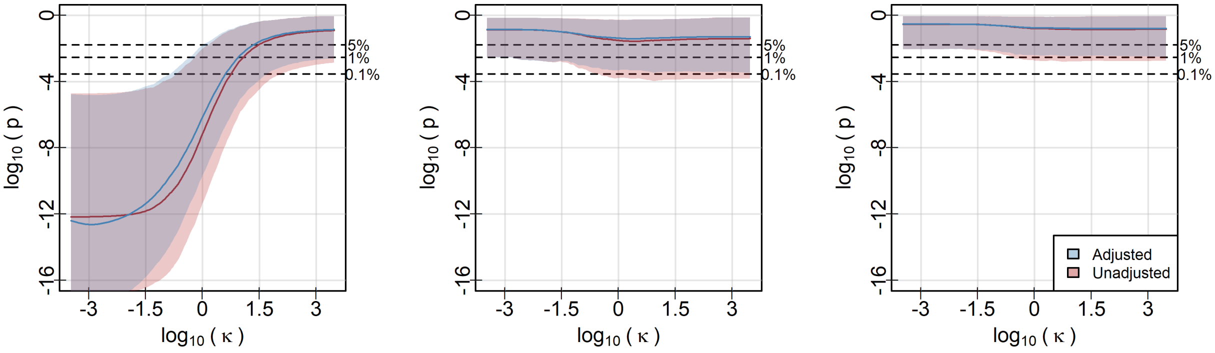
(b)

(c)

Figure 6.7: Central 0.95 quantiles and medians of adjusted and unadjusted pooled p-values by κ when (a) $\eta = 0.05$, (b) $\eta = 0.5$, and (c) $\eta = 1$ for markers spread across two chromosomes. The adjustment has increased the level of the curves and made their relative peaks and troughs less pronounced.

As desired, the adjustment for dependence increases the pooled p-value at every κ for every η, and so seems to have accounted for the dependence in the p-values. Additionally, the conclusions of the curves have not changed drastically, suggesting that the minimum of the κ curve is not affected by this adjustment. Still, these curves do not look like those from the independent case, and so the impact of dependence has not been completely removed. It is probably the case that it is easier to detect any association between $\mathbf{Z}$ and $\mathbf{y}$ when there is dependence present, as dependence causes the related tests to become more significant together, making the p-values all smaller as a group.

To ensure that the method does not spuriously change independent curves, the sample correlation matrix displayed in Figure 6.1 was also used to adjust κ curves for the independent data. The central 95% quantiles and median are displayed in Figure 6.8 for both the adjusted and unadjusted curves. Applying this method barely changes the quantiles or median, suggesting it behaves correctly and does not change the p-values in the independent case.



(a)

(b)

(c)

Figure 6.8: Central 0.95 quantiles and medians of adjusted and unadjusted pooled p-values by κ when (a) $\eta = 0.05$, (b) $\eta = 0.5$, and (c) $\eta = 1$ for the case of independent markers. The adjustment barely impacts the shape and value of the curve under independence.

6.4 Real genotype and phenotype data

To test this method on real data, the ideal would be a set of individuals with their genotypes and phenotypes both measured. However, data of this kind is hard to come by and typically requires an application or association with an institution with access. This limits the ability of researchers to test new methods.

In order to circumvent this, an improvised paired genotype and phenotype dataset was created from the Mouse Phenome Database (MPD), a public repository of genotype and phenotype data for inbred strains of mice (Bogue et al., 2020). These inbred strains are effectively genetically identical after hundreds of generations of inbreeding, and so sequencing data from an individual can be treated as the generic sequence data for every individual within a strain. Downloading and using this data requires no proposal or account and so serves as a useful test case. To demonstrate its use, the UCLA SNP panel (Eskin, 2023) was joined to the blood serum data of Svenson et al. (2007), both downloaded using the MPD API.

The blood serum data measures plasma lipids (e.g. cholesterol, triglycerides) in 10 week old mice that had been fed a standard laboratory mouse diet since weaning with a distribution by inbred strain and sex summarized in Table 6.2. Briefly, an average of 26 mice from each of 43 different inbred strains roughly balanced over the sex of the mice are recorded, though certain strains have many more observations than others. The blood serum measurements are not the final goal of the study; at 10 weeks the mice were given a diet conducive to the formation of plaque within arterial walls in order to determine the effectiveness of lab mice as a model of human plaque formation. The data explored here are merely the baseline.

Table 6.2: Distribution of mice in the blood serum data of Svenson et al. (2007) by sex and strain identifier. The data are roughly balanced by sex, though the strains have very different numbers of observations.

Jackson Lab Strain ID	Females	Males	Total
PL/J	27	23	50
RF/J	23	21	44
AKR/J	29	14	43
129S1/SvImJ	24	18	42
BALB/cByJ	21	21	42
C57BL/10J	30	10	40
C57L/J	15	22	37
C57BL/6J	25	11	36
C58/J	10	20	30

Continued on next page

Table 6.2 – continued from previous page

Jackson Lab Strain ID	Females	Males	Total
NOD/ShiLtJ	10	20	30
MOLF/EiJ	14	15	29
RIIIS/J	13	16	29
LP/J	17	11	28
DBA/1J	11	16	27
C3H/HeJ	11	15	26
CAST/EiJ	12	14	26
FVB/NJ	10	16	26
SWR/J	10	16	26
I/LnJ	12	13	25
NON/ShiLtJ	10	15	25
SM/J	14	11	25
C57BR/cdJ	11	13	24
CZECHII/EiJ	11	13	24
MSM/MsJ	12	12	24
PWK/PhJ	11	13	24
DBA/2J	10	13	23
A/J	10	12	22
C57BLKS/J	10	12	22
MA/MyJ	11	11	22
NZB/BlNJ	12	10	22
NZW/LacJ	12	10	22
BTBR T^+ $Itpr3^{tf}$/J	11	10	21
BUB/BnJ	11	10	21
CE/J	11	10	21
JF1/MsJ	9	12	21
SPRET/EiJ	11	10	21

Continued on next page

Table 6.2 – concluded from previous page

Jackson Lab Strain ID	Females	Males	Total
CBA/J	10	10	20
KK/HlJ	9	11	20
SJL/J	10	10	20
WSB/EiJ	10	9	19
SEA/GnJ	8	10	18
PERA/EiJ	8	9	17
BALB/cJ	0	5	5
Total	566	573	1139

After joining the UCLA SNP panel data to the blood serum data, the joined genotype-phenotype data are cleaned to remove mostly incomplete markers and markers with no variation. The format and limits of API extraction of the data organizes markers into groups corresponding to each API request, and so a subsample from these groups is taken as they roughly span the genome. The SNP with the most complete data is selected within each request group, and the others are dropped to give a final count of 1052 SNPs spanning all 20 chromosomes measured for each of the 1139 mice.

Within strains, the X/Y chromosome (which controls sex in mice in the same way as in humans) is a source of genetic variation that must be controlled. To avoid complications that result from this chromosome, the data are further restricted to contain only the 550 female mice at all 1052 locations. The measurement of total lipoproteins that are not high density (non-HDL) is selected as the target trait, and each SNP is tested against it using a recursive binning test restricted to a depth of two to match earlier simulations. The resulting p-values are adjusted using the modification of Brown's method outlined earlier in this section. Figure 6.9 displays the $chi\,(\mathbf{p};\kappa)$ curves adjusted using the theoretical correlation of the SNPs computed according to their centiMorgan distances as in Chapter 5. Theoretical correlation is used instead of an observed correlation matrix because the genotype matrix is artificially generated by repeating entries for each strain from a different data set.

The adjusted curve in Figure 6.9 increases from a minimum pooled p-value of roughly $10^{-3.4}$ at the smallest κ to a maximum of 1 for all $\kappa > 0.3$. Indeed, the majority of κ values in Figure 6.9 lie above the null 5% cutoff and only those less than $\kappa \approx 10^{-0.75}$ are below it. Comparing this to the adjusted curves from the simple linear models tested earlier suggests that only a small proportion of markers are associated with the concentration of total non-HDL in mouse blood serum.

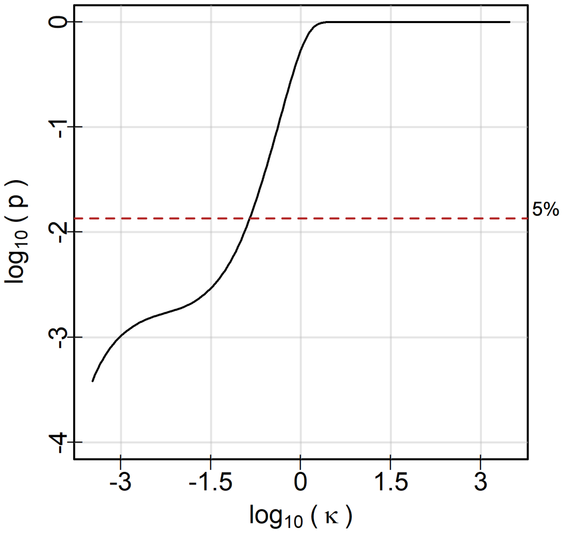


Figure 6.9: The adjusted $chi\,(\mathbf{p};\kappa)$ curves for the female measurements of total concentration of non-HDL in blood serum. The curve is below the 5% threshold and slowly increasing for all $\log_{10}\kappa < -0.75$, with a flatter section between $\kappa = 10^{-3}$ and $10^{-1.5}$.

More precisely, the simulations outlined in Section 4.7.2 can be used to identify the plausible region of alternative hypotheses to H_0 that corresponds with the curve in Figure 6.9. In this section, 100,000 examples of samples of p-values following an alternative to H_0 were generated with a set proportion of non-null p-values (η) following a beta distribution with a particular Kullback-Leibler divergence from uniform ($D(a,w)$). The power of different κ in $chi\,(\mathbf{p};\kappa)$ to detect these alternatives was computed and summarized across the range of possible beta alternative parameter values to identify the regions in η and $D(a,w)$ where each κ is most powerful. This was connected to the $chi\,(\mathbf{p};\kappa)$ curves by observing that the most powerful κ corresponds to the value minimizing the $chi\,(\mathbf{p};\kappa)$ curve. Two possible perspectives were displayed, one spanning the beta parameter w evenly and one spanning $D(a,w)$ evenly.

For this case, we want to filter the SNPs to identify those that are likely associated with blood serum non-HDL concentration without assuming much about the alternative distribution. Choosing to span $D(a,w)$ evenly over-represents extreme beta parameter values in the alternative, which corresponds to a non-uniform prior over the alternative beta distribution. Instead, we choose to be agnostic and take the perspective of a uniform prior on the beta parameter w. Taking all κ values lying below the 5% threshold in the adjusted case as our range of plausible κ values, we generate the map of plausible alternatives for $\kappa \leq 0.18$ shown in Figure 6.10(a).

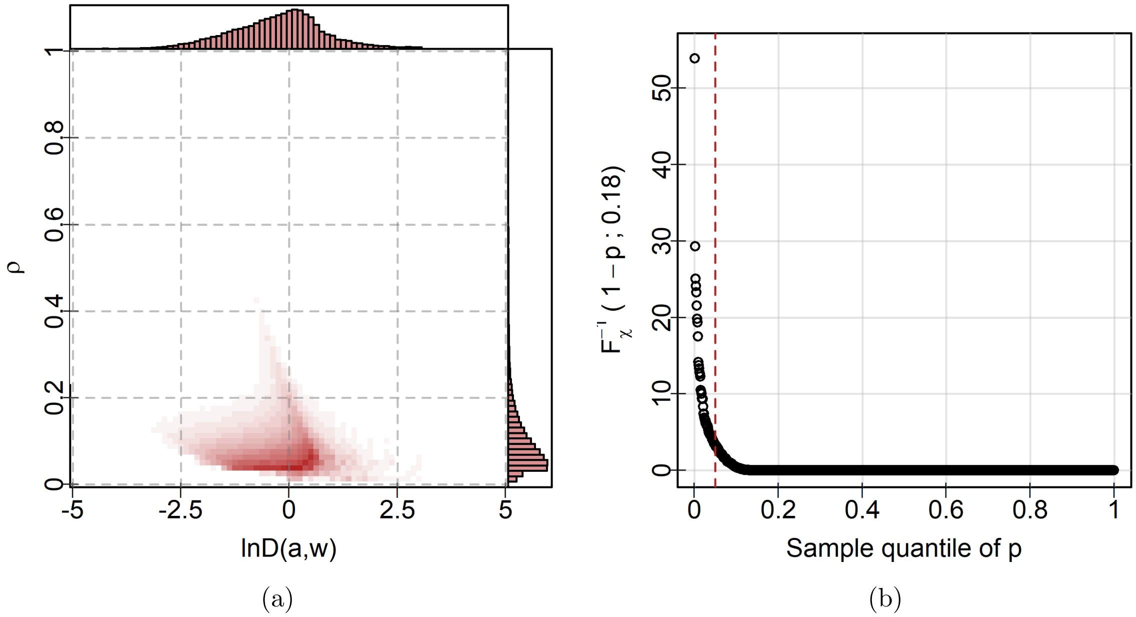


Figure 6.10: (a) The map of plausible alternatives for $\kappa \leq 0.17$. The rotated histogram on the right axis suggests a region of plausible alternatives has $\rho \in (0.05, 0.15)$. (b) $F_\chi^{-1}(1 - p_i; 0.18)$ by the sample quantile p_i. For the chosen cutoff 0.05, markers associated with p-values left of the vertical red line are considered the markers of interest.

This figure displays a heatmap of the region of alternative hypotheses where κ values less than $10^{-0.75}$ are most powerful. Darker saturation indicates more parameter settings where these κ are most powerful. Histograms with the same resolution as the heatmap display a scaled sum of this power for the corresponding margin. In particular, the histogram on the right side of the plot suggests that these κ are most powerful for $\eta \in (0.05, 0.15)$ with a peak in power around 0.05. Noting, in addition, the similar shape of the curve in this case to the simulated linear traits, we hypothesize that roughly 5% of the SNPs are relevant to determining non-HDL concentration in blood serum. We therefore focus on finding the $0.05M = 0.05 \times 1052 \approx 53$ marker locations which produce the largest value of

$$F_\chi^{-1}(1 - p; 0.18).$$

As the same κ value is applied to all tests these will simply be the smallest p-values, but converting to the $\chi^2_{0.18}$ quantile is a useful visual transformation to display the tests of interest, as demonstrated in Figure 6.10(b).

Considering the 53 SNP markers chosen by this cutoff, a natural question is their distribution across the genome. The line plot of Figure 6.11 displays the frequency of the 53 identifed SNPs by chromosome. Though only a small number of factors have been identified by the procedure, they are spread widely. This result is consistent with previous QTL analysis, which did not localize the genes responsible for non HDL cholesterol to a single chromosome, or even a handful.

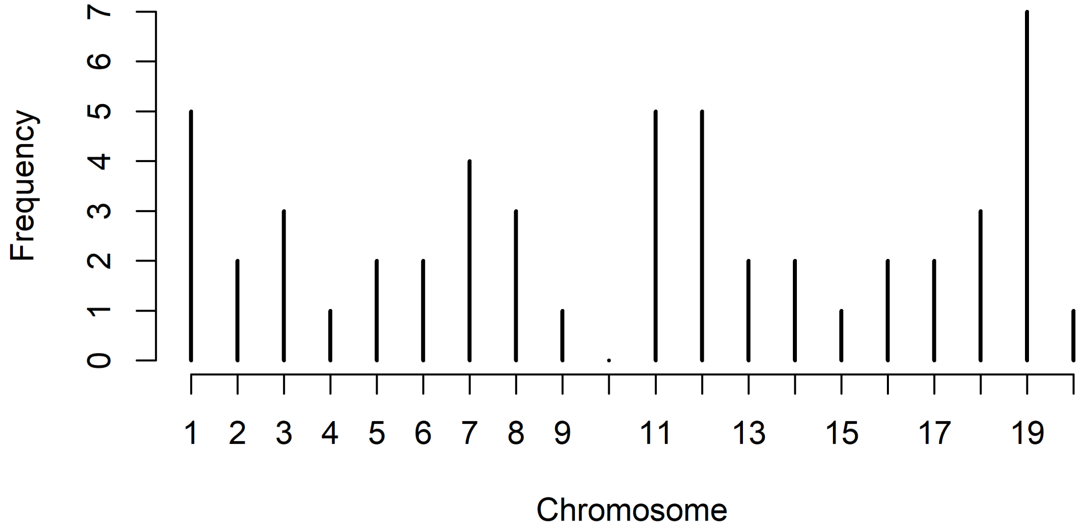


Figure 6.11: Frequency of identified markers by chromosome. The levels of non HDL cholesterol are strongly associated to a small number of factors widely spread on different chromosomes.

Indeed, the review of factors affecting blood serum lipoproteins in Wang and Paigen (2005) can be summarized into a similar plot by counting the number of identified regions on each chromosome which are associated with non-HDL blood serum concentrations in mice. This this review used the QTL method of Lander and Botstein (1989), it identified a very similar pattern on the genome to that identified here, in particular chromosomes 1, 19, and 11. The agreement of the novel method of pooled p-values presented here with the regions of interest identified by this more established method indicates the promise of the new method as a means to select regions in genomic studies. This is despite the data limitations present in this investigation that were not present in other studies, where phenotype and genotype data for individuals was collected and compared directly.

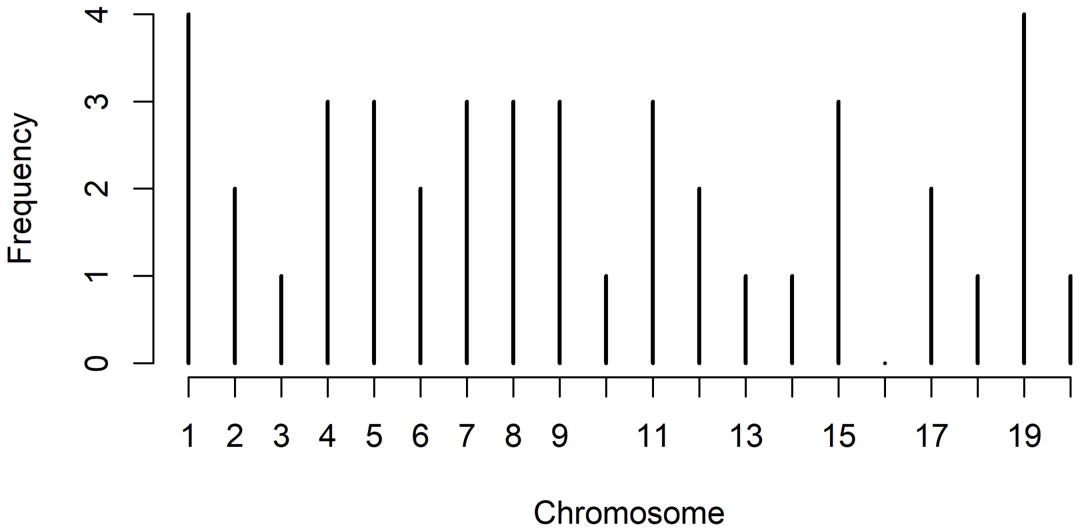


Figure 6.12: Frequency of markers by chromosome in Wang and Paigen (2005). The pattern between the two is rather similar, in particular in the peaks on chromosomes 1, 19, and 11.

6.5 Conclusions

An obvious application of central and marginal rejection in *chi* $(\mathbf{p}; \kappa)$ comes from considering the simple linear model $\mathbf{y} = \boldsymbol{\beta}\mathbf{Z} + \epsilon$ to relate a trait $\mathbf{y}$ to summarized encodings at M markers $\mathbf{Z}$ in a population, where $\boldsymbol{\beta} \in 0, 1^M$. The inheritance of traits through genetic information exists on a scale of monogenic to oligogenic, in the former case $\beta_m = 1$ for exactly one m in $\{1, \ldots, M\}$ while in the most extreme version of the latter $\beta_m = 1$ for all $m \in \{1, \ldots, M\}$. Increasing the proportion of coefficients which are one corresponds to increasing the η value of the alternative hypothesis without controlling the strength of the individual departures. Therefore, *chi* $(\mathbf{p}; \kappa)$ can be applied to identify the most plausible alternative hypotheses and corresponding range of η, indicating the proportion of markers contributing to a trait.

The dependence present in $\mathbf{Z}$ complicates this by producing smaller values of $chi\,(\mathbf{p};\kappa)$ across all κ for the same pattern of coefficients. Though the identification of η through the minimum location of this curve appears to be robust to dependence, thresholds which apply to the independent case cannot be used, making the significance of results unclear. Rather than perform simulations tailored to the specific correlation pattern and data to determine the correct thresholds for every investigation, a Satterthwaite approximation is proposed. Polynomials summarizing the relationship between genetic correlation and the correlation of χ^2 computed using Monte Carlo integration are used to approximate the distribution of $l_\chi(\mathbf{p};\kappa)$ and adjust $chi\,(\mathbf{p};\kappa)$. This adjustment places $chi\,(\mathbf{p};\kappa)$ in a more appropriate range when the values of $\mathbf{p}$ are dependent and correctly leaves independent p-values unadjusted. Applying this to real data, certain patterns in the genome identified using other methods are reproduced with adjusted $chi\,(\mathbf{p};\kappa)$ curves as a guide to selection.

Chapter 7

Example Application: Meta-Analysis

Pooled p-values are not new to meta-analysis, it is standard to combine individual p-values into a global p-value with a pooling function (Sinha et al., 2011). The framework of centrality and χ^2 pooling function $chi\,(\mathbf{p};\kappa)$ outlined in Section 4 can be directly applied to this case. The larger the centrality quotient of a pooling function of p-values (equivalently the larger the κ in $chi\,(\mathbf{p};\kappa)$), the more its pooled p-value depends on the mean of combined p-values and ignores the extremes.

A novel use of pooling functions is explored in the following section. By identifying a region of values which all individual estimates agree are plausible for the global parameter, a pooled p-value can be used to combine individual parameter estimates into a single global estimate of the unknown parameter. This presents a new perspective of combining estimates where each estimate provides an individual measure of the plausibility of a global estimate before the collection of individual measures is judged by a pooling function, say $chi\,(\mathbf{p};\kappa)$. This method seems to produce regions which are conservative in the inclusion of the true parameter and which simultaneously test the equality of study parameters in their construction. Though this method does not outperform classic confidence intervals, simulations suggest it provides richer information and is a promising avenue for future investigation.

7.1 Combining parameter estimates

Suppose we have a collection of M independent studies investigating a population parameter θ with respective sample sizes $n_1, \ldots, n_M$ and define the corresponding parameter values for each study as $\theta_1, \dots, \theta_M$. A key assumption of meta-analysis is homogeneity of study parameters, that is

$$\theta_1 = \cdots = \theta_M = \theta.$$

If the homgeneity assumption is false, meaningful differences exist between studies, perhaps due to differences in design or the sampled population, and their combination requires more careful analysis (if it makes sense at all).

Of course, we cannot know $\theta_1, \dots, \theta_M$ and only observe estimates $\widehat{\theta}_1, \dots, \widehat{\theta}_M$ with corresponding estimators $\widetilde{\theta}_1, \dots, \widetilde{\theta}_M$. Let the observed standard errors of these estimates be $s_1, \dots, s_M$, where each s_m is an estimate of the corresponding theoretical standard deviation $\sigma_1, \dots, \sigma_M$. The classic method of combining the study estimates into a single parameter estimate takes a weighted combination of the individual estimates

$$\widehat{\theta} = \frac{\sum_{m=1}^{M} \widehat{\theta}_m s_m^{-2}}{\sum_{m=1}^{M} s_m^{-2}}, \tag{7.1}$$

where the weight of each estimate is inversely proportional to its variance. Commonly, parameter estimator $\widetilde{\theta}_i$ will be approximately or asymptotically normal and so

$$\widetilde{\theta} \sim N\left(\theta, \frac{1}{\sum_{m=1}^{M} s_m^{-2}}\right)$$

if the homogeneity assumption is true. This admits a test of homogeneity based on the statistic

$$X^2 = \sum_{m=1}^{M} \frac{\left(\widehat{\theta}_m - \widehat{\theta}\right)^2}{s_m^2}$$

which is approximately χ^2_{M-1} distributed if all of the previous assumptions hold and all n_m are large enough (Sinha et al., 2011). We reject homogeneity at level α if X^2 is greater than the $1 - \alpha$ quantile of the χ^2_{M-1} distribution.

7.2 Defining an evidential estimate

To depart from the classic approach and use pooled p-values, note that each term of X^2 is the squared standardized difference of $\widehat{\theta}$ from $\widehat{\theta}_m$, defined for the m^{th} study as

$$d_m(x) = \frac{x - \widehat{\theta}_m}{s_m}.$$

The standardized difference gives the distance between a hypothesized x and $\widehat{\theta}_k$ in units of the estimated standard deviation of $\widetilde{\theta}_m$, and so represents the distance between them scaled by the standard deviation of the estimator. Rather than take the square of these distances, the evidential estimate converts each $d_m(x)$ to a p-value and reports the set of x where a pooled p-value computed using these is above a chosen threshold.

Assuming $\widetilde{\theta}_m \sim N(\theta_m, \sigma_m^2)$ and s_m^2 is an unbiased estimate of σ_m^2, $d_m(x)$ will have a non-central t distribution with $n_m - 1$ degrees of freedom and non-centrality parameter $x - \theta_m$ for any constant x. Let $t_\nu(\mu)$ be a random variable following a non-central t distribution with ν degrees of freedom and non-centrality parameter μ, so $t_\nu(0)$ is a central t-distributed random variable on ν degrees of freedom, denoted t_ν. Hypothesizing $x = \theta_m$, the non-centrality parameter is $x - \theta_m = 0$ and

$$p_m(x) = P\big(|t_{n_m-1}| \geq |d_m(x)|\big) = 2P\big(t_{n_m-1} \leq -|d_m(x)|\big)$$

is the p-value of $d_m(x)$ testing $H_{0m} : \theta_m = x$. As t_ν is normal in the limit of large ν, an approximate p-value for large n_m is given by

$$p_m(x) = P\big(|z| \geq |d_m(x)|\big) = 2P\big(z \leq -|d_m(x)|\big)$$

where $z \sim N(0,1)$. This converts the standardized differences $d_1(x), \ldots, d_M(x)$ into M p-values $\mathbf{p}(x) = (p_1(x), \ldots, p_M(x))^\mathsf{T}$ which can be pooled using $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ to test the simultaneous hypothesis

$$H_0 = \cap_{m=1}^{M} H_{0m}$$

and evaluate the overall evidence that $x = \theta_1 = \cdots = \theta_M = \theta$.

To establish a region of plausible θ values, a sequence of x values can be proposed and used to generate a sequence of $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ values in this manner. Those x where $\mathit{chi}\,(\mathbf{p}(x); \kappa) > a$ for a threshold $a \in (0,1)$ define a set in which $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ fails to reject H_0 at level a. Within this set, there is not enough evidence to reject any individual H_{0k} and so all $\widehat{\theta}_k$ agree on the plausibility of $x = \theta$. Therefore define the *evidential region* for threshold a to be the set

$$E_a = \{x : \mathit{chi}\,(\mathbf{p}(x); \kappa) > a\} \tag{7.2}$$

and take the x value maximizing $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ as a point estimate of θ, as it gives the weakest evidence against H_0 as measured by $\mathit{chi}\,(\mathbf{p}(x); \kappa)$. Call this point estimate the evidence-minimizing estimate (EME) of θ defined as

$$\widehat{\theta}^{(E)} = \arg\max_{x \in \mathbb{R}} \mathit{chi}\,(\mathbf{p}(x); \kappa)\,. \tag{7.3}$$

While this derivation took a particular choice of $d_m(x) \sim t_{n_m - 1}$, any function which admits p-values based on $\widehat{\theta}_1, \dots, \widehat{\theta}_M$ and x can be used in this way to define an evidential region. If E_a is convex, its boundaries are given by the minimum and maximum value of E_a.

It is possible that $\max_{x \in \mathbb{R}} \mathit{chi}\,(\mathbf{p}; \kappa) \leq a$, in which case E_a is the empty set denoted $\varnothing$ and there are no x which all studies agree are plausible for $x = \theta = \theta_1 = \cdots = \theta_M$ at a. The corresponding conclusion is that $\theta_1, \dots, \theta_M$ are not all equal and so homogeneity cannot be assumed. Evidential intervals therefore simultaneously perform an implicit test of homogeneity during their construction.

7.3 Choosing a centrality parameter

Expanding Equation (7.3) gives

$$\widehat{\theta}^{(E)} = \underset{x \in \mathbb{R}}{\arg\max} \left[1 - F_\chi \left(\sum_{m=1}^{M} F_\chi^{-1} \big(1 - p_m(x); \kappa \big); M\kappa \right) \right]$$

and Equation (7.2) gives

$$E_a = \left\{ x : 1 - F_\chi \left(\sum_{m=1}^{M} F_\chi \big(1 - p_m(x); \kappa \big); M\kappa \right) \geq a \right\}.$$

The mechanics of these expressions are dictated by the choice of κ, which determines the impact of individual p-values, and $p_m(x) = f(|d_m(x)|)$ where f gives the p-value of an observed standardized difference. For any discrepancy measure $|d_m(x)|$, such as the standardized difference, larger values provide stronger evidence against the null hypothesis and a value of zero indicates perfect agreement. Therefore $f(\cdot)$ is monotonically decreasing function with a maximum of $p_m(x) = f(|d_m(x)|) = 1$ at $|d_m(x)| = 0$.

To evaluate the choice of κ, recall from Section 4.5.3 that the centrality quotient of $\mathit{chi}\,(\mathbf{p}; \kappa)$ increases in the centrality parameter κ: the larger the value of κ the less any individual p-value affects the pooled p-value. When $\kappa \approx 0$, $\mathit{chi}\,(\mathbf{p}; \kappa)$ is a non-decreasing function of the minimum of $\mathbf{p}$ alone while $\kappa \to \infty$ gives a $\mathit{chi}\,(\mathbf{p}; \kappa)$ which is only weakly affected by the minimum. As the $\widehat{\theta}_m$ farthest from x give the largest discrepancy measures and therefore the smallest p-values, this suggests the value of $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ will be dominated by the most distant $\widehat{\theta}_m$ from x when $\kappa \approx 0$ and will largely ignore these distant $\widehat{\theta}_m$ when κ is large.

On the other hand, as x approaches $\widehat{\theta}_m$, $p_m(x) \to 1$ for each $m \in \{1, \ldots, M\}$. When κ is large $F_\chi(x;\kappa) \approx \Phi(x)$, the standard normal CDF, and $chi\,(\mathbf{p}(x);\kappa) \approx Sto(\mathbf{p}(x))$ which gives

$$\lim_{x \to \widehat{\theta}_m} F_\chi^{-1}(1 - p_m(x)) \approx \lim_{p_m(x) \to 1} \Phi^{-1}(1 - p_m(x)) = \Phi^{-1}(0) = -\infty,$$

and implies

$$\begin{aligned} \lim_{x \to \widehat{\theta}_m} chi\,(\mathbf{p}(x);\kappa) &\approx 1 - \lim_{p_m(x) \to 1} \Phi\left(\frac{1}{\sqrt{M}} \sum_{i \neq m} \Phi^{-1}(1 - p_i(x)) + \frac{1}{\sqrt{M}} \Phi^{-1}(1 - p_m(x))\right) \\ &= 1 - \Phi(-\infty) = 1 \end{aligned}$$

whenever $x = \widehat{\theta}_m$ for every $m \in \{1, \ldots, M\}$. This suggests spikes in $chi\,(\mathbf{p}(x);\kappa)$ at each $\widehat{\theta}_m$ when κ is large and that $Sto(\mathbf{p}(x))$ will not have a unique maximum value of $\widehat{\theta}^{(E)}$. Indeed, it may produce evidential regions which are not convex for certain choices of a.

Such spikes are avoided when $\kappa \approx 0$, in which case $chi\,(\mathbf{p}(x);\kappa) \approx Tip(\mathbf{p}(x))$ ignores the single large p-value that occurs whenever $x = \widehat{\theta}_m$. This is also an appealing choice intuitively, as it constructs an EME and evidential interval that bound the strongest evidence against H_0 by a. In this case

$$\widehat{\theta}^{(E)} \approx \arg\max_{x \in \mathbb{R}} Tip(\mathbf{p}(x)) = \arg\max_{x \in \mathbb{R}} p_{(1)}(x)$$

The value of x which maximizes $p_{(1)}(x)$ for a discrepancy measure $d_m(x)$ is the value of x which minimizes $|d_{(M)}(x)|$, so

$$\arg\max_{x \in \mathbb{R}} p_{(1)}(x) = \arg\min_{x \in \mathbb{R}} |d_{(M)}(x)| = \arg\min_{x \in \mathbb{R}} \max_{m \in \{1,\ldots,M\}} \frac{|\widehat{\theta}_m - x|}{s_m}.$$

The x which satisfies this expression must be a value within the range of $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$. To see this, first suppose x is outside the range of $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_K\}$. Then note that increasing or decreasing x to move closer to the estimates will simultaneously decrease all $d_m(x)$ and so also decrease their maximum. Therefore, the value of x which minimizes $|d_{(M)}(x)|$ must lie within the range of study means. As it is an internal point, it will be determined by only the largest value and smallest values of $\widehat{\theta}_m / s_k m$.

Let l be the index in $\{1, \ldots, M\}$ of the minimum of $\widehat{\theta}_1/s_1, \ldots, \widehat{\theta}_M/s_M$ and u that of the maximum. Then either $|d_l(x)|$ or $|d_u(x)|$ will be the maximum of $|d_1(x)|, \ldots, |d_M(x)|$

for any internal point x. To simultaneously minimize these discrepancies, the EME must satisfy

$$\left|d_l\left(\widehat{\theta}^{(E)}\right)\right| = \left|d_u\left(\widehat{\theta}^{(E)}\right)\right| \implies \frac{\widehat{\theta}^{(E)}}{s_l} - \frac{\widehat{\theta}_l}{s_l} = \frac{\widehat{\theta}_u}{s_u} - \frac{\widehat{\theta}^{(E)}}{s_u}$$

to give

$$\widehat{\theta}^{(E)} = \frac{\widehat{\theta}_u s_u^{-1} + \widehat{\theta}_l s_l^{-1}}{s_u^{-1} + s_l^{-1}}$$

which is similar to $\widehat{\theta}$ of Equation (7.1) but considers only the most extreme estimates and uses estimated standard deviations instead of estimated variances. The evidential intervals and EME for small κ are therefore expected to follow the centre of the range of $\widehat{\theta}_1/s_1, \ldots, \widehat{\theta}_M/s_M$, generating an internal interval closer to whichever of $\widehat{\theta}_l$ and $\widehat{\theta}_u$ has the smaller estimated variance.

Finally, a moderate κ value can be considered with $chi\,(\mathbf{p}; 2) = Fis(\mathbf{p}(x))$, or

$$chi\,(\mathbf{p}(x); 2) = 1 - F_\chi\left(-2\sum_{m=1}^{M} \ln p_m(x); 2M\right)$$

as $F_\chi^{-1}(1-p; 2) = -2\ln p$. Not only is this κ interesting because it corresponds to Fisher's method, but as F_χ is one-to-one the EME is given by

$$\arg\max_{x\in\mathbb{R}} \left[1 - F_\chi\left(-2\sum_{m=1}^{M} \ln p_m(x); 2M\right)\right] = \arg\max_{x\in\mathbb{R}} \sum_{m=1}^{M} \ln p_m(x) = \arg\max_{x\in\mathbb{R}} \prod_{m=1}^{M} p_m(x).$$

The evidential estimate when $\kappa = 2$ thereby maximizes the geometric mean of $p_1(x), \ldots, p_M(x)$ and E_a defines a set where the geometric mean of $p_1(x), \ldots, p_M(x)$ is greater than

$$-\exp\left[\left(F_\chi^{-1}(1-a)\right)^{1/M}\right]/2.$$

The behaviour of this function is not clear in advance. The spikes at each estimate and tendency to chase the centre of the range are straightforward to show for large and small κ respectively, but the shape and location of moderate κ are less clear.

To gain intuition about the behaviour of E_a and $\widehat{\theta}^{(E)}$ for these less obvious cases, curves of $chi\,(\mathbf{p}(x); \kappa)$ by x for different values of κ and patterns of observed study estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$ can be inspected. In each of the following examples, eight study means and their variances are chosen deliberately to illustrate the behaviour of E_a and $\widehat{\theta}^{(E)}$ for three κ which span the range of κ values, specifically $\kappa = 10^{-3.5} \approx 0.0003, 10^{0.3} \approx 2$, and

$10^{3.5} \approx 3162$. First, consider symmetric patterns about the mean of estimates shown in the matrix of plots in Figure 7.1.

Each subplot in the matrix displays curves of $chi\,(\mathbf{p}(x); \kappa)$ for a range of x values overtop vertical gray lines at each $\widehat{\theta}_m$ and a thicker vertical gray line at $\widehat{\theta}$ computed using Equation (7.1). A horizontal dashed line at $a = 0.05$ serves as an example threshold illustrating how the curves determine evidential intervals: E_a is defined by the region where the curve is above the threshold. Below the plot of curves, approximate 95% confidence intervals are added to each individual estimate at staggered heights so all intervals are visible. Finally, a point at $\widehat{\theta}$ is added below these intervals along with a black line giving its 95% confidence interval.

So, for example, the plot in the top right shows three coloured curves symmetric about 0. The green curve displays $chi\,(\mathbf{p}(x); 10^{-3.5})$, the orange curve displays $chi\,(\mathbf{p}(x); 2)$, and the purple curve displays $chi\,(\mathbf{p}(x); 10^{3.5})$. As expected, the purple curve spikes at each of the grey lines indicating individual estimates and the green curve peaks at 0, the centre of the range of estimates. The orange curve displays neither of these tendencies, and instead smoothly and symmetrically increases above the threshold in the centre of the range. All three methods produce similar evidential intervals for this plot.

Looking below the curves in this subplot, the individual estimates and their 95% confidence intervals can be inspected by viewing the horizontal grey lines. The estimates are symmetric about zero and have identical variance, with a greater concentration near zero than elsewhere. Many of the individual confidence intervals overlap, though there is no region where all overlap. The curves of $chi\,(\mathbf{p}(x); \kappa)$ are larger when more of the intervals overlap and smaller when fewer overlap, demonstrating the connection between individual assessments of plausibility and the pooled evidential estimate. Looking lower still, a point and a vertical line are plotted at $\widehat{\theta}$ and its associated 95% confidence interval is drawn with a horizontal black line. In this case, the evidential regions at $a = 0.05$ are intervals of similar width that nearly match the 95% confidence interval about $\widehat{\theta}$ for every κ.

For the leftmost subplot in the second row, the same pattern of estimates is displayed again but the identical variance of every estimate is smaller, which can be seen immediately by comparing the individual confidence intervals below the plot of curves. By decreasing the width of the intervals, their overlap is reduced and the value of $chi\,(\mathbf{p}(x); \kappa)$ decreases at all x for every κ. For the threshold plotted, this leads to $E_a = \varnothing$ for every κ.

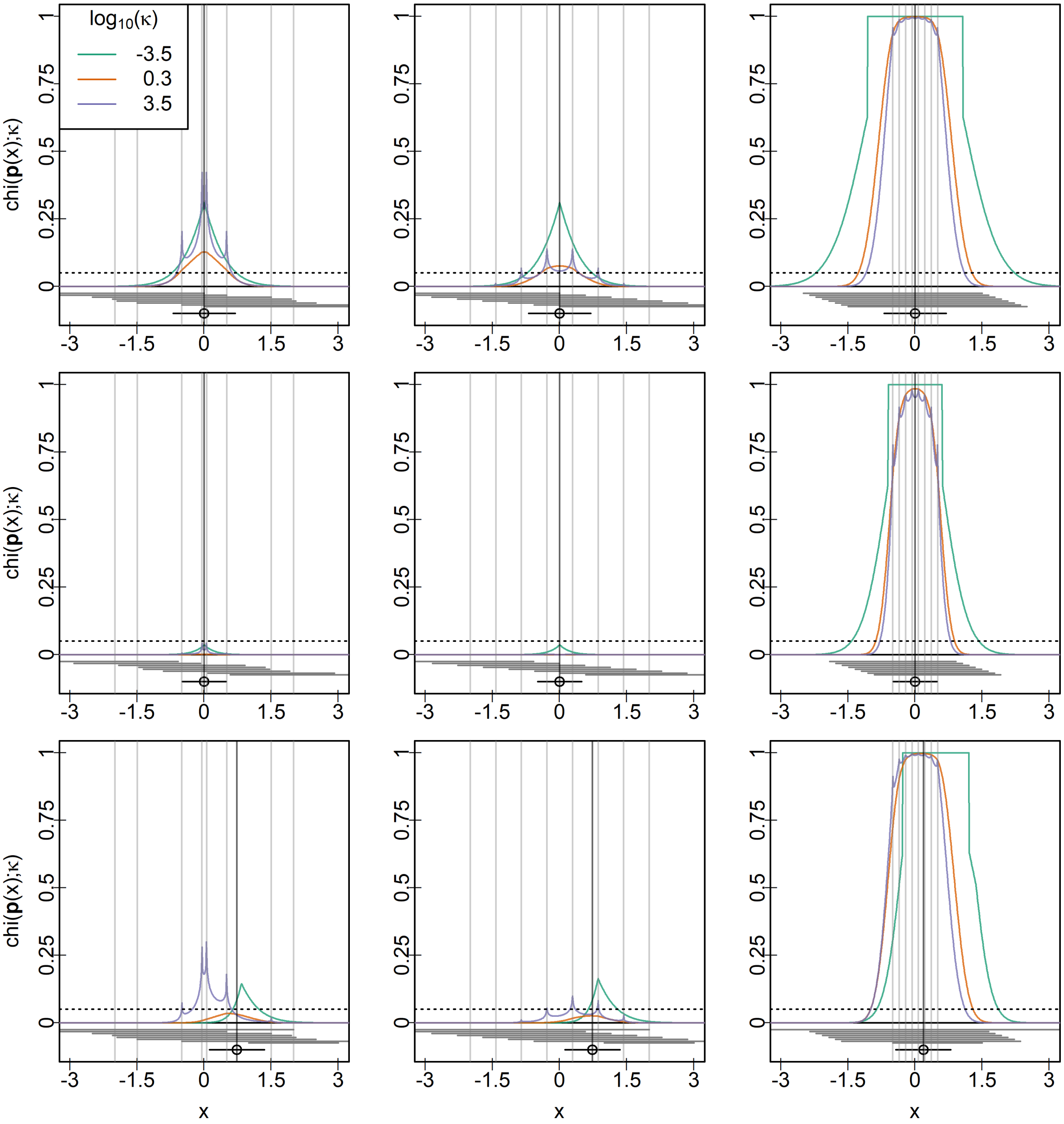


Figure 7.1: Canonical examples of $\widehat{\theta}_m$ symmetrically located about the centre of their range.

In the third row the variance is no longer constant for every estimate. Rather, the leftmost estimate has twice the variance of the estimates in the middle and the rightmost estimate has half their variance. This both lowers $chi\,(\mathbf{p}(x);\kappa)$ by reducing the overlap of intervals and pulls the maximum of each curve to the right, towards the estimate with smaller variance. The shift is particularly dramatic for $\kappa = 10^{-3.5}$, which has a maximum nearly three quarters of the range above the leftmost estimate, rather than perfectly centred as before. Though the shape of $chi\,(\mathbf{p}(x);10^{3.5})$ is distorted towards this estimate, its location and peak are still nearly at zero. The moderate $\kappa = 2$ seems to attempt a balance of these two extreme κ and peaks somewhere between them at a lower value. It matches $\widehat{\theta}$ and its confidence interval most closely, while the large and small κ form intervals almost entirely above and below $\widehat{\theta}$.

In the second column, the variance patterns in each row are the same as are the left- and rightmost estimates, but the internal estimates are now spread evenly across the range of estimates, leading to less central overlap in their confidence intervals. This does not impact $chi\,(\mathbf{p}(x);10^{-3.5})$ at all, as it only considers the most extreme estimates and so has curves identical to those in the first column. In comparison, the curves for $\kappa = 10^{3.5}$ and 2 are lower for the evenly spaced estimates than the centrally concentrated pattern. By decreasing the concentration of estimates, the average p-value in the central overlapping region is reduced and so these estimates with moderate and large centrality quotients give a smaller pooled p-value. The confidence intervals around $\widehat{\theta}$ are unchanged, as the weighted mean of the evenly spaced pattern is identical to the centrally concentrated one.

Finally, the third column displays the same progression of variances with all estimates highly concentrated near their centre so their individual confidence intervals nearly completely overlap. This high concentration of estimates results in $chi\,(\mathbf{p}(x);\kappa) \approx 1$ near the centre of the estimates for all κ, and creates evidential intervals which extend beyond the range of estimates. The width of these intervals decreases as the variances decrease, and all intervals are pulled towards the extreme estimate with the smallest variance. The abrupt step of the curve to one observed when $\kappa = 10^{-3.5}$ is a result of limits in the computation of the corresponding χ^2 quantiles.

Another informative setting is asymmetric, where one or two estimates disagree with the others. Figure 7.2 displays a matrix of subplots similar to Figure 7.1 where the estimate values are constant in each column and the variance values are constant in each row. While the previous case highlighted the difference between symmetric and asymmetric patterns, these explore increasing distances between a single estimate and the others by decreasing the smallest estimate in successive rows while the others remain constant. As before, the estimates in the first and second row have equal variances with the second row having the smaller variance of the two. In the final row, the leftmost estimate has a smaller variance than the others while the rightmost estimate has a larger variance.

In these cases, the difference between small and large κ choices is stark. Consider the first subplot in the first row, where the evidential interval for $\kappa = 10^{-3.5}$ is highest in the centre of the range of estimates despite the high concentration of estimates on the right. In contrast, the curve for $\kappa = 10^{3.5}$ ignores the smallest estimate and produces an evidential region around the cluster of other estimates, suggesting robustness to the single dissenting estimate. Attempting to balance these two inconsistent patterns of evidence, the moderate $\kappa = 2$ has a lower level than either extreme κ and a location somewhere between them, though still favouring the cluster of estimates.

Moving across the first row, observe that the heights of the curves for all κ are reduced as the dissenting estimate moves away from the others. This reduction is least consequential for $\kappa = 10^{3.5}$, which still produces an evidential interval for the farthest case when the others are well below the example threshold of 0.05. Comparing subplots down the rows shows that reducing the variance of the estimates, in particular the smallest estimate, has a similar impact.

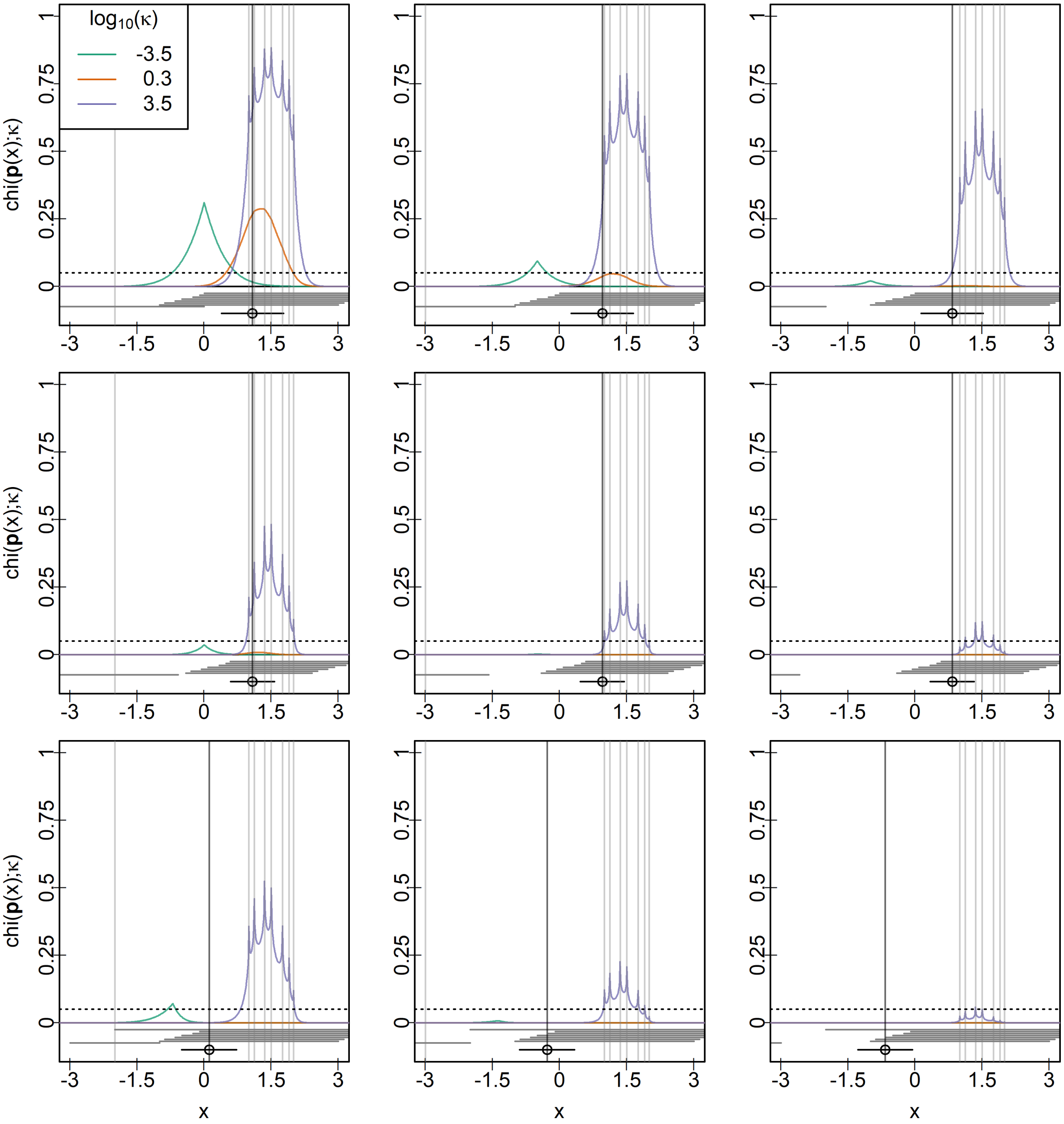
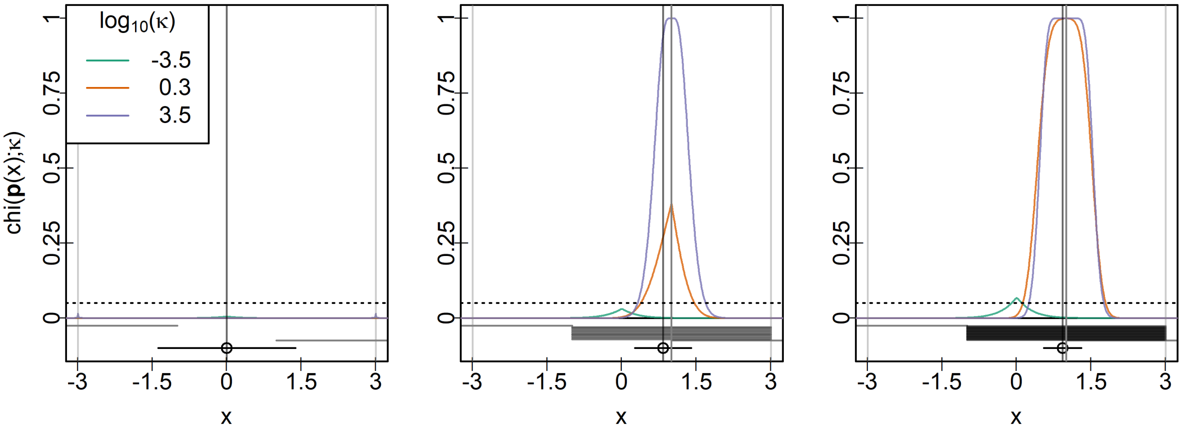


Figure 7.2: Canonical examples where most values of $\widehat{\theta}_m$ are concentrated on the upper end of their range.

A final illustrative case adds identical off-centre estimates between two fixed extreme estimates, shown in Figure 7.3. As expected, $\kappa = 10^{3.5}$ and 2 both produce $\mathit{chi}\,(\mathbf{p}(x);\kappa)$ curves which are largest around the repeated internal estimates while small $\kappa = 10^{-3.5}$ is largest in the centre of the range no matter the number of repeated estimates. Though the location of this maximum doesn't move, the curve for $\kappa = 10^{-3.5}$ still produces a non-empty evidential interval when M grows large enough, suggesting the implicit test of homogeneity is useful even when the chosen κ does not follow the pattern of evidence. When there are enough measurements in agreement, $\mathit{chi}\,(\mathbf{p}(x);\kappa)$ will ignore disagreement by one observation for any κ.



(a)

(b)

(c)

Figure 7.3: Canonical examples demonstrating the impact of adding (a) 0, (b) 10, or (c) 25 study estimates at a non-central point between two fixed estimates. While the confidence intervals decrease in width for every additional mean, the evidential intervals widen.

Figure 7.3 also demonstrates the different behaviour of evidential intervals and confidence intervals when $\widehat{\theta}_1 = \widehat{\theta}_2 = \cdots = \widehat{\theta}_M$. As the variance of the confidence interval is estimated by

$$\frac{1}{\sum_{m=1}^{M} s^{-2}},$$

the width of a confidence interval always decreases when M increases, provided all estimates have finite variance. For any choice of κ and $p_m(x)$, however, adding a new mean does not necessarily decrease the size of evidential regions. Indeed, in the case where $\widehat{\theta}_1 = \cdots =$

$\widehat{\theta}_M = y$ and $s_1 = \cdots = s_M = s$ the width of evidential intervals actually increases in M.[1] This behaviour is understood by considering the level of central rejection for $\mathit{chi}\left(\mathbf{p}(x);\kappa\right)$.

Recall Equation (4.12), which expresses the central rejection level for the χ^2 pooled p-value as

$$p_c = 1 - F_\chi\left(\frac{1}{M}F_\chi^{-1}(1-\alpha;M\kappa);\kappa\right) \to 1 - F_\chi\left(\kappa;\kappa\right)$$

as $M \to \infty$. This limit implies a harsh rejection boundary at $p_c = 1 - F_\chi(\kappa;\kappa)$ for large M such that

$$\mathit{chi}\left((p,\ldots,p)^\intercal;\kappa\right) = \begin{cases} 1 & \text{when } p > 1 - F_\chi(\kappa;\kappa) \\ 1/2 & \text{when } p = 1 - F_\chi(\kappa;\kappa) \\ 0 & \text{when } p < 1 - F_\chi(\kappa;\kappa). \end{cases}$$

In meta-analysis, this creates an asymptotic evidential interval based on $F_\chi(;\kappa)$ which does not depend on the threshold a, and within which $\mathit{chi}\left(\mathbf{p}(x);\kappa\right) = 1$. The bounds of this interval will be the two values of x for which $p_m(x) = 1 - F_\chi(\kappa;\kappa)$, that is where

$$P\left(\frac{|x-\widetilde{\theta}_m|}{s} < \frac{|x-y|}{s}\right) = F_\chi(\kappa;\kappa)$$

for $m = 1,\ldots,M$. As M increases, the evidential intervals begin to approach a piecewise constant function that is one within these bounds and zero outside of them. Effectively, the method does not distinguish between any of the values within this range of the repeated study estimates, and the only way to decrease the range is to decrease the variance of the corresponding estimators.

This behaviour seems counter-intuitive; such close agreement of many estimates suggests that the global parameter has been found. Indeed, that is the conclusion supported by the decreasing width of the confidence intervals as more estimates are added. That evidential intervals behave so differently in this case is instructive.

Recall the observation from Figure 7.1 that the curves of $\mathit{chi}\left(\mathbf{p}(x);\kappa\right)$ are highest in regions where multiple 95% confidence intervals for individual estimates overlap. This is a visual representation of the way these curves are constructed: each x is individually compared to each estimate for plausibility using $p_m(x)$ before a pooling function combines these individual votes of plausibility. When numerous estimates perfectly coincide, all agree on the plausibility of nearby values and pooling functions which are based on the majority

[1] Note E_a defines an interval for any a and κ in this case because the only localized spike in $\mathit{chi}\left(\mathbf{p}(x);\kappa\right)$ occurs at $x = y$ and $\mathit{chi}\left(\mathbf{p}(x);\kappa\right)$ decreases monotonically as $|x-y|$ increases.

vote, such as $\mathit{chi}\,(\mathbf{p}(x); 10^{3.5})$, therefore take large values for these nearby points. Only decreasing the individual variances so that $p_m(x)$ decreases faster in $|\widehat{\theta}_m - x|$ shortens the plausible regions. Adding more identical estimates only increases the number of concurring votes for the plausibility of nearby values.

All of this suggests that centrality is a useful concept when choosing κ to produce evidential intervals in meta-analysis. Large values of κ with large centrality quotients tend to produce evidential intervals which follow groups of tightly clustered estimates, often ignoring any extreme estimates that disagree. Just as an individual p-value is not generally enough to reject when using large κ to pool tests, an individual estimate barely affects $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ for large κ when combining estimates in meta-analysis. Small κ display the opposite tendency, ignoring any central patterns of estimates in favour of a region that balances the evidence against the most extreme estimates. A small centrality parameter suggests that only the smallest p-values matter in testing, and similarly that only the most extreme estimates matter in meta-analysis. Moderate κ present something of a balance, which often leads them to reject homogeneity when the extremes and central patterns are not consistent even though larger or smaller κ fail to reject.

For any κ, evidential regions provide richer information than confidence intervals. Primarily, this is due to the implicit test of homogeneity. While a confidence interval can always be constructed and does not have a built in feature that communicates when the interval may not be meaningful, the curve of $\mathit{chi}\,(\mathbf{p}(x); \kappa)$ provides information about both the location of a pooled estimate and consistency of all individual estimates. By changing κ, consistency can be evaluated along a range from the majority vote for large κ to only the most extreme votes for small κ. A secondary consequence of this is a more informative behaviour of evidential regions about the pattern of estimates than confidence intervals. Large κ produce evidential regions that chase clusters and ignore single dissenting estimates while small κ consider only the extreme estimates and so ignore clusters entirely. Confidence intervals cannot be so easily tuned to create such a broad range of behaviours.

7.4 Simulating coverage and rejection probabilities

Once κ is chosen, with the above results suggesting a moderate κ to balance the behaviour of the extremes, the construction of E_a requires the choice of a threshold a. Changing a changes both $\alpha(a) = P\,(E_a = \varnothing)$, the probability of rejecting H_0, and the evidential intervals created when $E_a \neq \varnothing$. Of equal interest to testing H_0 and thus homogeneity is the coverage probability $\pi(a) = P\,(\theta \in E_a |\, E_a \neq \varnothing)$. The former justifies inference by indicating whether a common θ is plausible while the latter indicates our confidence that

E_a includes the common θ should it exist. Both pieces of information are contained in E_a and so both are controlled by the threshold a. To investigate the behaviour of $\alpha(a)$ and $\pi(a)$, a simulation study follows.

Two generative models are commonly used in meta-analysis: the fixed-effect model and random-effects model of study means (Normand, 1999; Sinha et al., 2011). Assuming normality, the fixed-effect model asserts

$$\widetilde{\theta}_m \sim N(\theta, \sigma_m^2)$$

where $\sigma_m^2 > 0$ is the variance of $\widetilde{\theta}_m$ arising from random sampling of the population. The random-effects model takes

$$\widetilde{\theta}_m | \theta_m \sim N(\theta_m, \sigma_m^2)$$

where

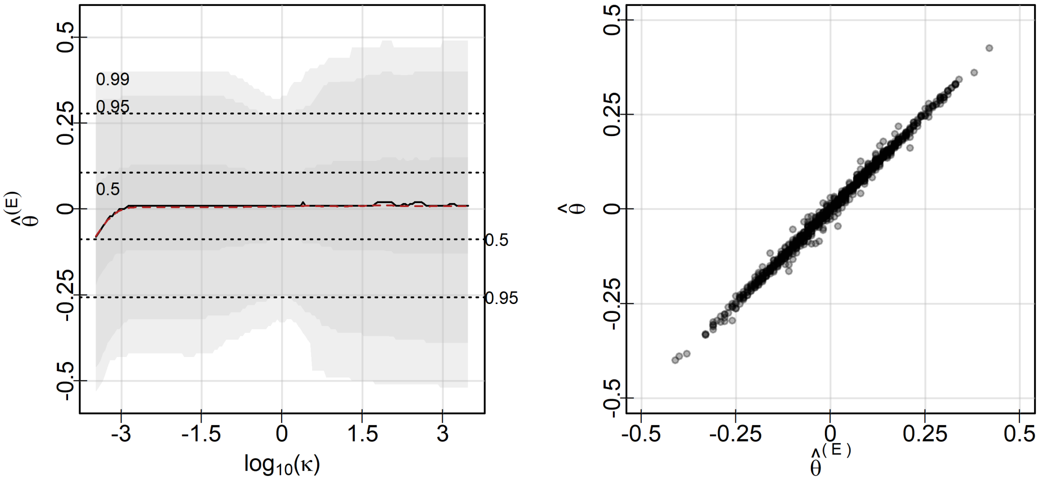
$$\theta_m | \theta \sim N(\theta, \tau^2)$$

for some specified $\tau > 0$, where $\sigma_m^2 > 0$ again represents the variation in $\widetilde{\theta}_m$ arising from random sampling. If $\tau > 0$, then $P(\theta_1 = \cdots = \theta_M) = 0$, and so homogeneity is violated. The magnitude of this violation increases in τ, very large τ will tend to produce study means far apart while $\tau \approx 0$ will generate $\theta_1 \approx \cdots \approx \theta_M$ and so behaves similarly to the fixed-effect case. Indeed, the fixed-effect model is a special case of the random-effects model with the parameter choice $\tau = 0$.

7.4.1 Fixed-effect with equal variances

As E_a incorporates a test of $\theta_1 = \cdots = \theta_M = \theta$ in its construction and this assumption is explicitly violated under random-effects with $\tau > 0$, the coverage and rejection probabilities were first computed assuming the fixed-effect model. Explicitly, 240 study observations are generated independently from a $N(0, 4)$ distribution split into 8 groups of $n = 30$ observations each. Within each group, the observed mean $\widehat{\theta}_m$ and mean standard deviation s_m^2 are computed and recorded. This was repeated 1000 times. For each repetition, the t_{29} distribution is then used to compute $p_m(x)$ from $d_m(x) = \frac{x - \widehat{\theta}_m}{s_m}$ and these are pooled using $\mathit{chi}\left(\mathbf{p}(x); \kappa\right)$ for every $\ln \kappa \in \{-8, -7.9, -7.8, \ldots, 8\}$ over x from -1.5 to 1.5 at increments of 0.01. For each κ, evidential intervals E_a are constructed for every threshold $a \in \{0.01, 0.02, \ldots, 0.2\}$, a point estimate $\widehat{\theta}^{(E)}$ is determined by the location of the maximum of $\mathit{chi}\left(\mathbf{p}(x); \kappa\right)$, and the inclusion of the true mean of 0 in E_a is determined.

Consistent with the different behaviours seen in Section 7.3, the choice of κ seems consequential to the variability of $\widehat{\theta}^{(E)}$. Figure 7.4(a) displays the median, mean, and



(a)

(b)

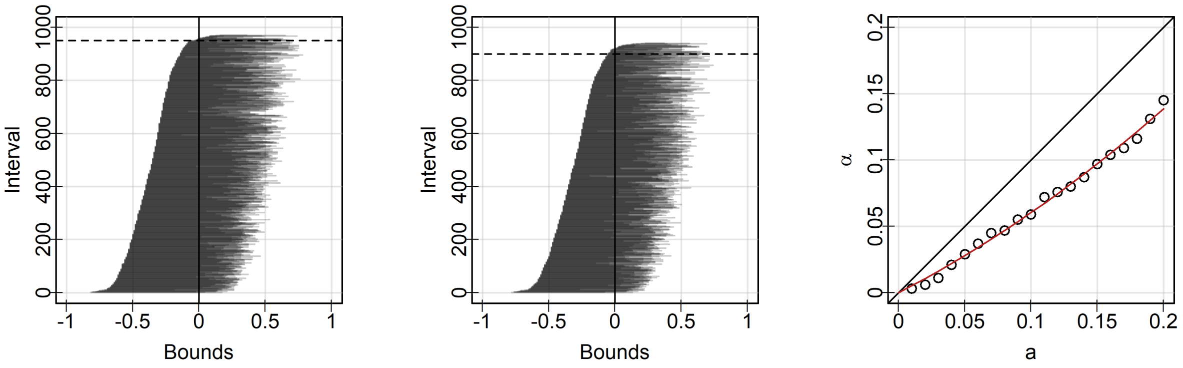
Figure 7.4: (a) The median (black line), mean (dashed red line), and central quantiles (labelled polygons) of $\widehat{\theta}^{(E)}$ compared with the central 0.5 and 0.95 quantiles of $\widehat{\theta}$ (dotted black lines). The region near 1 in κ has minimal variance in $\widehat{\theta}^{(E)}$ and a very similar distribution to $\widehat{\theta}$, as shown in (b) for $\kappa = 1$.

central quantiles of $\widehat{\theta}^{(E)}$ across all values of κ and compares them with the central quantiles of $\widehat{\theta}$ from Equation (7.1) displayed by dotted lines. With the exception of very small values of κ, $\widehat{\theta}^{(E)}$ is unbiased and symmetrically distributed[2] though the variation of $\widehat{\theta}^{(E)}$ changes considerably in κ. The local spikes in $chi\left(\mathbf{p}(x);\kappa\right)$ whenever $x = \widehat{\theta}_m$ for some $m \in \{1, \ldots, M\}$ for large κ make the choice of a maximum dependent on which sampled x is closest to an observed estimate. Similarly, the increased variance for small κ can be seen as a result of $chi\left(\mathbf{p}(x);\kappa\right)$ considering only the largest and smallest standardized distances, meaning $\widehat{\theta}^{(E)}$ is effectively based on only two observations. Moderate values in the range $[1/2, 2]$ have neither of these quirks and so appear to minimize the variance of $\widehat{\theta}^{(E)}$. They also produce estimates which are nearly identical to $\widehat{\theta}$, seen in Figure 7.4(b). This latter property is not necessarily desirable in principle, but is an interesting feature.

Choosing $\kappa = 2$ due to its generally symmetric and smooth curves across all settings in Figures 7.1 and 7.2 and because it uses the classical $Fis(\mathbf{p})$ to construct evidential

[2] The negative bias in these cases is an artifact of computation for extremely small degrees of freedom in the χ^2 distribution.

regions, all 1000 instances of E_a were plotted as stacked lines for $a = 0.05$ and $a = 0.01$ in Figures 7.5(a) and (b).[3] These plots order the intervals by their left bounds as a visual aid and plot empty intervals as space at the top of each stack, allowing the reader to simultaneously observe $\alpha(a)$ and every evidential interval. For a more complete depiction of the relationship between α and a, Figure 7.5(c) plots the two across all thresholds a with a black reference line added at $\alpha(a) = a$.



(a)

(b)

(c)

Figure 7.5: E_a for $\kappa = 2$ under H_0 when (a) $a = 0.05$ and (b) $a = 0.1$. The space at the top of the vertical axis without intervals corresponds to the cases where $E_a = \varnothing$, the black dashed horizontal line indicates $1 - a$. In all cases $\alpha < a$, indicating the threshold a defines a conservative test with a true level less than the chosen threshold a, shown in more detail in (c).

At every threshold a, $\alpha(a) < a$, indicating that the rejection rule $chi\,(\mathbf{p}(x); 2) \leq a$ swept over all x is a conservative test of H_0 at level a. To estimate $\alpha(a)$ based on these simulations, a polynomial is fit using least squares with stepwise selection excluding the intercept (as $P(chi\,(\mathbf{p}(x); 2) < 0) = 0$) to the observed $a, \alpha(a)$ pairs. This gives the quadratic

$$\alpha(a) = 0.514a + 0.891a^2$$

with a coefficient of determination $R^2 = 0.998$ and highly significant individual coefficients with p-values both less than 0.000035. The fitted line is shown in red on Figure 7.5(c).

[3] Though all evidential regions are intervals for $\kappa = 2$, this would not necessarily be the case for larger κ. As demonstrated earlier, non-convex evidential regions are possible for large κ.

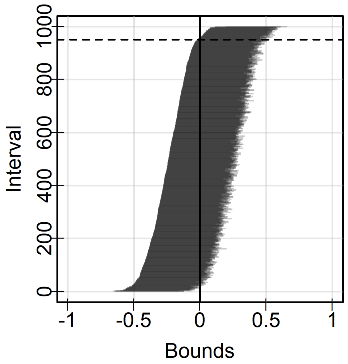


Figure 7.6: The 95% confidence intervals of $\widehat{\theta}$, which tend to be shorter and less variable in length than E_a.

Another insight of Figure 7.5 is that evidential intervals, where they exist, have more variable widths than the corresponding confidence intervals of $\widehat{\theta}$ displayed in Figure 7.6. The evidential intervals have widths which are much less regular with a tendency to be shorter when the left bound is larger, compared to the relatively consistent widths independent of location for the confidence intervals. Figure 7.7 makes this clearer by plotting the widths of 95% confidence intervals against evidential intervals with $a = 0.05$ for each of the 1000 simulated data sets in a scatterplot. Marginal histograms are added to both margins with constant bin widths so that the marginal distributions can be compared directly along with the joint distributions. E_a tends to produce much wider intervals than the confidence interval most of the time, and these widths tend to be more variable. The range of the confidence interval widths is roughly 0.15 compared to the range of 1 for evidential intervals. Despite this, the widths seem related with larger confidence intervals associated with larger evidential intervals, likely because both increase as any of $s_1, \dots, s_M$ increase.

The examples in Section 7.3 give some insight into this observation. It was noted there that the pattern of $\widehat{\theta}_1, \dots, \widehat{\theta}_M$ affects the size of E_a, with close agreement of estimates (relative to their respective variances) leading to larger evidential regions and estimates far from each other (relative to their variances) leading to smaller evidential regions. Figure 7.8 plots three particular simulations of the 1000 that match these earlier behaviours. As before, the width of the confidence interval does not change greatly for different patterns of estimates while $chi\left(\mathbf{p}(x); 2\right)$ produces a wide interval in Figure 7.8(c) when the estimates are tightly clustered, a very narrow interval in Figure 7.8(b) when two estimates seem to disagree with the others, and a moderate interval when the estimates do not display either

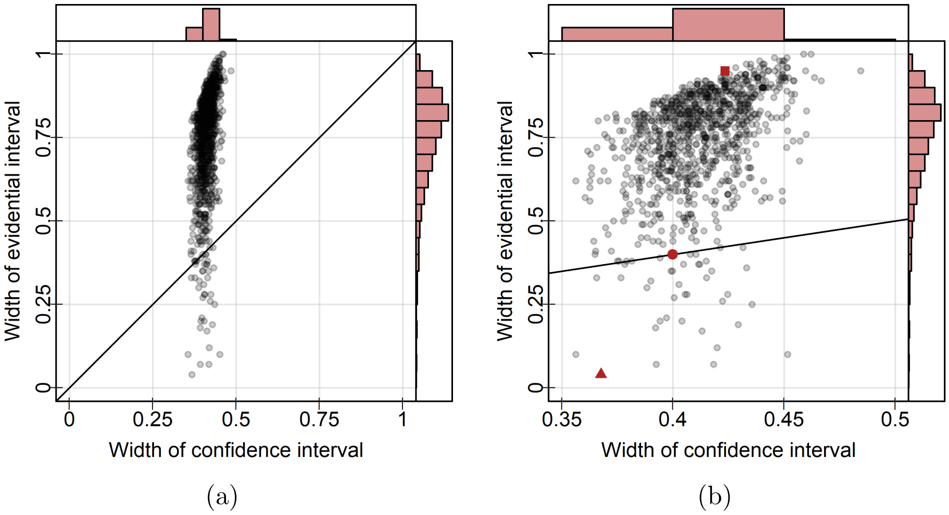
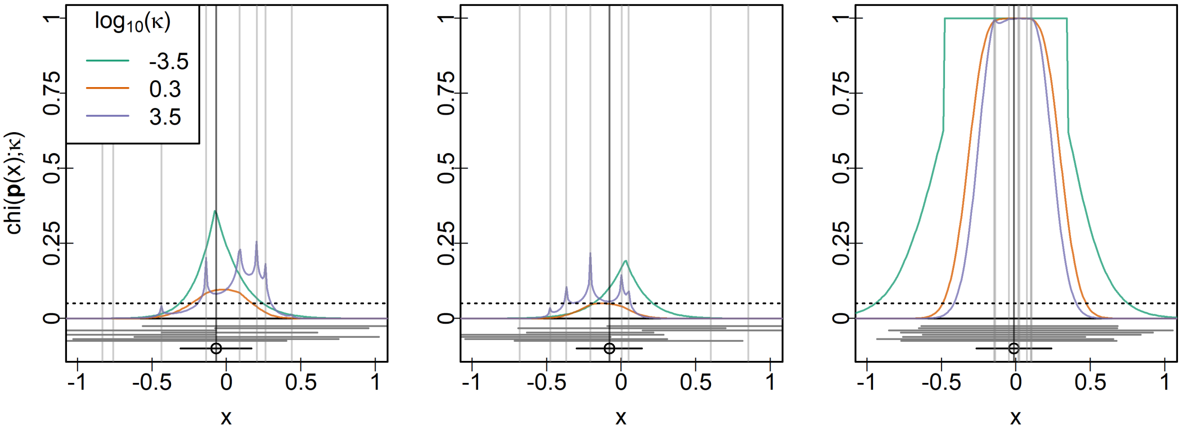


Figure 7.7: 95% confidence interval widths against evidential interval widths when $a = 0.05$ for all 1000 simulated meta-analyses. (a) uses equal axis ranges while (b) restricts the range horizontally to make details visible. The widths of evidential intervals vary more than confidence intervals and are wider generally, though a positive relationship exists between the two.

pattern as in Figure 7.8(a). These examples correspond to the red square, the red triangle, and the red circle in Figure 7.7 respectively.

The behaviour of these intervals is the same in this simulated case as in the constructed ones. When all means are close together relative to their variances, values of x outside of the range of means will still produce $p_m(x)$ values large enough to be considered plausible by E_a for all $m \in \{1, \ldots, M\}$. The opposite is true when study means are spread widely relative to their variances, in which case the region where $chi\,(\mathbf{p}(x); \kappa) \geq a$ may be very small and so E_a will be shorter than the confidence interval around $\widehat{\theta}$.

Though the behaviour of $chi\,(\mathbf{p}(x); 2)$ contains information about the distribution of study estimates, the resulting variability in the length of E_a means the coverage probability $\pi(a) = P(\theta \in E_a | E_a \neq \varnothing)$ is not obvious. The inclusion of $\theta = 0$ in each of the 1000 E_a for each κ is therefore determined and plotted by κ in Figures 7.9(a) and (b) for $a = 0.05$ and 0.1. A vertical line at $\kappa = 2$ provides reference, as well as horizontal lines at $1 - a$ and $1 - a/2$. Grey polygons plotted around the line give approximate 95% confidence intervals for each coverage probability based on a normal approximation to the binomial distribution. Figure 7.9(c) plots the observed π by a when $\kappa = 2$.



(a) (b) (c)

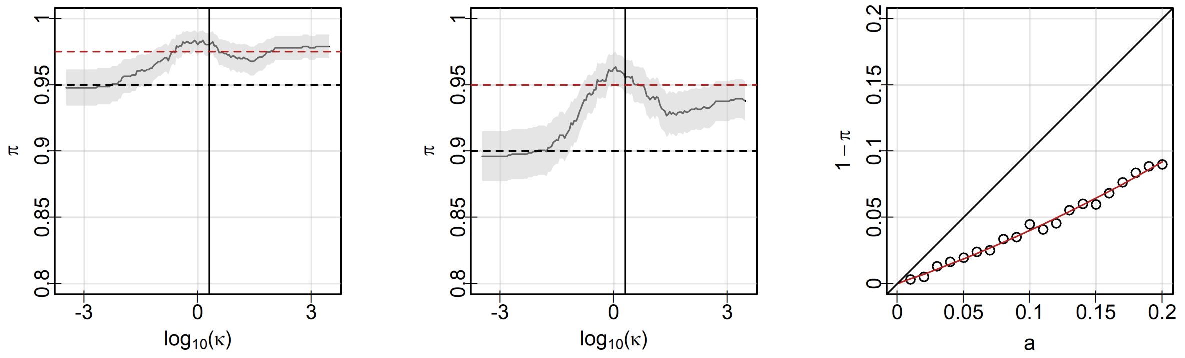
Figure 7.8: Plots showing the pattern of study estimates when the ratio of widths of the evidential interval with $a = 0.05$ for $\kappa = 2$ to 95% confidence intervals (a) is nearly one, (b) is the smallest observed over 1000 repetitions, and (c) is the largest observed over 1000 repetitions. The patterns here match those of the constructed examples earlier.

A consistent pattern in π and κ is observed across different thresholds a. When $\kappa \approx 0$, $\pi(a)$ is slightly less than $1 - a$, it then increases to a maximum above $1 - a/2$ when $\kappa \approx 1$ before decreasing to values greater than $1 - a$ for $\kappa \approx 10$ and increasing again as κ grows larger. As $1 - \pi(a) > a$ for all but small κ, it seems the evidential regions produced are conservative for all thresholds $a \in \{0.01, 0.02, \ldots, 0.2\}$. More precisely, $1 - \pi(a)$ was estimated using the observed $a, 1 - \pi(a)$ pairs using stepwise linear regression without an intercept, as $a = 0$ generates the evidential interval $E_0 = \mathbb{R}$ which implies $\pi = 1$ by definition. The fitted model is

$$\pi(a) = 1 - 0.342a - 0.588a^2$$

with $R^2 = 0.998$ and individual coefficients with p-values both less than 0.00016. A red line displays the model on Figure 7.9(c).

In the case of the fixed-effect model with equal estimator variances, it seems E_a produces regions and an implicit test homogeneity which are both conservative at level a. That is to say, the coverage probability of E_a is greater than or equal to a and the probability that falsely rejects the null hypothesis is less than or equal to a. For both, inference is supported with at least a confidence. Just as in the constructed case, the intervals produced contain



(a) (b) (c)

Figure 7.9: Coverage probabilities π of E_a when $\kappa = 2$ for studies with equal estimator variance when (a) $a = 0.05$ (b) $a = 0.1$. Black horizontal dashed lines indicate $1 - a$ while red horizontal dashed lines indicate $1 - a/2$. (c) plots 1 - $\pi(a)$ by a when $\kappa = 2$. Coverage probabilities increase in κ until $\kappa \approx 1$, after which they decrease slightly, suggesting these most stable estimates also uniquely maximize the coverage probability.

information about the overlap of individual estimate confidence intervals, and so their widths change considerably in the concentration of estimates.

7.4.2 Fixed-effect with unequal variances

Commonly $s_1^2 \neq \ldots \neq s_M^2$ due to differences in study sample size or design. To address this case, the fixed-effect equal variance simulation was repeated with unbalanced study sizes to investigate the how unequal variances of $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_M$ affect evidential intervals. As before, each repetition in this second simulation study generates 240 observations from a $N(0, 4)$ distribution and divides them into 8 groups, but it does not assume a fixed sample size for the studies. Instead, 240 labels are randomly sampled with replacement from a set with repeated entries, explicitly $\{1, 2, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8\}$, and assigned to each observation in order. This creates a wide variety of study sizes $n_1, \ldots, n_M$ summarized in the histogram of Figure 7.10. Group sizes ranged from 7 to 81 with a large peak in frequency near 25 and and a smaller one near 60.

Once groups were assigned, the analysis for each of 1000 simulations was identical to the case of equal variances. Group means $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$ and mean variances $s_1, \ldots, s_M$ were

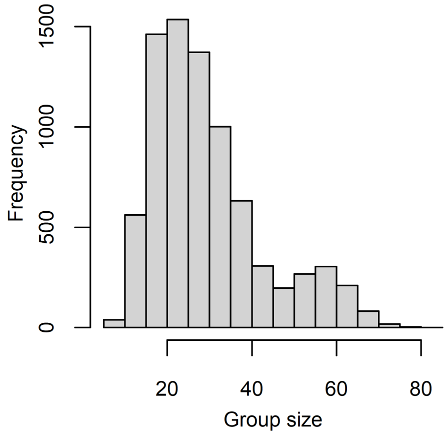


Figure 7.10: Histogram of the frequency of different group sizes over 1000 repeated samples.

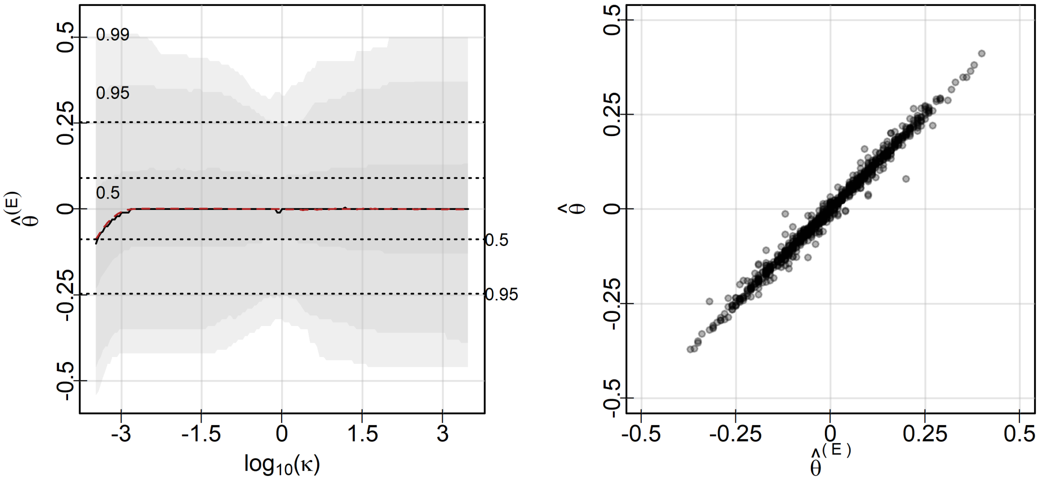
computed and used to generate $p_m(x)$ based on $d_m(x)$ and the t_{n_m-1} distribution, and these were pooled for a range of κ values. Evidential intervals were constructed for a range of thresholds a. All parameter choices were the same as in the case of equal variances.

As before, the first investigation was of the median, mean, and central quantiles of $\widehat{\theta}^{(E)}$ across all values of κ, displayed in Figure 7.11. Though it is still unbiased, $\widehat{\theta}^{(E)}$ displays greater variation for small κ in this case compared to the case of equal variance. The EME for small κ values is less variable than large κ in the case of equal study variances but is more variable for this simulation with unequal and random variances. The constructed examples of Section 7.3 shed some light on this pattern. When κ is small its maximum is pulled towards whichever of the two bounding estimates has smaller variance, and if the variances of these bounding estimates are not fixed then the extra variation in these variances translates to a more variable EME for small κ. In contrast, large κ values tend to chase the cluster of estimates nearest to each other, which is less affected by the increased variation of individual study means.

Despite this, κ values between $[1/2, 2]$ are still the most stable and produce EMEs with very similar values to $\widehat{\theta}$ as seen in Figure 7.11(b). Tangentially, this suggests that in both the equal variance and unequal variance settings, $\widehat{\theta}$ approximately minimizes

$$\sum_{m=1}^{M} F_{\chi}^{-1}\left(1 - 2F_t\left(-\frac{|x-\theta_k|}{s_k}; n_m - 1\right); 1\right)$$

over $x \in \mathbb{R}$, where $F_t(x;k)$ denotes the CDF of a t_k random variable.



(a)

(b)

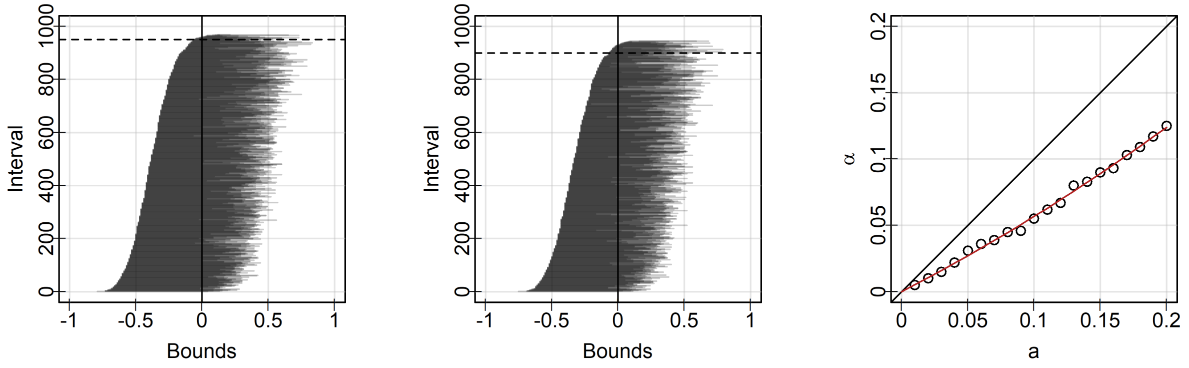
Figure 7.11: (a) The median (black line), mean (dashed red line), and central quantiles (labelled polygons) of $\widehat{\theta}^{(E)}$ compared with the central quantiles of $\widehat{\theta}$ (dotted black lines). The region near 1 in κ with minimal variance for $\widehat{\theta}^{(E)}$ has a very similar distribution to $\widehat{\theta}$, as shown in (b) for $\kappa = 1$.

Focusing again on $\kappa = 2$ due to the stability of $\widehat{\theta}^{(E)}$ for this κ and others near one in both the equal and unequal variance settings, all 1000 E_a are displayed for $\kappa = 2$ when $a = 0.05$ and 0.1 in Figures 7.12(a) and (b). These intervals look similar to the equal variance case. They seem to have roughly the same variability in length and are again conservative with $\alpha(a) < a$, a pattern shown in more detail in Figure 7.12(c). Once again, stepwise selection of parameters in linear regression is used to fit a polynomial without an intercept to estimate $\alpha(a)$. The quadratic

$$\alpha(a) = 0.515a + 0.525a^2$$

is selected with coefficient of determination $R^2 = 0.999$ and individual coefficient p-values both less than 0.00008. As before, the threshold a defines a conservative test at level a against homogeneity, as it has a type I error less than a.

Comparing $E_{0.05}$ to the 95% confidence intervals for the fixed-effect model with unequal variances gives similar insight to the equal variance case. The evidential intervals and confidence intervals have similar marginal distributions to the equal variance case and a



(a) (b) (c)

Figure 7.12: E_a for $\kappa = 2$ under H_0 for studies with unequal estimator variance when (a) $a = 0.05$ and (b) $a = 0.1$. (c) displays $\alpha(a)$ when $\kappa = 2$, demonstrating the conservative test at level a established by the threshold a.

similar joint distribution. The evidential intervals are generally wider and more variable than the confidence intervals, and the two seem positively associated.

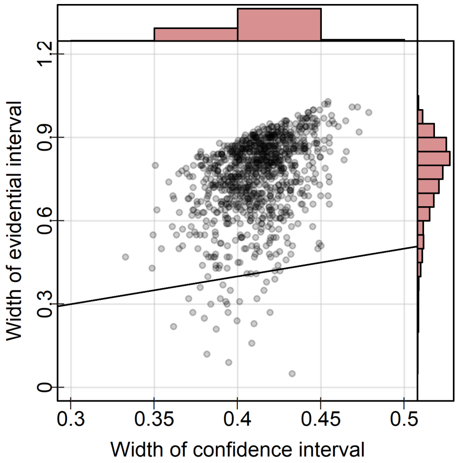
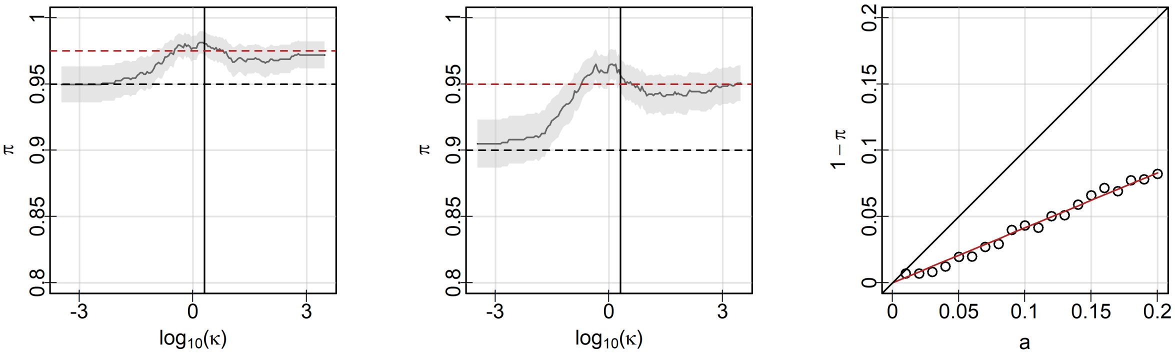


Figure 7.13: 95% confidence interval widths against evidential interval widths when $a = 0.05$ for all 1000 simulated meta-analyses with unequal variances. The pattern is the same as for equal variances.

Finally, $\pi(a)$ in the case of unequal variances is plotted by κ for several values of a in Figure 7.14. The pattern observed across all a is nearly identical to the case of equal variances: the coverage probability is lower for small κ, achieves a maximum above $1-a/2$ for κ between 1/2 and 2, drops slightly around $\kappa \approx 10$, and then increases to just less than $1-a/2$ for large κ. The observed inclusion probability is plotted by a when $\kappa = 2$ in Figure 7.14(c), displaying a similar pattern to the equal variance case yet again.



(a) (b) (c)

Figure 7.14: Coverage probabilities π of E_a when $\kappa = 2$ for studies with unequal estimator variance when (a) $a = 0.05$ (b) $a = 0.1$. Black horizontal dashed lines indicate $1-a$ while red horizontal dashed lines indicate $1-a/2$. The pattern of π in κ is nearly identical to the case of equal variances. (c) plots 1 - $\pi(a)$ by a when $\kappa = 2$.

In contrast to the equal variance case, however, stepwise selection of a polynomial model without an intercept for $\pi(a)$ when variances are not equal chooses a linear model

$$\pi(a) = 1 - 0.415a$$

with $R^2 = 0.996$ and a coefficient p-value less than the floating point numerical precision of R . Though it is once again conservative, this is different than the quadratic fit found in the fixed variance case.

Generally, changing from equal variance in study means to unequal and random variances does not seem to change much about the behaviour of evidential intervals. All of the same patterns observed for the equal variance case are repeated here. The only noteworthy exception is the variability of the EME for small κ, which is higher in this case owing to the increased variability of the standard deviations for the largest and smallest estimates.

7.4.3 Random-effects

In contrast to these two examples of the fixed-effect model, which assume the hypothesis of homogeneous estimates, $H_0 : \theta_1 = \cdots = \theta_M$, the random-effects model takes

$$\widetilde{\theta}_m | \theta_m \sim N(\theta_m, \sigma_m^2)$$

with

$$\theta_m | \theta \sim N(\theta, \tau^2)$$

for $\tau > 0$, $\sigma_m^2 > 0$. This explicitly violates H_0, as $\theta_M \sim N(\theta, \tau^2)$ implies $P(\theta_k = \theta_l) = 0$ for all $k \neq l$. Therefore, we expect E_a to produce empty intervals more frequently under random-effects than the fixed-effect model where H_0 is true. In the random-effects case, the proportion of empty intervals produced is the power of E_a to detect violations of homogeneity under this alternative.

To simulate the power of evidential intervals at detecting random-effects as a function of τ, 8 study means θ_{mi} were generated independently and identically $N(0, \tau_i^2)$ for every $\tau_i \in \{0.1, 0.2, \ldots, 1.9\}$. For the m^{th} study mean under variance τ_i, 30 observations were generated independently and identically following $N(\theta_{mi}, 4 - \tau_i^2)$ in order to keep the total variance comparable to the fixed-effect cases already investigated. The construction of evidential intervals proceeded by ignoring these known quantities and estimating $\widehat{\theta}_{mi}$ and its standard deviation from each group separately before pooling these for the same range of κ values as in the fixed case ($\ln \kappa \in \{-8, -7.9, \ldots, 7.9, 8\}$) and thresholding these by the same a ($a \in \{0.01, 0.02, \ldots, 0.2\}$). This procedure was repeated for 1000 repetitions, and the evidential intervals and EME were recorded for every case.

Figure 7.15 displays example evidential intervals for $a = 0.05$ when $\kappa = 2$ and $\tau = 1/\sqrt{2}$. As anticipated, these are empty much more frequently than in the fixed-effect case (Figure 7.5(a)). In a vast majority of repetitions, $chi\,(\mathbf{p}(x); 2)$ produces an empty evidential region when the means are generated from a normal distribution with variance 1. When $E_a \neq \varnothing$, the resulting intervals are more variable and narrow than in the fixed case.

Not all κ values proved this powerful, however. Figure 7.16(a) displays filled contours of the power surface of $E_{0.05}$ as a function of $\tau^2/4$ (the proportion of total variance explained by τ) and κ. The power at detecting the random-effects model is generally high, when the proportion of explained variance is only 0.2 the least powerful κ choices have a power of nearly 0.8. The most powerful choices seem to be κ in the range $[1/2, 2]$, with $\kappa = 1$ having the greatest power only narrowly. Figure 7.16(b) compares the power curve of $E_{0.05}$ for $\kappa = 2$, the cross-section of the surface in Figure 7.16(a) at the black line, to the power of

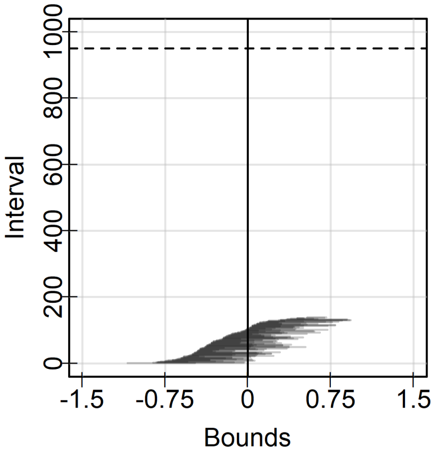


Figure 7.15: $E_{0.05}$ when $\kappa = 2$ for the random effects case with $\tau = 1/\sqrt{2}$. As expected, the method detects the inhomogeneity in most repetitions, achieving a power of 0.86.

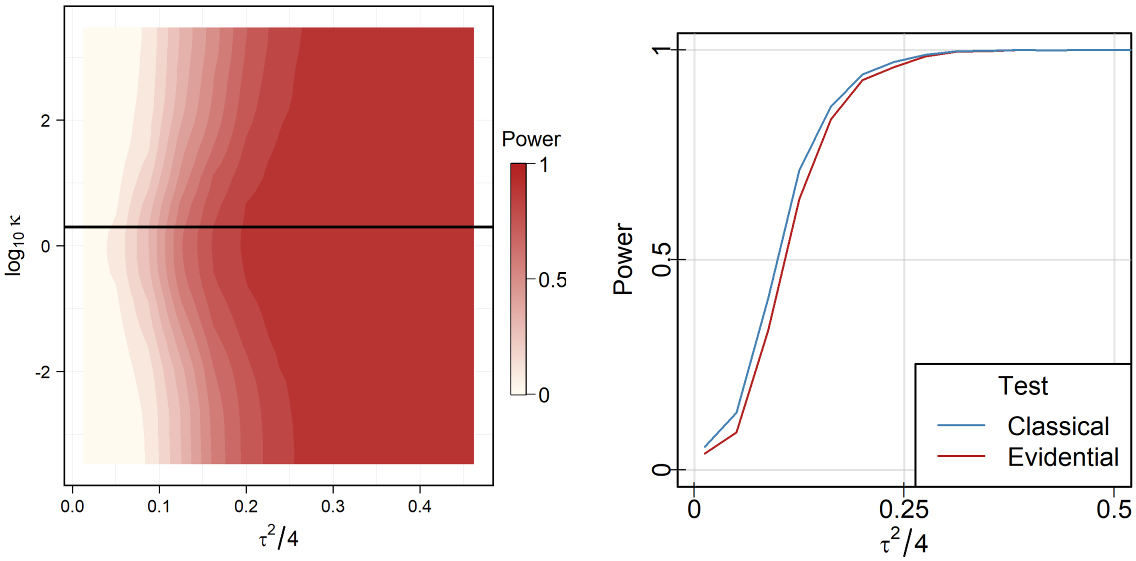
the classical test

$$X^2 = \sum_{m=1}^{M} \frac{(\widehat{\theta}_m - \widehat{\theta})^2}{s_m^2} \geq \chi^*_{M-1}(\alpha(0.05)) \approx \chi^*_{M-1}(0.514a + 0.891a^2) = \chi^*_{M-1}(0.028)$$

where $\chi^*_k(\alpha)$ is the $1 - \alpha$ critical value for the χ^2_k distribution. Though $E_{0.05}$ derived from $chi\,(\mathbf{p}(x); 2)$ does not attain the classical power in this setting, the two curves are quite close. Choosing $\kappa = 1$ tells a similar story. The classical test is slightly more powerful in this setting, but not by much.

7.5 Discussion

Given the observed reduction in the variance of the EME when κ is chosen in the interval $[1/2, 2]$ and the relatively high power of κ values in this range, it seems these κ values should be preferred in the construction of E_a using $chi\,(\mathbf{p}(x); \kappa)$. These patterns were nearly identical in the fixed-effect case both when all estimate variances were equal and when they were not, and are all interpretable in light of the demonstrated behaviour of $chi\,(\mathbf{p}(x); \kappa)$ for canonical examples. Moderate κ values in this range present a balance of the tendency to chase clusters of estimates for large κ and the tendency to care only about the bounding estimates for small κ.



(a) (b)

Figure 7.16: (a) power of $E_{0.05}$ by the portion of total variance explained by τ. Extreme choices of κ which are very small or large have low power relative to κ values in the interval $[1/2, 2]$. The power of $E_{0.05}$ when $\kappa = 2$ is compared to the classical test of homogeneity at level $0.514 \times 0.05 + 0.891 \times 0.05^2 = 0.028$ in (b). The implicit test based on evidential intervals is only slightly less powerful than the classical one.

The threshold $a \in (0,1)$ for κ values in this range additionally defines a conservative test for homogeneity at level a and generates conservative $1-a$ evidential intervals for θ when homogeneity is true. In particular, if $\kappa = 2$ is chosen and $a \leq 0.05$, the actual level of the test for homogeneity is well-approximated by $\alpha(a) = a/2$ and the coverage probability $\pi(a)$ is strictly greater than $1 - a/2$. For small α and $\kappa = 2$, simulations suggest the evidential interval $E_{2\alpha}$ simultaneously tests H_0 at level α and has coverage probability of at least $1-\alpha$, therefore simultaneously controlling both the probability of false rejection of H_0 and the probability of excluding of the true mean at α. This pattern is observed under simulation for both the case of equal variances and unequal variances in study estimators when H_0 is true.

Though it may just be a curiosity, there is a suggestive parallel between $chi\,(\mathbf{p}(x);2)$ and the likelihood ratio test. Define $L(\widehat{\theta}_m|x)$ as the likelihood of obtaining an observed study estimate of $\widehat{\theta}_m$ given a true study parameter of x. The joint likelihood of the sample

$\widehat{\theta}_1, \dots, \widehat{\theta}_M$ is

$$\prod_{m=1}^{M} L\left(\widehat{\theta}_m | x\right)$$

under the assumption of independent and homogeneous studies. The log-likelihood of the data is given by

$$\ln \prod_{m=1}^{M} L(\widehat{\theta}_m | x) = \sum_{i=1}^{M} \ln L(\widehat{\theta}_m | x),$$

and thresholding or maximizing this quantity is the basis of likelihood estimation. Recall

$$chi\,(\mathbf{p}(x); 2) = 1 - F_\chi \left(-2 \sum_{m=1}^{M} \ln p_m(x); 2M \right) \tag{7.4}$$

as $F_\chi^{-1}(1-p; 2) = -2 \ln p$. As F_χ is injective and monotonic, the threshold $chi\,(\mathbf{p}(x); 2) > a$ is equivalent to some threshold

$$\sum_{i=1}^{M} \ln p_m(x) > a'$$

for a' that is a monotonic transformation of a. The quantities $L(\widehat{\theta}_m | x)$ and $p_m(x)$ are closely related, as

$$p_m(x) = 1 - c \int_{-|\widehat{\theta}_m|}^{|\widehat{\theta}_m|} L(t|x) dt$$

for a normalizing constant c such that $c \int_{-\infty}^{\infty} L(t|x) dt = 1$. Though the specific thresholds will be different, both of these use model-defined likelihood functions to generate a statistic based on the sum of natural logarithms to make decisions about H_0. Moreover, if $p_m(x)$ is replaced by $L(\widehat{\theta}_m | x)$ in Equation (7.4), the expression becomes the asymptotic p-value of the log-likelihood ratio $-2 \sum_{m=1}^{M} \ln L_m(x)$ testing

$$H_0 : \theta_1 = \cdots = \theta_M = x$$

against

$$H_A : \theta_m \neq x \text{ for at least one } m \in \{1, \dots, M\},$$

so the substitution of this related quantity changes $chi\,(\mathbf{p}(x); 2)$ to the Neyman-Pearson UMP test.

Both evidential estimates and likelihood estimates depend entirely on the model-defined likelihood $L(\widehat{\theta}_m | x)$, but approach measuring the plausibility of $\widehat{\theta}_m$ given x differently. The likelihood takes the probability density of $\widehat{\theta}_m$ directly while $chi\left(\mathbf{p}(x);2\right)$ integrates the density in the tails beyond the observed value, complicating analysis of the behaviour of E_a in theory. Consider $\alpha(a) = P\left(E_a = \varnothing\right)$ and $\pi(a) = P\left(\theta \in E_a \middle| E_a \neq \varnothing\right)$, for example. Evaluating

$$\alpha(a) = P\left(E_a = \varnothing\right) = P\left(\max_{x\in\mathbb{R}} chi\left(\mathbf{p}(x);2\right) < a\right)$$

is not straightforward as it requires the distribution of

$$\max_{x\in\mathbb{R}}\left\{1 - F_{\chi}\left(-2\sum_{m=1}^{M}\ln\left[1 - c\int_{-|\widetilde{\theta}_m|}^{|\widetilde{\theta}_m|} L(t|x)dt\right];2M\right)\right\}. \tag{7.5}$$

The coverage probability $\pi(a)$ additionally requires evaluation of

$$1 - F_{\chi}\left(-2\sum_{m=1}^{M}\ln\left[1 - c\int_{-|\widetilde{\theta}_m|}^{|\widetilde{\theta}_m|} L(t|\theta)dt\right];2M\right)$$

conditioned on the distribution of the quantity in Equation (7.5). The results of the preceding simulations and the similar form of $chi\left(\mathbf{p};2\right)$ to the likelihood ratio support further investigation of both expressions to better characterize evidential intervals to combine studies in meta-analysis. Though it is less powerful than the classical method for the case of equal variances and random-effects, it may prove more powerful for other settings.

7.6 Recommendations

When using evidential intervals in practice, the preceding simulations suggest $chi\left(\mathbf{p}(x);1\right)$ or $chi\left(\mathbf{p}(x);2\right)$ is prudent unless there is a compelling reason to choose otherwise. These settings produce the least variable EME that balances the tendencies of larger and smaller κ to ignore certain estimates. Large κ may be chosen if a small proportion of estimates which disagree with the others and a pooled estimate ignoring these is desired. Small κ are harder to justify, as they ignore all but the smallest and largest estimates.

If $chi\left(\mathbf{p}(x);2\right)$ is chosen, simulations indicate that a threshold of a produces E_a which have coverage probability $\pi(a) \geq a$ and tests $\widehat{\theta}_1 = \cdots = \widehat{\theta}_M$ at level $\alpha(a) \leq a$. To choose a threshold when sample sizes are not equal, the polynomials $\alpha(a) = 0.515a + 0.525a^2$ and $\pi(a) = 1 - 0.415a$ can be solved for a given a desired $\alpha(a)$ or $\pi(a)$ if the estimates are computed from samples of different size. Alternatively, a threshold can be chosen in advance and these polynomials used to report the empirical coverage probability and type I error rate. If sample sizes are not equal, the corresponding polynomials are instead $\alpha(a) = 0.514a + 0.891a^2$ and $\pi(a) = 1 - 0.342a - 0.588a^2$.

7.7 Real data

The `metadat` package in R (White et al., 2022) contains several meta-analysis data sets reporting statistical summaries of studies collected to address a common theme. One such data set from Konstantopoulos (2011) addresses the performance of students in the United States under modified school calendars. In contrast to the standard school calendar, which gives students two months of holiday every summer, modified calendars maintain the same number of instructional days but distribute vacation throughout the year rather than concentrating it in July and August. The studies in Konstantopoulos (2011) all record the standarized difference in mean performance on a common test given to students receiving a modified schedule and others receiving a standard schedule in the same school. White et al. (2022) provide these differences in means, their pooled standard errors, and the district of each school in R, and Figure 7.17 displays the differences in means with intervals given by their standard errors for each school organized and coloured by district.

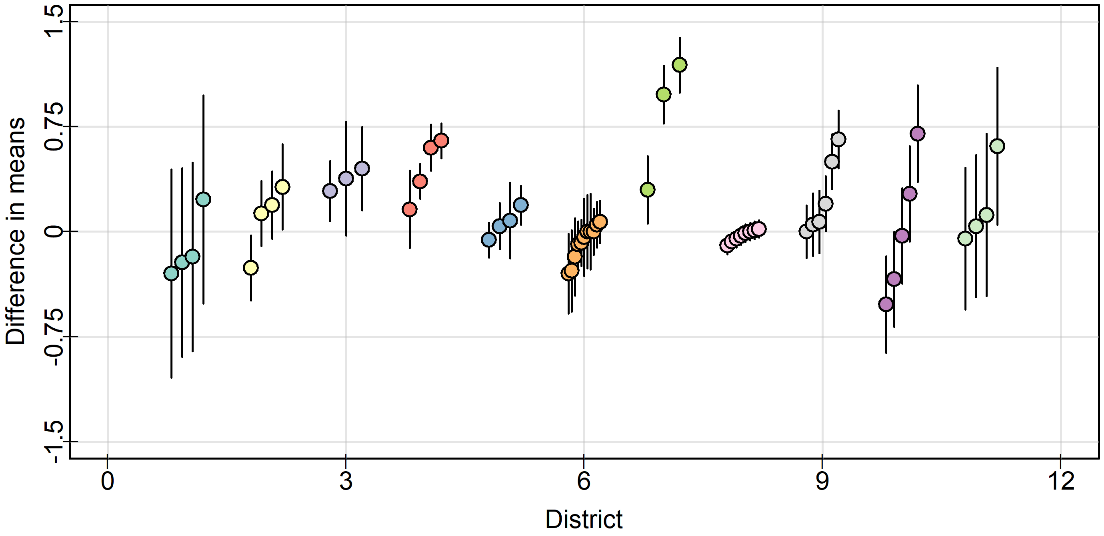


Figure 7.17: Difference in mean performance between students on a modified calendar and those on the standard calendar by district. Each point represents a different school within each district. The range of points varies widely by district.

Immediately obvious is the strong impact of district on the observed mean differences. Most districts form distinct clusters with overlapping intervals with the exception of the

7^{th} and 10^{th} districts, which have mean differences spread over a wide range by school. This has considerable impact on the evidential intervals for the data. Due to its good performance in simulations, $chi\,(\mathbf{p}(x); 2)$ was applied to the entire data set in an attempt to estimate E_a at $a = 0.05$ to produce an approximate 97.5% evidential interval and test homogeneity at approximately 0.025. Sweeping x values from -2 to 2, no region gave $chi\,(\mathbf{p}(x); 2) \geq 0.05$. Indeed, the maximum of 8.7×10^{-83} at 0.04 provides strong evidence against the hypothesis of homogeneity. Given the clear groups of performance by school district in Figure 7.17, this is unsurprising.

Noting that the effects within the districts tend vary less than between them, however, $chi\,(\mathbf{p}; 2)$ was used estimate $\hat{\theta}^{(E)}$ and E_a for each district separately and plot the resulting evidential intervals at $\alpha = 0.05$ in Figure 7.18. Even within districts, H_0 is rejected in 5 instances which therefore produce no evidential regions at $\alpha = 0.05$, though $\hat{\theta}^{(E)}$ still exists and is recorded.

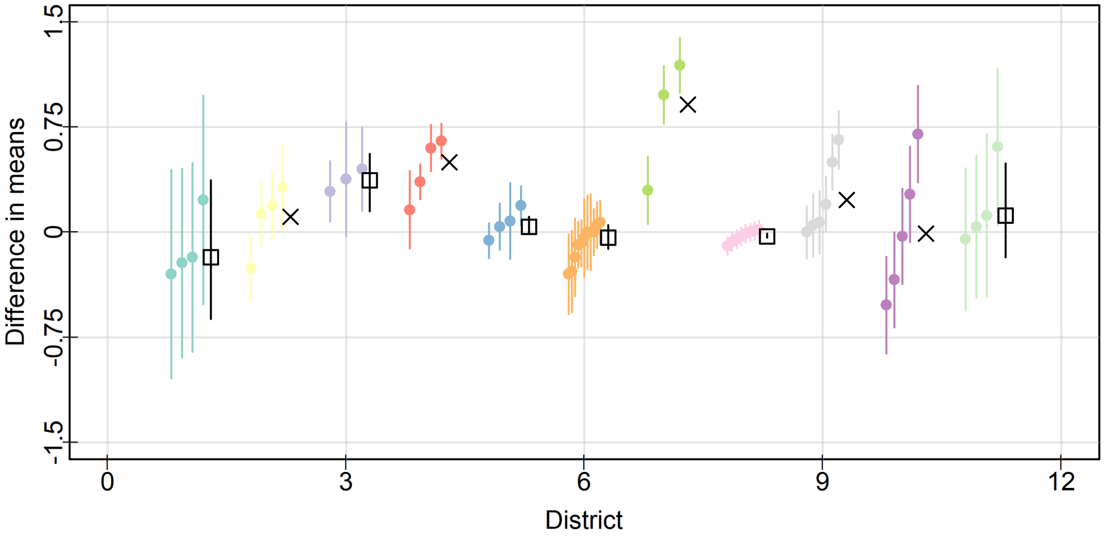


Figure 7.18: Difference in mean performance between students on a modified calendar and those on the standard calendar by district with $E_{0.05}$ and $\hat{\theta}^{(E)}$ plotted over them. "X" points indicate $\hat{\theta}^{(E)}$ in districts where the implicit test in E_a rejected homogeneity within the district.

The districts which produce evidential intervals are split evenly above and below zero, and four of the six intervals include zero. These results are consistent with the statistics reported in Konstantopoulos (2011) by district. Though the original study reported a barely significant positive effect for the the mean difference it is not clear this combined estimate is appropriate. The largest positive mean differences have evidence of heterogeneity at 0.05 and the least variable evidential intervals and estimates all occur near zero and tend to include zero in their evidential intervals and confidence intervals. In short, there is not enough consistent evidence that the switch away from a summer vacation to year-round schooling has an impact on student performance in general, though it may have impacts locally.

7.8 Conclusions

The combination of study estimates by thresholding at a and maximizing $chi\left(\mathbf{p}(x);\kappa\right)$ over a sequence of candidate x values to obtain E_a and $\widehat{\theta}^{(E)}$ is a promising method for further investigation. The EME $\widehat{\theta}^{(E)}$ is unbiased and has variability similar to $\widehat{\theta}$ for $\kappa \in [1/2, 2]$. Constructing an evidential region E_a for a κ in this range produces a region that implicitly tests homogeneity conservatively at a, and produces a region with coverage probability greater than $1 - a$. This is despite having a width which changes considerably in the pattern of study estimates, with clustered estimates relative to their standard deviations giving wider regions and spread estimates giving narrower regions. All of this can be understood by considering the set of x values which all estimates agree are plausible, and considering how the choice of κ impacts which estimates are favoured and which are relatively ignored. Small κ effectively only heed the smallest and largest estimates, while large κ ignore extreme estimates in favour of large groups of close estimates. These behaviours are exactly as expected given the corresponding centrality quotients.

Of course, these results do not show the evidential method to be optimal for any setting, and have not analyzed the coverage probability of E_a or probability of rejecting H_0 when it is true. The simulations completed here are also incomplete, there are many other settings under both the assumption of homogeneity and when it is violated than could be investigated. Despite these limitations, the results obtained here are promising and suggest further investgiations may give interesting results.

There are also obvious extensions to the simple idea presented here. Methods exist to estimate and account for τ from $\hat{\theta}_1, \dots, \hat{\theta}_K$ and $s_1, \dots, s_K$ in order to estimate θ, such as restricted maximum likelihood and Bayesian estimation (Normand, 1999; Sinha et al., 2011). Though a cautious approach calls for more data or more detailed analysis of each study when homogeneity is rejected, the computation of $p_m(x)$ could be modified to account for τ by computing it based on a more complex likelihood before sweeping x. Any model which admits p-values for all x can be used with the preceding framework to construct E_a and $\widehat{\theta}^{(E)}$.

Chapter 8

Conclusions

This thesis has addressed several aspects of the problem of detecting associations in large data sets. A new measure based on recursive binary splits was established and explored as a way to measure association. It was proven that splits only need to be considered at the coordinates of observed points to produce a binning maximizing scores based on the χ^2 and mutual information. Simulation studies characterized the null distribution of the χ^2 statistic on bins produced by the algorithm under different splitting regimes and demonstrated its power at detecting patterns in both constructed and real data. Random splitting was shown to be as powerful at detecting a range of patterns as maximized splitting, but with a null distribution conservatively approximated by the appropriate χ^2 distribution if bin size is limited. To detect association, random splitting therefore provides conservative p-values and a conservative test without any simulation. Maximized splitting, in contrast, requires simulation to model its null distribution, but produces final bins which can be plotted to effectively visualize a pairwise relationship. The recursive binning algorithm was implemented in the R package `AssocBin`, providing easy access to this promising method of analyzing pairwise dependence.

A framework to analyze pooled p-values was then developed. Central and marginal rejection represent two meaningful patterns of significance that arise in collections of M p-values. After defining both quantities and demonstrating their relevance to the behaviour of the uniformly most powerful test for a restricted beta family, the central rejection level was proven to always be greater than or equal to the marginal rejection level. The centrality quotient capitalizes on this proof to define a measure between 0 and 1 that communicates whether marginal or central rejection is favoured by a pooling function. It was proven that the centrality quotient of the pooled p-value based on quantiles of the χ^2_κ distribution, $chi\left(\mathbf{p};\kappa\right)$, can be controlled by changing the degrees of freedom κ, and expressions for the centrality quotient of $chi\,(\mathbf{p};\kappa)$ were derived. Simulation studies demonstrated the robustness of $chi\,(\mathbf{p};\kappa)$ to mis-specification of its parameters compared to the uniformly most powerful pooled p-value. A link between the pattern of significance and the degrees of freedom that minimize $chi\,(\mathbf{p};\kappa)$ was leveraged to generate a map that conveys a region of most plausible alternative hypotheses given a curve generated by applying the pooled p-value to a sample for different degrees of freedom. Code to compute central rejection, marginal rejection, the χ^2 poooled p-value, the uniformly most powerful pooled p-value, and generate these maps was combined into the R package `PoolBal` to facilitate its use.

The thesis next took a deeper dive into the important motivating example of genome-wide association studies. A mathematical model was derived from first principles and then used to derive expressions for the correlation between genetic markers and these were compared to real data using a custom plot matrix, the correlation test plot. This plot indicated a good fit to the observed data. The full model was implemented in the `toyGenomeGen` package to fill an unmet need for software that can perform fully customizable genetic experiments in R.

All previous results were combined to demonstrate how monogenic and oligogenic traits can be separately identified by recursive binning and the χ^2 pooled p-value adjusted to account for genetic correlation. Simulations showed that oligogenic and monogenic data lead to distinct patterns in the curve of $chi\,(\mathbf{p};\kappa)$ by κ for linear effects of genetic information on traits. A Satterthwaite approximation was developed to adjust these curves to account for known correlations between pairs of genetic markers to great effect. When applied to a real data set, adjusted curves produced results similar to those previously seen in the literature.

Finally, the use of $chi\,(\mathbf{p};\kappa)$ to combine parameter estimates in meta-analysis was explored. By considering many candidate estimates, a plausible region based on the set of values where $chi\,(\mathbf{p};\kappa)$ is above some threshold is obtained. The resulting evidential region thereby simultaneously combines estimates and tests them for homogeneity. Calibration of coverage probabilities and test levels were determined for the cases of equal and unequal estimate variances. In simulation, these regions were shown to be responsive to patterns in the data which traditional confidence intervals ignore without sacrificing much power or coverage probability. Indeed, by changing κ, different treatment of outliers can be obtained for the same combination method. The results of this exploration were intriguing, and recommendations for future investigation and use were provided.

In short, this thesis presents new methods in measuring association and combining p-values, with the latter supported by a novel framework proven relevant for the choice of pooled p-values. A complete tutorial on the basics of genetics results in a simple expression for genetic correlation which matches observation reasonably well. Several software tools implement all of these results to make them accessible for any R user. They are then immediately applied to one established problem and to make first developments into one exploratory and novel method in meta-analysis. A common theme of using many pairwise tests to gain insight about data ties these topics together. Each chapter addresses a different part a complicated puzzle: some motivate, others develop methods, and the last few demonstrate these methods in action.

8.1 Future directions

Recursive binning has a number of desirable qualities. Under random splitting the distribution of the χ^2 statistic over the final bins can be conservatively approximated, and maximized binning provides a convenient visualization of the data at the end. Both result in more flexible bin shapes than are possible through other, marginal binning methods such as Reshef et al. (2011), Jiang et al. (2015), and Heller et al. (2016). Comparing random binning to these earlier methods may prove interesting, as the greedy nature of bin-based methods has proven a point of contention in the past (Gorfine et al., 2012; Kinney and Atwal, 2014a; Reshef et al., 2014; Kinney and Atwal, 2014b; Simon and Tibshirani, 2014).

The conversion to ranks is a natural way to focus on the association between variables independent of their margins. Recursive binning is but one way to tesselate the unit square, and others invite further investigation. For one, it allows for different grid shapes to be used, such as as hexagonal bins (Carr et al., 1987), or custom bins designed to identify density in specific regions. Scott (1988) notes that hexagonal bins introduce less bias in two-dimensional histograms than rectangular ones, but the impact this would have on measuring association is not clear. Additionally, it is not obvious how to implement non-rectangular bins into the recursive algorithm. Alternatively, particular patterns of interest may be captured by specific binnings applied to data, or different split logic that changes bin density in certain regions of interest. In the S&P 500 data, for example, the corner bins have a particularly relevant interpretation that suggests increasing the number of corner bins to model tail dependence. This also suggests further investigations into copulas as a way to simulate rank patterns.

Deeper exploration of central and marginal dependence could also prove interesting. Though the minimum pooled p-value was proven to uniquely produce a centrality quotient of 0, at least two pooled p-values produce centrality quotients of 1 despite defining different combination functions. The curvature of the rejection boundary therefore seems relevant to further characterize this balance of marginal and central rejection. This is rather easy to visualize in two dimensions, but quite tricky in three or more. Similarly, it is not obvious how these concepts apply to more complicated sequential rejection methods, for example those of Simes (1986) and Hochberg (1988), or to methods designed to control the false discovery rate, such as Genovese and Wasserman (2002).

Within the χ^2 quantile function, solidifying the theory behind the curves generated by sweeping κ in the χ^2 pooling function would lend greater confidence to their application. Though the empirical use of these curves was demonstrated, little was proven about them. More fully exploring the theory of these curves would allow for stronger statments to be made, perhaps shedding more light on the plausible alternatives they imply. Simulations also suggested that the χ^2 pooled p-value could prove powerful in detecting alternative beta densities which are not strictly decreasing, a setting which is thus far relatively unexplored in the literature. Extending pooled p-values to this, perhaps, more realistic case would be useful and could pave the way to even more general alternative distributions which are not part of the beta family.

Though the focus of the genetic example here was diploid organisms, the genetic model outlined easily accounts for other ploidy with the addition or removal of columns from the annotation matrix. As several common plant species such as alfalfa and potato are tetraploid, this could expand the search for relevant data. Due to the relative ease of measurement and simpler ethical concerns, obtaining paired measurements of traits and genomes for these plants could prove simpler than for humans or mice. This could potentially garner more complete data on which to apply the methods of Chapter 6.

Evidential regions as a way to combine parameter estimates were treated as an application here, but the intriguing early results obtained suggest they may be worthy of study in their own right. The simulation settings and theory presented in this work barely scratch the surface of the method, but already garnered results suggesting their promise. Identifying cases where evidential regions are preferred over classical confidence intervals would be very informative, and would provide guidance on where to look next and inspiration for attempts to understand them theoretically.

References

Alan Agresti. Measures of nominal-ordinal association. *Journal of the American Statistical Association*, 76(375):524–529, 1981.

Francis J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1): 17–21, 1973.

Jon A. Beck, Sarah Lloyd, Majid Hafezparast, Moyha Lennon-Pierce, Janan T. Eppig, Michael F.W. Festing, and Elizabeth M.C. Fisher. Genealogies of mouse inbred strains. *Nature Genetics*, 24(1):23–25, 2000.

Allan Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

Molly A Bogue, Vivek M Philip, David O Walton, Stephen C Grubb, Matthew H Dunn, Georgi Kolishovski, Jake Emerson, Gaurab Mukherjee, Timothy Stearns, Hao He, et al. Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. *Nucleic acids research*, 48(D1):D716–D723, 2020.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391): 580–598, 1985.

Karl W. Broman, Lucy B. Rowe, Gary A. Churchill, and Ken Paigen. Crossover interference in the mouse. *Genetics*, 160(3):1123–1131, 2002.

Morton B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992, 1975.

Carol J. Bult, Judith A. Blake, Cynthia L. Smith, James A. Kadin, Joel E. Richardson, and the Mouse Genome Database Group. Mouse genome database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806, 2019.

Luisa Canal. A normal approximation for the chi-square distribution. *Computational Statistics & Data Analysis*, 48(4):803–808, 2005.

Dan Cao, Yuan Chen, Jin Chen, Hongyan Zhang, and Zheming Yuan. An improved algorithm for the maximal information coefficient and its application. *Royal Society Open Science*, 8(2):201424, 2021.

Daniel B. Carr, Richard J. Littlefield, W.L. Nicholson, and J.S. Littlefield. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, 82(398): 424–436, 1987.

J Casellas. Inbred mouse strains and genetic stability: a review. *animal*, 5(1):1–7, 2011.

Tony F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988.

Yuan Chen, Ying Zeng, Feng Luo, and Zheming Yuan. A new algorithm to optimize maximal information coefficient. *PLoS One*, 11(6):e0157567, 2016.

James M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 2001.

Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

Ozan Cinar and Wolfgang Viechtbauer. *poolr: Methods for Pooling P-Values from (Dependent) Tests*, 2021. URL https://CRAN.R-project.org/package=poolr. R package version 1.0-0.

Ozan Cinar and Wolfgang Viechtbauer. The poolr package for combining independent and dependent p values. *Journal of Statistical Software*, 101:1–42, 2022.

William G. Cochran. The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, pages 315–345, 1952.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Karen N. Conneely and Michael Boehnke. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.

Harald Cramér. Remarks on correlation. *Scandinavian Actuarial Journal*, 1924(1):220–240, 1924.

James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, 1970.

Hongying Dai, J Steven Leeder, and Yuehua Cui. A modified generalized Fisher method for combining probabilities from dependent tests. *Frontiers in Genetics*, 5:32, 2014.

Lien Dejager, Claude Libert, and Xavier Montagutelli. Thirty years of Mus spretus: a promising future. *Trends in Genetics*, 25(5):234–241, 2009.

Michael Dewey. metap: Meta-analysis of significance values. r package version 1.8, 2022.

Apostolos Dimitromanolakis, Jingxiong Xu, Agnieszka Krol, and Laurent Briollais. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1):1–9, 2019.

A. Adam Ding, Jennifer G. Dy, Yi Li, and Yale Chang. A robust-equitable measure for feature ranking and selection. *The Journal of Machine Learning Research*, 18(1):2394–2439, 2017.

R.W. Doerge, Z.B. Zeng, and B.S. Weir. Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science*, 12(3):195–219, 1997.

Sebastian Dümcke, Ulrich Mansmann, and Achim Tresch. A novel test for independence derived from an exact distribution of ith nearest neighbours. *PloS One*, 9(10), 2014.

Eugene S. Edgington. An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363, 1972.

Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas. Technical report, ETH Zurich, 2001.

E. Eskin. SNP data, 132,000+ locations for 248 inbred and RI strains of mice. MPD:UCLA1. Mouse Phenome Database web resource (RRID:SCR_003212), The Jackson Laboratory, Bar Harbor, Maine USA, 2023. https://phenome.jax.org.

H. Fairfield Smith. On comparing contingency tables. *The Philippine Statistician*, 6:71–81, 1957.

Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. The international genome sample resource (igsr) collection of open human genomic variation resources. *Nucleic acids research*, 48(D1):D941–D947, 2020.

Ronald A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

Ronald Aylmer Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 1932.

Janis S. Fisler, Craig H. Warden, Mario J. Pace, and Aldons J. Lusis. Bsb: a new mouse model of multigenic obesity. *Obesity Research*, 1(4):271–280, 1993.

Jerome H. Friedman and Werner Stuetzle. John W. Tukey's work on interactive graphics. *Annals of Statistics*, pages 1629–1639, 2002.

Nicholas W. Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7): 559–568, 2009.

Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE transactions on Knowledge and Data Engineering*, 25(4):734–750, 2012.

Christian Genest and Bruno Rémillard. Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369, 2004.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

Leo A. Goodman and William H. Kruskal. *Measures of Association for Cross Classifications.* Springer, 1979.

Malka Gorfine, Ruth Heller, and Yair Heller. Comment on "Detecting Novel Associations In Large Data Sets". *Science*, pages 1–6, 2012.

Constantinos Goutis, George Casella, and Martin T. Wells. Assessing evidence in multiple hypotheses. *Journal of the American Statistical Association*, 91(435):1268–1277, 1996.

Earl L. Green, Douglas L. Coleman, Nathan Kaliss, Charles P. Dagg, Elizabeth S. Russell, John L. Fuller, John Staats, and Margaret C. Green. *Biology of the laboratory mouse*. Dover, 1966. Accessed via online adaptation at http://www.informatics.jax.org/greenbook/index.shtml.

Major Greenwood and G. Udny Yule. The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general, 1915.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 2007.

Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, 1998.

J.B.S. Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8(29):299–309, 1919.

Benjamin C Haller and Philipp W Messer. SLiM 4: Multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E000–E000, 2023.

Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4): e1000456, 2009.

Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):1–15, 2017.

Douglas M. Hawkins and R.A.J. Wixley. A note on the transformation of chi-squared variables to normality. *The American Statistician*, 40(4):296–298, 1986.

Nicholas A. Heard and Patrick Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018.

James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016.

Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

Ruth Heller, Yair Heller, Shachar Kaufman, Barak Brill, and Malka Gorfine. Consistent distribution-free k-sample and independence tests for univariate random variables. *The Journal of Machine Learning Research*, 17(1):978–1031, 2016.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

Wassily Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, pages 546–557, 1948.

Marius Hofert and R. Wayne Oldford. Visualizing dependence in high-dimensional data: An application to S&P 500 constituent data. *Econometrics and Statistics*, 8:161–183, 2018.

Marius Hofert and R. Wayne Oldford. Zigzag expanded navigation plots in R: The R package zenplots. *Journal of Statistical Software*, 95(1):1–44, 2020.

Marius Hofert, Kurt Hornik, and Alexander J. McNeil. *qrmdata: Data Sets for Quantitative Risk Management Practice*, 2022. URL https://CRAN.R-project.org/package=qrmdata. R package version 2022-05-31-1.

EA Housworth and FW Stahl. Crossover interference in humans. *The American Journal of Human Genetics*, 73(1):188–197, 2003.

Tadeusz Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

JAX. JAX Stock #000664. Technical report, The Jackson Laboratory, 2022. URL https://www.jax.org/strain/000664.

Bo Jiang and Jun S. Liu. Sliced inverse regression with variable selection and interaction detection. *arXiv preprint arXiv:1304.4056*, 652:17, 2013.

Bo Jiang, Chao Ye, and Jun S. Liu. Nonparametric k-sample tests via dynamic slicing. *Journal of the American Statistical Association*, 110(510):642–653, 2015.

James M Joyce. Kullback-Leibler divergence. In *International Encyclopedia of Statistical Science*, pages 720–722. Springer, 2011.

Harry Khamis. Measures of association: How to choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162, 2008.

Justin B. Kinney and Gurinder S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9): 3354–3359, 2014a.

Justin B. Kinney and Gurinder S. Atwal. Reply to Reshef et al.: Falsifiability or bust. *Proceedings of the National Academy of Sciences*, 111(33):E3364–E3364, 2014b.

Mikko Kivikoski, Pasi Rastas, Ari Löytynoja, and Juha Merilä. Predicting recombination frequency from map distance. *Heredity*, 130(3):114–121, 2023.

Theo A Knijnenburg, Lodewyk FA Wessels, Marcel JT Reinders, and Ilya Shmulevich. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.

Daniel C. Koboldt, Karyn Meltz Steinberg, David E. Larson, Richard K. Wilson, and Elaine R. Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

Mehmet Kocak. Meta-analysis of univariate p-values. *Communications in Statistics – Simulation and Computation*, 46(2):1257–1265, 2017.

Roger Koenker. *quantreg: Quantile Regression*, 2023. URL https://cran.r-project.org/web/package=quantreg. R package version 5.97.

Spyros Konstantopoulos. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, 2(1):61–76, 2011.

James A. Koziol and Michael D. Perlman. Combining independent chi-squared tests. *Journal of the American Statistical Association*, 73(364):753–763, 1978.

Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, 2009.

H.O. Lancaster. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3(1):20–33, 1961.

Eric S. Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.

Kenneth Lange. *Mathematical and statistical methods for genetic analysis*, volume 488. Springer, 2002.

Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, 2009.

Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

Seung-Chun Lee and Moon Yul Huh. A measure of association for complex data. *Computational statistics & data analysis*, 44(1-2):211–222, 2003.

J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

Albert M. Liebetrau. *Measures of association*. Sage, 1983.

Ramon C. Littell and J. Leroy Folks. Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66(336):802–806, 1971.

Ying Liu, Victor de la Pena, and Tian Zheng. Kernel-based measures of association. *WIREs Computational Statistics*, 10(2), 2018.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *arXiv preprint arXiv:1304.7717*, 2013.

Thomas M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004.

Philipp W Messer. Slim: simulating evolution with selection and linkage. *Genetics*, 194 (4):1037–1039, 2013.

Tamás F. Móri and Gábor J. Székely. Four simple axioms of dependence measures. *Metrika*, 82(1):1–16, 2019.

Govind S. Mudholkar and E.O. George. The logit statistic for combining probabilities – an overview. Technical report, Rochester University Department of Statistics, 1977.

Onuttom Narayan and B. Sriram Shastry. Generalized Toeplitz–Hankel matrices and their application to a layered electron gas. *Journal of Physics A: Mathematical and Theoretical*, 54(17):175201, 2021.

NCBI. NCBI dbSNP Build 155, 2021. URL https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi.

NCBI/NLM. dbGaP: database of Genotypes and Phenotypes, 2023. URL https://www.ncbi.nlm.nih.gov/gap.

Michael A Newton. Introducing the discussion paper by Székely and Rizzo. *The Annals of Applied Statistics*, 3(4):1233–1235, 2009.

Sharon-Lise T Normand. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999.

Dale R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

Art B Owen. Karl Pearson's meta-analysis revisited. *The Annals of Statistics*, 37(6B): 3867–3892, 2009.

Hervè Pagés. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs, 2023. URL https://bioconductor.org/packages/release/bioc/html/BSgenome.html.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Karl Pearson. On the χ^2 test of goodness of fit. *Biometrika*, 14(1/2):186–191, 1922.

Karl Pearson. Further note on the χ^2 test of goodness of fit. *Biometrika*, 14(3-4):418–418, 1923.

Karl Pearson. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25(3/4):379–410, 1933.

Bo Peng, Marek Kimmel, and Christopher I Amos. *Forward-time population genetics simulations: methods, implementation, and applications*. John Wiley & Sons, 2012.

Robin L. Plackett. Karl Pearson and the chi-squared test. *International Statistical Review*, pages 59–72, 1983.

William Poole and David Gibbs. *EmpiricalBrownsMethod: Uses Brown's method to combine p-values from dependent tests*, 2015. URL https://github.com/IlyaLab/CombiningDependentPvaluesUsingEBM.git. R package version 1.28.0.

William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard, and Theo A. Knijnenburg. Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics*, 32(17):i430–i436, 2016.

Adam Rahman. *Preserving measures structure during generation and reduction of multivariate point configurations*. PhD thesis, University of Waterloo, 2018. URL http://hdl.handle.net/10012/13365.

Matthew Reimherr and Dan L. Nicolae. On quantifying dependence: a framework for developing interpretable measures. *Statistical Science*, 28(1):116–130, 2013.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.

David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

David N. Reshef, Yakir A. Reshef, Michael Mitzenmacher, and Pardis C. Sabeti. Cleaning up the record on the maximal information coefficient and equitability. *Proceedings of the National Academy of Sciences*, 111(33):E3362–E3363, 2014.

David N. Reshef, Yakir A. Reshef, Pardis C. Sabeti, and Michael Mitzenmacher. An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1):123–155, 2018.

Kent A Riemondy, Ryan M Sheridan, Austin Gillen, Yinni Yu, Christopher G Bennett, and Jay R Hesselberth. valr: Reproducible genome interval analysis in R. *F1000Research*, 6, 2017.

L.B. Rowe, J.H. Nadeau, R. Turner, W.N. Frankel, V.A. Letts, J.T. Eppig, M.S.H. Ko, S.J. Thurston, and E.H. Birkenmeier. Maps from two interspecific backcross DNA panels available as a community genetic mapping resource. *Mammalian Genome*, 5(5):253–274, 1994.

Chris Salahub. *AssocBin: Measuring Association with Recursive Binning*, 2023a. URL https://cran.r-project.org/package=AssocBin. R package version 0.1.

Chris Salahub. *PoolBal: Balancing Central and Marginal Rejection of Pooled p-Values*, 2023b. URL https://cran.r-project.org/package=PoolBal. R package version 0.1.

Daria Salyakina, Shaun R. Seaman, Brian L. Browning, Frank Dudbridge, and Bertram Müller-Myhsok. Evaluation of Nyholt's procedure for multiple testing correction. *Human Heredity*, 60(1):19–25, 2005.

Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

Markus S Schröder, Aedín C Culhane, John Quackenbush, and Benjamin Haibe-Kains. survcomp: an r/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

Berthold Schweizer and Edward F. Wolff. On nonparametric measures of dependence for random variables. *Annals of Statistics*, 9(4):879–885, 1981.

David W. Scott. A note on choice of bivariate histogram bin shape. *Journal of Official Statistics*, 4(1):47, 1988.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Min Shi, David M Umbach, Alison S Wise, and Clarice R Weinberg. Simulating autosomal genotypes with realistic linkage disequilibrium and a spiked-in genetic effect. *BMC bioinformatics*, 19:1–10, 2018.

Karl Friedrich Siburg and Pavel A. Stoimenov. A measure of mutual complete dependence. *Metrika*, 71(2):239–251, 2010.

Andrew F. Siegel. The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. *Biometrika*, 66(2):381–386, 1979.

David Siegmund and Benjamin Yakir. *The statistics of gene mapping*. Springer Science & Business Media, 2007.

S.D. Silvey. On a measure of association. *The Annals of Mathematical Statistics*, pages 1157–1166, 1964.

R. John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

Noah Simon and Robert Tibshirani. Comment on "Detecting Novel Associations In Large Data Sets" by Reshef et al., Science dec 16, 2011. *arXiv preprint arXiv:1401.7645*, 2014.

Bimal K Sinha, Joachim Hartung, and Guido Knapp. *Statistical meta-analysis with applications*. John Wiley & Sons, 2011.

Abe Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pages 1–14, 1996.

Terry P Speed. Genetic map functions. *Encyclopedia of Biostatistics*, 3, 2005.

Michael A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

Stephen M Stigler. Francis Galton's account of the invention of correlation. *Statistical Science*, pages 73–79, 1989.

Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), vol. 1.* Princeton University Press, 1949.

Karen L Svenson, Randy Von Smith, Phyllis A Magnani, Heather R Suetin, Beverly Paigen, Jurgen K Naggert, Renhua Li, Gary A Churchill, and Luanne L Peters. Multiple trait measurements in 43 inbred mouse strains capture the phenotypic diversity characteristic of human populations. *Journal of applied physiology*, 102(6):2369–2378, 2007.

Gábor J Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, pages 1236–1265, 2009.

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

Olivier Thas and Jean-Pierre Ottoy. A nonparametric test for independence based on sample space partitions. *Communications in Statistics-Simulation and Computation*, 33 (3):711–728, 2004.

Leonard Henry Caleb Tippett. *The Methods of Statistics*. London: Williams & Norgate Ltd., 1931.

Dag Tjøstheim, Håkon Otneim, and Bård Støve. Statistical dependence: Beyond pearson's ρ. *Statistical science*, 37(1):90–109, 2022.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

Carl Veller, Nathaniel B Edelman, Pavitra Muralidhar, and Martin A Nowak. Variation in genetic relatedness is determined by the aggregate recombination process. *Genetics*, 216(4):985–994, 2020.

Murugesan Venkatapathi and M. Harlprasad. Circulant decomposition of a matrix and the eigenvalues of toeplitz type matrices. *arXiv preprint arXiv:2105.14805*, 2022.

Peter M. Visscher and Michael E. Goddard. From R.A. Fisher's 1918 paper to GWAS a century later. *Genetics*, 211(4):1125–1130, 2019.

W. Allen Wallis. Compounding probabilities from independent significance tests. *Econometrica, Journal of the Econometric Society*, 10(3/4):229–248, 1942.

Xiaosong Wang and Beverly Paigen. Genome-wide search for new genes controlling plasma lipid concentrations in mice and humans. *Current opinion in lipidology*, 16(2):127–137, 2005.

Carrie L. Welch, Yu-Rong Xia, Ishaiahu Shechter, Robert Farese, Margarete Mehrabian, Shahab Mehdizadeh, Craig H. Warden, and Aldons J. Lusis. Genetic regulation of cholesterol homeostasis: chromosomal organization of candidate genes. *Journal of Lipid Research*, 37(7):1406–1421, 1996.

Margareta Westberg. Combining independent statistical tests. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 34(3):287–296, 1985.

Thomas White, Daniel Noble, Alistair Senior, W. Kyle Hamilton, and Wolfgang Viechtbauer. *metadat: Meta-Analysis Datasets*, 2022. URL https://CRAN.R-project.org/package=metadat. R package version 1.2-0.

Bryan Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156, 1951.

Leland Wilkinson and Graham Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17(2):473–491, 2008.

Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164. IEEE, 2005.

Daniel J. Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

Sewall Wright. Inbreeding and homozygosis. *Proceedings of the National Academy of Sciences*, 19(4):411–420, 1933.

James J Yang, Jia Li, L Keoki Williams, and Anne Buu. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics*, 17:1–11, 2016.

Lynn Yi and Lior Pachter. *aggregation: p-Value Aggregation Methods*, 2018. URL https://cran.r-project.org/package=aggregation. R package version 1.0.1.

G. Udny Yule. On the application of the χ^2 method to association and contingency tables, with experimental illustrations. *Journal of the Royal Statistical Society*, 85(1):95–104, 1922.

Dmitri V. Zaykin, Lev A. Zhivotovsky, Wendy Czika, Susan Shao, and Russell D. Wolfinger. Combining p-values in large-scale genomics experiments. *Pharmaceutical Statistics*, 6(3): 217–226, 2007.

Hong Zhang, Tiejun Tong, John Landers, and Zheyang Wu. Tfisher: a powerful truncation and weighting procedure for combining p-values. *Annals of Applied Statistics*, 14(1):178–201, 2020.

Hongyu Zhao and Terence P. Speed. On genetic map functions. *Genetics*, 142(4):1369–1377, 1996.

Jing Hua Zhao. gap: Genetic analysis package. *Journal of Statistical Software*, 23:1–18, 2008.

Wenjun Zheng, Dejian Lai, and K. Lance Gould. A simulation study of a class of nonparametric test statistics: a close look of empirical distribution function-based tests. *Communications in Statistics-Simulation and Computation*, pages 1–17, 2021.

APPENDICES

Appendix A

Structural means for robustness to missing data

The methods adjusting for dependence in Chapter 6 all require a genetic correlation or linkage disequilibrium (LD) matrix $\mathbf{\Sigma}$ though missing data often forces us to use the pairwise correlation/LD matrix $\mathbf{\Sigma}'$ which may not be positive definite. To minimize the impact of this missing data, it is natural to reduce the difference between $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$ in a suitable norm. In this appendix, it is shown that a simple pre-processing step which converts $\mathbf{\Sigma}'$ to a Frobenius-optimal matrix approximating $\mathbf{\Sigma}$ given a known theoretical correlation structure drastically improves robustness to missing data at the expense of some bias when the data are complete. In the following,$\mathbf{\Sigma}$ and the term LD matrix will be used interchangeably.[1]

Recall that assessing the impact of a region of the genome on a measured trait requires adjusting test statistics to account for the LD between marker regions (Lander and Botstein, 1989; Uffelmann et al., 2021), that is the correlation of genetic markers over successive generations. For regions i and j on chromosomes c_i and c_j, $LD(i,j)$ can be computed using the equation

$$[\mathbf{\Sigma}]_{ij} = LD(i,j) = I(c_i = c_j)\gamma e^{-\frac{d(i,j)}{50}}$$

where $d(i,j)$ is an additive measure of genetic distance between markers i and j and γ is a constant determined by population characteristics. $LD(i,j)$ is constant in $d(i,j)$, implying a theoretical structure to the LD matrix which can be computed in advance. Despite

[1] A version of this appendix is currently under review for submission to IEEE Transactions on Computational Biology and Bioinformatics with a co-author, Jeffrey Uhlmann.

this theoretical knowledge, this matrix is still computed on genomic samples in order to determine the observed correlation between tests, $\hat{\mathbf{\Sigma}}$, and use this to adjust for dependence in multiple testing. When data are missing, this is computed pairwise to give $\hat{\mathbf{\Sigma}}'$ which may differ further from $\mathbf{\Sigma}$, adding additional error.

Rather than using $\hat{\mathbf{\Sigma}}'$ directly, the theoretical structure of $\mathbf{\Sigma}$ can be leveraged to compute a Frobenius-optimal approximating matrix given $\hat{\mathbf{\Sigma}}'$ by taking means along the theoretical level sets of $\mathbf{\Sigma}$. This guarantees adherence to the prescribed theoretical structure and reduces the impact of individual missing entries by taking means over the observed correlations. First, consider the problem generally.

A.1 Approximating a general matrix

Suppose we would like to approximate the $n \times n$ matrix

$$\mathbf{M} = \begin{bmatrix} m_{00} & m_{01} & \dots & m_{0,n-1} \\ m_{10} & m_{11} & \dots & m_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-1,0} & m_{n-1,1} & \dots & m_{n-1,n-1} \end{bmatrix} \tag{A.1}$$

$m_{ij} \in \mathbb{C}$ with a matrix $\mathbf{T} \in \mathbb{C}^{n \times n}$ with a particular structure for computational or analytical reasons, for example circulant $\mathbf{T}$ for preconditioning Chan (1988); Venkatapathi and Harlprasad (2022) and Toeplitz-Hankel $\mathbf{T}$ for physical modelling Narayan and Shastry (2021). For any application, the approximating matrix $\mathbf{T}$ should be optimal by some measure, commonly the Frobenius norm of the difference of matrices

$$||\mathbf{T} - \mathbf{M}||_F = \sqrt{\text{trace}\left((\mathbf{T} - \mathbf{M})^*(\mathbf{T} - \mathbf{M})\right)}, \tag{A.2}$$

where $\mathbf{A}^*$ is the conjugate matrix of $\mathbf{A} \in \mathbb{C}^{n \times n}$. Minimizing this for a circulant $\mathbf{T}$ is the express goal of Chan (1988) and was noted as a positive feature of the approximation in Venkatapathi and Harlprasad (2022). Both of these are particular cases of a more general result that can construct an optimal matrix approximation for an arbitrary structure.

Structure in $\mathbf{T} \in \mathbb{C}^{n \times n}$ is limited here to cases where equality of the entries t_{ij}, $0 \leq i, j \leq n-1$ follows a pattern in i and j. Explicitly, given an index function $f : \{0, 1, \dots, n-1\}^2 \mapsto \{0, 1, 2, \dots, K\}$ that provides membership of the i, j entry to a level set, $\mathbf{T}$ is structured with index function $f(i, j)$ if

$$t_{ij} = t_{f(i,j)}. \tag{A.3}$$

This implies a k^{th} index set $\mathcal{T}_k = \{(i,j)|f(i,j) = k\}$ with cardinality $|\mathcal{T}_k| = n_k > 0$.

The index function $f(\cdot,\cdot)$ defines the structure of $\mathbf{T}$ by defining the $\mathcal{T}_k$. Common structures and corresponding index functions are shown in Table A.1, though these functions are not unique. Many candidate functions define identical index sets, Hankel matrices for example can take either $f(i,j) = j + i$ or $f(i,j) = 2(n-1) - j - i$.

Table A.1: Some common examples of structured index functions.

Structure	$f(i,j)$
Circulant	$(i-j) \bmod n$
Toeplitz	$j - i + n$
Hankel	$i + j$

A.2 Optimizing the Frobenius norm

Using the notation defined above, consider the following theorem.

Theorem 4 (Means minimize $||\mathbf{T} - \mathbf{M}||_F$). *$\mathbf{T}_M$ with*

$$t_{ij} = t_{f(i,j)} = \overline{m}_{f(i,j)} \tag{A.4}$$

is the Frobenius-optimal structured matrix with index function $f(i,j)$ that approximates $\mathbf{M}$, where

$$\overline{m}_k := \frac{1}{n_k} \sum_{\mathcal{T}_k} m_{ij}. \tag{A.5}$$

is the mean of entries in $\mathbf{M}$ over the k^{th} index set. The minimum error, $||\mathbf{T}_M - \mathbf{M}||_F$, is proportional the total within-group standard deviation of entries in $\mathbf{M}$ over all index sets.

Proof. Take $\overline{m}_k$ to be the mean of entries in $\mathbf{M}$ for the k^{th} index set as in Equation A.5, define the vector of all such means

$$\overline{\mathbf{m}} = (\overline{m}_0, \overline{m}_1, \ldots, \overline{m}_K)^{\mathsf{T}}.$$

Further, denote the vector of unique t_k as

$$\mathbf{t} = (t_0, t_1, \ldots, t_K)^{\mathsf{T}}$$

and the diagonal matrix of n_k as

$$\mathbf{N} = \operatorname{diag}(n_0, n_1, \ldots, n_K).$$

As Equation A.2 is always positive, any $\mathbf{T}$ which minimizes $||\mathbf{T} - \mathbf{M}||_F$ will also minimize $||\mathbf{T} - \mathbf{M}||_F^2$. Expanding gives

$$\begin{aligned} \operatorname{trace}\left((\mathbf{T} - \mathbf{M})^*(\mathbf{T} - \mathbf{M})\right) = & \operatorname{trace}\mathbf{M}^*\mathbf{M} \\ & - \operatorname{trace}\mathbf{M}^*\mathbf{T} \\ & - \operatorname{trace}\mathbf{T}^*\mathbf{M} \\ & + \operatorname{trace}\mathbf{T}^*\mathbf{T}. \end{aligned} \tag{A.6}$$

$\mathbf{M}^*\mathbf{M}$ is constant in $\mathbf{T}$, so can be ignored. The latter three terms can be considered individually to give $\operatorname{trace}\mathbf{T}^*\mathbf{T} = \sum_{k=0}^{K} n_k t_k^* t_k$, $\operatorname{trace}\mathbf{M}^*\mathbf{T} = \sum_{k=0}^{K} n_k t_k \overline{m}_k^*$, and $\operatorname{trace}\mathbf{T}^*\mathbf{M} = \sum_{k=0}^{K} n_k t_k^* \overline{m}_k$. So we seek to minimize

$$F(\mathbf{t}) = \sum_{k=0}^{K} n_k t_k^* t_k - \sum_{k=0}^{K} n_k t_k^* \overline{m}_k - \sum_{k=0}^{K} n_k t_k \overline{m}_k^*,$$

which can be written in matrix form as

$$\begin{aligned} F(\mathbf{t}) &= \mathbf{t}^*\mathbf{N}\mathbf{t} - \mathbf{t}^*\mathbf{N}\overline{\mathbf{m}} - \overline{\mathbf{m}}^*\mathbf{N}\mathbf{t} \\ &= (\mathbf{t} - \overline{\mathbf{m}})^*\mathbf{N}\,(\mathbf{t} - \overline{\mathbf{m}}) - \overline{\mathbf{m}}^*\mathbf{N}\overline{\mathbf{m}}. \end{aligned}$$

As $n_k > 0$ for all $k = 0, 1, \ldots, K$, $\mathbf{N}$ is positive definite, and so the quadratic form $\mathbf{x}^*\mathbf{N}\mathbf{x}$ has a minimum of zero when $\mathbf{x} = \mathbf{0}$. Therefore $F(\mathbf{t})$ is minimized for $\mathbf{t} = \overline{\mathbf{m}}$ and has a minimum of

$$F(\overline{\mathbf{m}}) = -\overline{\mathbf{m}}^*\mathbf{N}\overline{\mathbf{m}} = -\sum_{k=0}^{K} n_k ||\overline{m}_k||^2. \tag{A.7}$$

So $\mathbf{T}_M$ is the Frobenius-optimal structured matrix with index function $f(i,j)$ approximating $\mathbf{M}$. The residual $\mathbf{T}_M - \mathbf{M}$ has a squared Frobenius norm of

$$\begin{aligned} ||\mathbf{T}_M - \mathbf{M}||_F^2 &= \sum_{k=0}^{K} n_k \left(\sum_{\mathcal{T}_k} \frac{||m_{ij}||^2}{n_k} - ||\overline{m}_k||^2 \right) \\ &= \sum_{k=0}^{K} n_k \sigma_k^2 \end{aligned} \tag{A.8}$$

where $\sigma_k^2 = \frac{1}{n_k} \sum_{\mathcal{T}_k} (m_{ij} - \overline{m}_k)^2$ is the variance of the m_{ij} in the index set $\mathcal{T}_k$. Therefore we have

$$\frac{1}{\sqrt{n}} ||\mathbf{T}_M - \mathbf{M}||_F = \sqrt{\sum_{k=0}^{K} \frac{n_k}{n} \sigma_k^2},$$

which is the total within-group standard deviation from the structured means. □

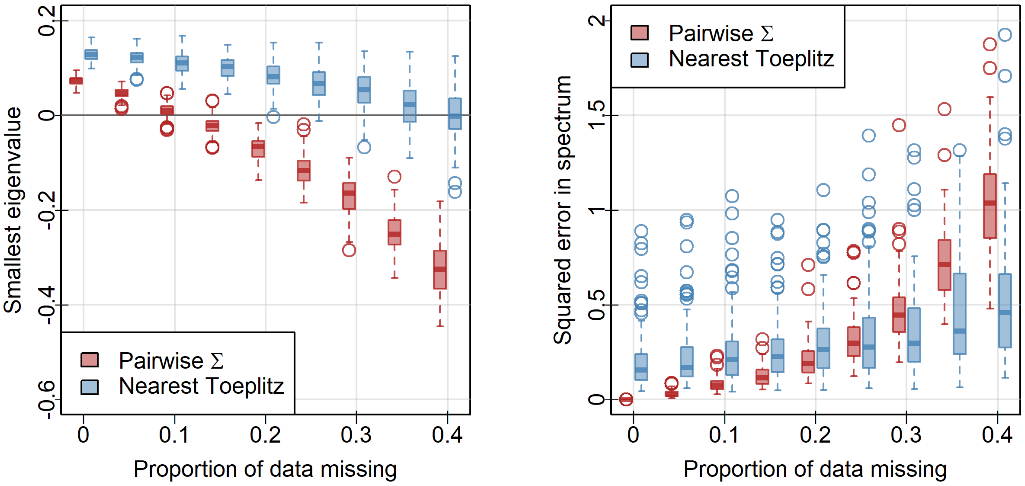
A.3 Applying the optimal approximation to simulated data

The first test of this proposal is a simulation study where 100 synthetic populations of 100 individuals were generated measured at 20 markers with $d(i, j) = 15$ cM and $c_i = c_j = 1$ for each pair. This setting gives, in theory, a Toeplitz LD matrix. For each simulated population, the complete data was first used to compute $\hat{\mathbf{\Sigma}}$ and determine its eigenvalues. Increasing proportions of observations were then removed completely at random from the data to simulate different data completeness, and at each proportion the $\hat{\mathbf{\Sigma}}'$ was computed and used to generate the nearest Toeplitz matrix based on Theorem4. The minimum eigenvalue and sum of squared errors in the ordered eigenvalues from the complete data LD matrix were computed and recorded for both the pairwise LD matrix $\hat{\mathbf{\Sigma}}'$ and the nearest Toeplitz matrix $\hat{\mathbf{T}}$. Figure A.1 displays the result for proportions missing ranging from 0 to 0.4.

Figure A.1(a) shows that the nearest Toeplitz is more robust to negative eigenvalues than the pairwise LD matrix. A vast majority of the 100 simulated populations have nearest Toeplitz matrices with no negative eigenvalues until more than a third of the data is missing, and even then only about 25% produce negative eigenvalues. In contrast, the pairwise LD matrix has negative eigenvalues more than 75% of the time when as little as 15% of the data is missing. Figure A.1(b) indicates this robustness also extends to the sum of squared differences between ordered eigenvalues, which does not depend greatly on the completeness of the data for the nearest Toeplitz matrix but does for the pairwise LD. The cost of this robustness is potential bias when the data is (mostly) complete.

A.4 Using real data

Following the simulated study of the previous section, we now corroborate our findings by replicating the experiment on real data. The JAX BSB data from Rowe et al. (1994) in-



(a)

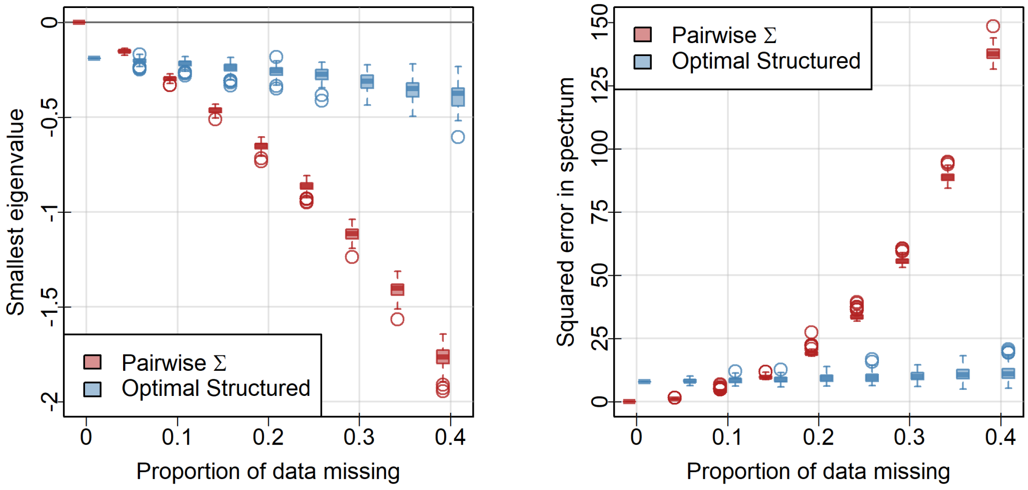
(b)

Figure A.1: Paired boxplots of the (a) minimum eigenvalues and (b) sum of squared errors in the ordered eigenvalues for the pairwise LD matrix and the nearest Toeplitz matrix by the proportion of data missing. The nearest Toeplitz, displayed to the right of the corresponding proportion for each pair of boxplots, is more robust to missing data than the pairwise LD matrix, displayed to the left for each pair, but is biased when the data are complete.

cluded in the `toyGenomeGenR` package contains partial measurement of 5951 markers across all chromosomes on 94 mice. Complete observations present on 2624 of the markers and were used to create a 2624×2624 observed correlation matrix. In contrast to the previous simulated example, these markers are not positioned uniformly across chromosomes, the cM distances between adjacent markers differ greatly.

Instead of focusing on all of the markers, however, we make the real data comparable to the simulated case by considering only those markers on chromosome 1. This leaves 199 markers measured across the chromosome with adjacent distances ranging from less than 0.01 cM to 4.37 cM. The cM distances were used to generate a theoretical correlation/LD matrix, the theoretical correlations were rounded to two decimal places, and the rounded correlations were treated as the level sets of the theoretical structured matrix. Note that unlike the examples illustrated above, these level sets do not correspond with any named structured matrix. Nonetheless, Theorem 4 dictates that the optimal structured approximation in the Frobenius norm for this unnamed structure is given by means computed over the level sets. An unfortunate consequence of the lack of correspondence with a named matrix is that there is no guarantee that the result will be a valid correlation matrix.

These steps give us the observed matrix $\hat{\boldsymbol{\Sigma}}$ on the full data and the theoretical correlation matrix $\boldsymbol{\Sigma}$ based on cM distances. Once again, marker measurements were then removed completely at random from the JAX BSB data 100 times for each of a range of proportions of missingness, the pairwise LD matrix and the optimal structured approximation based on the theoretical level sets were generated, and both were compared to the observed matrxi $\hat{\boldsymbol{\Sigma}}$ using the minimum eigenvalue and sum of squared errors. Figure A.2 results.



(a) (b)

Figure A.2: Paired boxplots of the (a) minimum eigenvalues and (b) sum of squared errors in the ordered eigenvalues for the pairwise LD matrix (on the left) and the optimal structured matrix (on the right) by the proportion of data missing. The bias in the optimal structured matrix is more serious in this case than the simulated example: negative eigenvalues are produced for the complete data.

First, note that the larger correlation matrix for the real data example has resulted in less variability in the summaries of eigenvalues for both the pairwise LD and optimal structured matrices, accentuating differences in perfomance. Just as in the simulated case, the optimal structured matrix proves far more robust to missing data. Both the minimum eigenvalue and the sum of squared errors barely change on average as the proportion of missing data increases. In contrast, the pairwise LD matrix has an error that grows in the proportion of missing data and a minimum eigenvalue which continues to decrease as data is removed. When as little as 15% of the data is missing, the optimal structured matrix is better by both metrics.

The bias of the optimal approximation is more severe in the real data than the simulated data, however. For a mostly complete data set, the optimal structured matrix produces negative eigenvalues and it does not correspond with a correlation matrix. In cases with nearly complete data, an optimal structured approximation should only be used if the structure is Toeplitz, circulant, or of another class with known qualities to its eigenspectrum. When large proportions of the data are missing, however, applying structural means greatly improves the stability of estimates.

Appendix B

Demonstrating toyGenomeGen

The following is a short introduction to the features of `toyGenomeGen` that loosely follows the demos included in the package.

B.1 Constructing genomes

We begin by creating a genome using the default options and specifying only the number of markers on each chromosome.

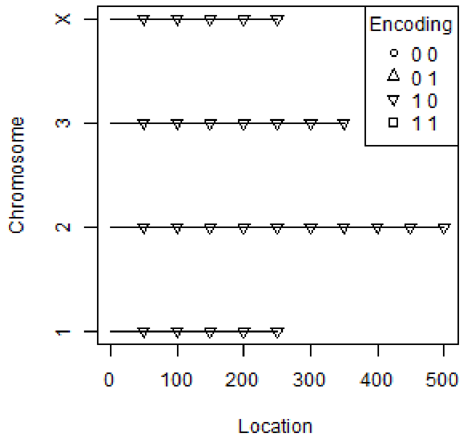
```
## define the marker distribution by chromosome
markerDist <- c("1" = 5, "2" = 10, "3" = 7, "X" = 5)
## construct a genome based on this distribution
g <- simGenome(markerDist)
```

The `print.genome` method gives a basic summary of `g`.

```
> g
A genome object encoding 30 markers across 4 chromosomes, distributed:
  5 10 7 5
```

The `plot.genome` method shows more detail.

```
plot(g, add.legend = TRUE)
```



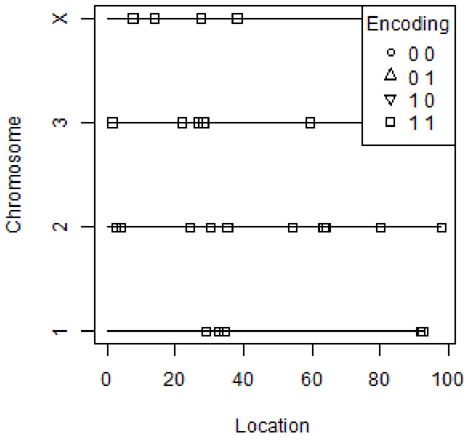
Different point types quickly communicate the encodings to the user, lines group the chromosomes, and point locations show where markers have been measured. By changing the `elevs` and `epch` arguments, both the presumed encodings and points displaying them can be changed. Note that the encoding strings matched back to `elevs` are generated from `g` by pasting the columns of `g$encoding` together row-wise with a space between them. By default `simGenome` presumes evenly-spaced markers across each chromosome with 50 cM of separation and generates **X** such that one column contains only ones and the other only zeros, though the order of these columns is random.

Very different genomes can be obtained by changing the defaults of `simGenome`. Inspection shows that the default marker generation function is `markerHybrid` and the default location generation function is `locationRegular`:

```
> formals(simGenome)
$nmark


$alleles


$markerFuns
markerHybrid

$locFuns
locationRegular
```

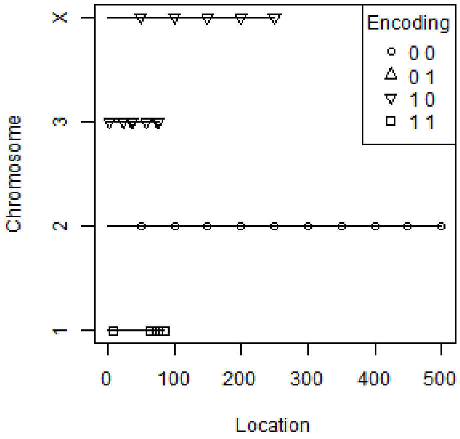
Calling `simGenome` again but using one of the other marker functions, say `markerPureDom`, and the other location generation function, `locationUniform`, gives a radically different genetic structure.

```
g <- simGenome(markerDist, markerFuns = markerPureDom,
               locFuns = locationUniform)
plot(g, add.legend = TRUE)
```



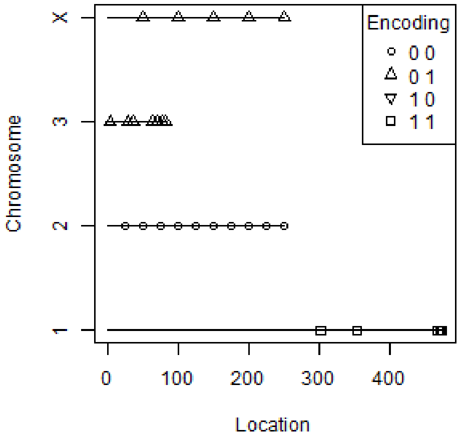
`locationUniform` generates markers randomly along each chromosome based on a uniform distribution between 0 and 100 by default and `markerPureDom` produces $\mathbf{X}$ with every entry 1. The complementary function `markerPureRec` produces $\mathbf{X}$ with every entry 0. For even greater control, we can supply a list of functions to apply to each chromosome separately.

```
g <- simGenome(markerDist,
               markerFuns = list(markerPureDom, markerPureRec,
                                 markerHybrid, markerHybrid),
               locFuns = list(locationUniform, locationRegular,
                              locationUniform, locationRegular))
plot(g, add.legend = TRUE)
```



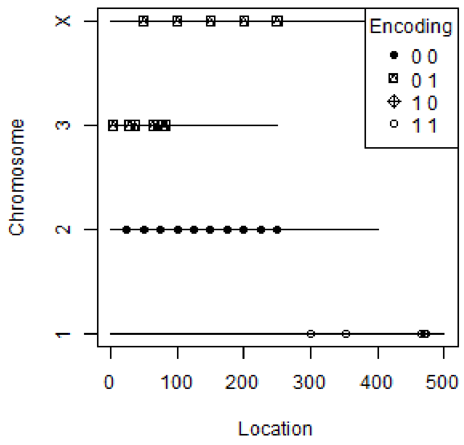
The default behaviours can be modified by wrapping the appropriate function in a closure.

```
uni100500 <- function(n) locationUniform(n, min = 100, max = 500)
reg25 <- function(n) locationRegular(n, delta = 25)
g <- simGenome(markerDist,
               markerFuns = list(markerPureDom, markerPureRec,
                                 markerHybrid, markerHybrid),
               locFuns = list(uni100500, reg25,
                              locationUniform, locationRegular))
plot(g, add.legend = TRUE)
```



Finally, plot characteristics such as the length of chromosomes and the point type to use for encodings can be changed by specifying the `chrLens` and `epch` arguments.

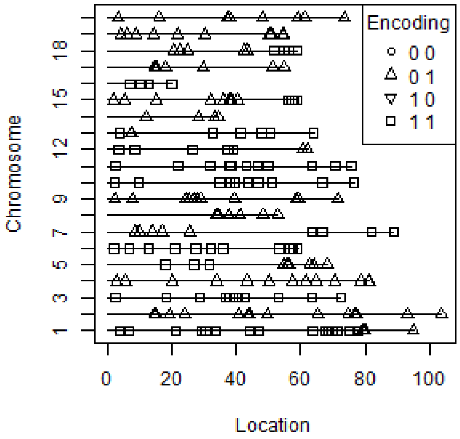
```
plot(g, add.legend = TRUE, chrLens = c(500, 400, 250, 400),
     epch = c(19, 14, 9, 1))
```



Users can define their own marker and location functions to pass into simGenome for complete control over the generation of a genome object. They must accept a single argument n and return a two-column matrix with n (for custom marker functions) or a numeric vector of length n (for custom location functions). Other helpers exist to convert a data.frame to a genome and to create a genome directly from provided slots with checks to ensure conformity of arguments. These are outlined in the basicDemo.R demo in the toyGenomeGen package.

genomes based on real data are provided by the 11 data sets included in toyGenomeGen. These report the results of backcross experiments carried out with different known strains of inbred mice by several different labs. Loading and inspecting a sample genome from one of these experiments is easy in the package.

```
> data(ucla_bsb)
> ucla_bsb
A population of 67 genomes encoded at 223 markers across 20
chromosomes, distributed:
 21 15 11 16 10 12 11 7 18 13 13 9 8 5 14 4 7 9 12 8
Roughly 3 % of the data is missing.
> x1 <- selectGenome(ucla_bsb, ind = 1)
> plot(x1, add.legend = TRUE)
```



This plot is much more complicated than the earlier examples of `g`, but still displays an important feature of each genome. Recall that in a backcross experiment, we cross the encodings

$$\mathbf{F}_X = \begin{bmatrix} f & f \\ \vdots & \vdots \\ f & f \end{bmatrix}, \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}.$$

where $f \in \{0, 1\}$. In this case only the encodings `0 1` and `1 1` are possible as $f = 1$. If crossing over did not occur every chromosome would be identically `0 1` or `1 1` at every marker position, and the encodings within each chromosome would be identical. Every time a change in point type is observed along a chromosome in this plot, recombination (an odd number of cross overs) occurred in the region between the different point types. See, for example, the change from triangle to square in chromosome 7. Not only does this plot give an overview of the markers in `genome`, it also gives a peek into the distribution of their recombination.

B.2 Simulating recombination

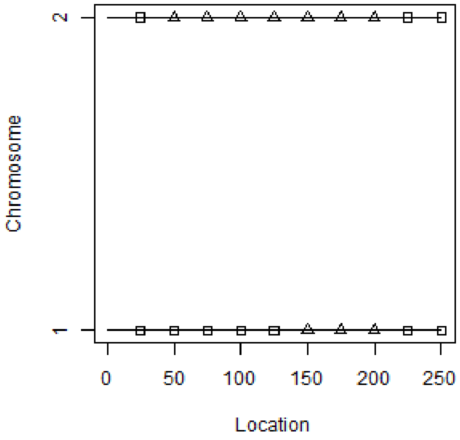
Before simulating a cross, we generate two `genome`s corresponding with the N_2 backcross setting.

```
## start by defining the locations and chromosomes
locs <- rep(list(cumsum(rep(25,10))), 2)
alls <- rep(list(c("A", "a")), 20)
```

```
chr <- factor(rep("1", "2"), each = 10)
## use settings to generate two genomes
g1 <- makeGenome(locs, alls, chr)
g2 <- makeGenome(locs, alls, chr, enc = markerPureDom(20))
```

Generating a new genome based on the cross between them is accomplished by calling the wrapper function `sex` on the pair.

```
o1 <- sex(g1, g2)
plot(o1) # can see recombination in point changes
```

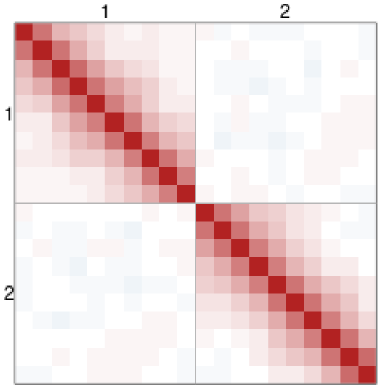


Whole populations can be generated with `replicate` from base R .

```
popsize <- 1000
pop <- asPopulation(replicate(popsize, sex(g1, g2), simplify = FALSE))
```

These populations can be visualized using the `corrImg` function and `addChromosomeLines`.

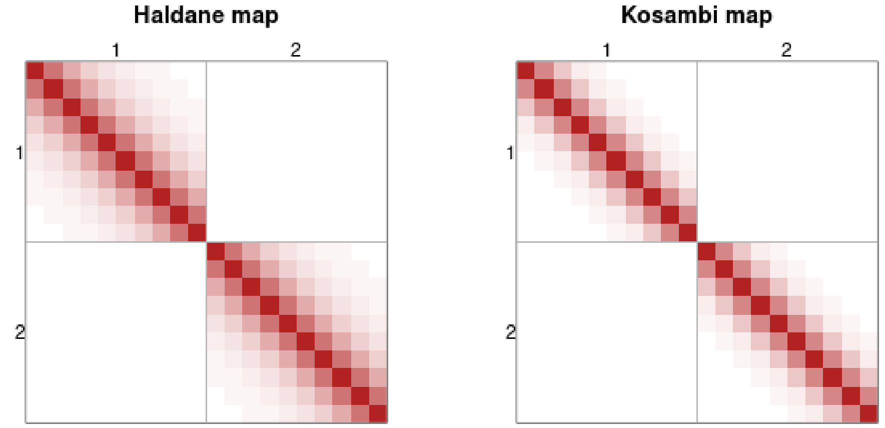
```
## define some constants to control the visualization
pal <- colorRampPalette(c("steelblue", "white", "firebrick"))(49)
corBrks <- seq(-1, 1, length.out = 50)
## plot correlation
corrImg(popCorrelation(pop), col = pal, breaks = corBrks,
        xaxt = "n", yaxt = "n")
addChromosomeLines(pop)
```



To read this plot, note that the hue of this heatmap at the i, j square counted from the top left is given by the diverging palette based on the corresponding break in `corBrks` containing correlation between z_i and z_j. Darker colours indicate stronger correlation, red-shaded cells have positive values, and blue-shaded cells have negative values. The lines demarcating the chromosomes are helpful primarily as guides to separate the intra- and inter-chromosome patterns of correlation.

Theoretical correlations based on a map function are generated with `theoryCorrelation` with the default using Haldane's map and generating correlations by Equation 5.18.

```
corrImg(theoryCorrelation(pop), col = pal, breaks = corBrks,
        xaxt = "n", yaxt = "n", main = "Haldane␣map")
addChromosomeLines(pop)
corrImg(theoryCorrelation(pop, map = mapKosambi), col = pal,
        breaks = corBrks, xaxt = "n", yaxt = "n",
        main = "Kosambi␣map")
addChromosomeLines(pop)
```



As the default settings of sex assume the Haldane map function and independent recombination, it is no surprise that the Haldane map theoretical correlation matrix looks closer to the measured correlation matrix on pop than the Kosambi map.

To change the default recombination behaviour, we pass relevant arguments to the helper meiose inside of sex by setting the optional arguments to sex.

```
> formals(sex)
$genome1


$genome2


$probs1
NULL

$probs2
NULL

$map
mapHaldane

$crossFun
crossIndep
```

The arguments probs1 or probs2 allow users to provide custom probabilities of recombination for each adjacent marker pair, for example gathered experimentally. These probabilities override the provided map function and are passed to crossFun along with the locations of markers on the genome. crossFun defines recombination within mieose, the function that recombines the copies within genome1 and genome2 before mixing them to generate offspring. The default function, crossIndep, assumes independence between crossovers and so is effectively a wrapper for runif:

```
> crossIndep
function (probs, locs)
{
  breaks <- runif(length(probs))
  which(breaks < probs)
}
<bytecode: 0x0000024ab6fd68b8>
<environment: namespace:toyGenomeGenR>
```

The independent assumption underlying this function, though simple, does not always provide a good fit in practice (Broman et al., 2002) and crossovers are often modelled by a renewal process with scaled χ^2 holding times (Housworth and Stahl, 2003; Lange, 2002). crossChi defined below provides an example of how such an renewal process can be implemented in toyGenomeGen.

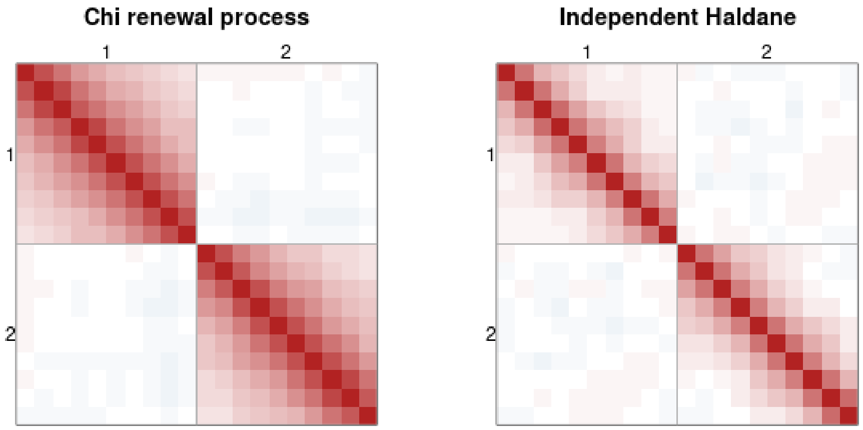
```
crossChi <- function(probs, locs) { # must accept two arguments
  locLen <- length(locs) # largest location
  crosses <- numeric() # cross indices
  currPos <- 100*rchisq(1, df = 2) # current position of process
  while (currPos < locs[1]) { # only care within marker range
    new <- 100*rchisq(1, df = 2) # next holding time
    currPos <- currPos + new # update break spot
  }
  while (currPos <= locs[locLen]) { # breaks within markers
    crosses <- c(crosses, sum(locs < currPos)) # index of split
    new <- 100*rchisq(1, df = 2)
    currPos <- currPos + new
  }
  crosses
}
```

Using crossChi, we can generate a second population using the renewal process model of cross over recombination.

```
pop2 <- asPopulation(replicate(popsize,
                     sex(g1, g2, crossFun = crossChi),
                     simplify = FALSE))
```

Just as before, the correlation matrix under the additive map can be used to visualize this and compare it to the previous population generated assuming independent cross overs and Haldane's map.

```
corrImg(popCorrelation(pop2), col = pal, breaks = corBrks,
        xaxt = "n", yaxt = "n", main = "Chi␣renewal␣process")
addChromosomeLines(pop2)
corrImg(popCorrelation(pop), col = pal, breaks = corBrks,
        xaxt = "n", yaxt = "n", main = "Independent␣Haldane")
addChromosomeLines(pop)
```



Changing the dynamics of recombination in `toyGenomeGen` is as simple as defining a function like `crossChi`.