RICE UNIVERSITY

Minimum Distance Estimation in Categorical Conditional Independence Models

by

David John Kahle

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Javier Rojo, Chair
Professor of Statistics

Hadley Wickham
Assistant Professor and Dobelman Family
Junior Chair of Statistics

Brendan Hassett
Professor and Chair of Mathematics

Yin Zhang
Professor of Computational and Applied
Mathematics

Houston, Texas

August, 2011

ABSTRACT

Minimum Distance Estimation in Categorical Conditional Independence Models

by

David John Kahle

One of the oldest and most fundamental problems in statistics is the analysis of cross-classified data called contingency tables. Analyzing contingency tables is typically a question of association – do the variables represented in the table exhibit special dependencies or lack thereof? The statistical models which best capture these experimental notions of dependence are the categorical conditional independence models; however, until recent discoveries concerning the strongly algebraic nature of the conditional independence models surfaced, the models were widely overlooked due to their unwieldy implicit description. Apart from the inferential question above, this thesis asks the more basic question – suppose such an experimental model of association is known, how can one incorporate this information into the estimation of the joint distribution of the table?

In the traditional parametric setting several estimation paradigms have been developed over the past century; however, traditional results are not applicable to arbitrary categorical conditional independence models due to their implicit nature. After laying out the framework for conditional independence and algebraic statistical models, we consider three aspects of estimation in the models using the minimum Euclidean (L2E), minimum Pearson chi-squared, and minimum Neyman modified chi-squared distance paradigms as well as the more ubiquitous maximum likelihood

approach (MLE). First, we consider the theoretical properties of the estimators and demonstrate that under general conditions the estimators exist and are asymptotically normal. For small samples, we present the results of large scale simulations to address the estimators' bias and mean squared error (in the Euclidean and Frobenius norms, respectively). Second, we identify the computation of such estimators as an optimization problem and, for the case of the L2E, propose two different methods by which the problem can be solved, one algebraic and one numerical. Finally, we present an R implementation via two novel packages, `mpoly` for symbolic computing with multivariate polynomials and `catcim` for fitting categorical conditional independence models. It is found that in general minimum distance estimators in categorical conditional independence models behave as they do in the more traditional parametric setting and can be computed in many practical situations with the implementation provided.

Dedicated to the Glory of God

– the Lord Jesus Christ –

on account of whose death and resurrection you and I are offered life,

and

my parents

– Warren and Gail Kahle –

whose tireless efforts and innumerable sacrifices for their family

are a testament to the God-given human capacity to love.

Acknowledgements

There are several people who have made the current work possible, and it is my pleasure to try to honor them here.

There are many graduate students and post-docs who have befriended, challenged and instructed me over the years. In particular I would like to thank Alejandro Cruz-Marcelo, Jonathan Lane, Stephanie Hicks, Eric Chi, Joseph Egbulefu, Garrett Grolemund, Blair Christian, Josue Salazar and Birnur Guven for countless discussions over the past five years.

Over the course of my life I have been extraordinarily blessed with a number of outstanding teachers and professors who I would like to acknowledge – Tony Sirignano, Chuck Garwood, Orson Cook, Douglas Sharp, Judy Edwards, Kurt Oehler and Carole Platt of St. John's School in Houston, Texas; Della Fenster, Lester Caudill, Jim Davis and Bill Ross of the University of Richmond; and David Scott, Jim Thompson, Dennis Cox, Kathy Ensor, and Hadley Wickham of Rice University.

I owe a debt of gratitude I can never repay to two former teachers omitted from the above list. First to Wendall Zartman, who refused to believe that I could not do better academically and (more importantly) truly cared for me personally when I was a young man and still does to this day. Second to Doug Elliott, who cared to see a curious spark not unlike his own in a nervous red headed boy and worked to provide fuel and a controlled environment. These two men inspired me then and, to an even greater degree, now; they can be found at St. John's.

For the past year I have found a research home and financial support from Devika Subramanian, Bob Stein and Leonardo Dueñas-Osorio, all of Rice University. Dedicated to understanding the physical and human dimensions of hurricane related disasters, the group has been one of the most rewarding experiences I have undertaken while at Rice, and I thank them dearly for the opportunities they have provided me.

I would also like to thank my thesis committee – Brendan Hassett, Yin Zhang, Hadley Wickham, and Javier Rojo – for many interesting and useful discussions from

which the following thesis has benefited. Javier Rojo, my advisor and friend of many years, I thank especially. He has truly made a profound impact on my life and the lives of many others through his hard work.

Several close personal friends have been by my side throughout the process as well. Jacob Buck, Przemek Polaski, Warren Bellows, Nick Dhesi, Josh Buck, Will Bryant, Elisa Cortez and Javier Valdez have each in their own way added real joy to my complex life which I will be forever grateful for.

I am happy to here publicly express my great appreciation of and admiration for Lieutenant Aaron Hawley, Majors Robert and Shannon Winters, and Lieutenants Joe and Maxie DeBlanc of The Salvation Army who have faithfully purposed to care for and support me spiritually for the past two and a half years. It is the likes of these of whom our Lord said, "Who then is the faithful and wise servant, whom his master has set over his household, to give them their food at the proper time? Blessed is that servant whom his master will find so doing when he comes. Truly, I say to you, he will set him over all his possessions." (Mt. 24:45-47, ESV)

I am dumbfounded in gratitude and love for my girlfriend Becky Guest. She has been my confidant and companion, mentor and friend throughout this process. She has laughed with me, cried with me, danced with me and prayed with me, and there is nobody else I would have wanted to be with me through these last years. She is quite simply the magic which makes defying gravity possible.

Language cannot express the gratitude and love that I have for my family. This work is dedicated in part to my parents Warren and Gail Kahle whose love for their children is unsurpassed in the whole world. My siblings Kathy, Richard, and Michael Kahle have challenged, encouraged, loved and sustained me throughout this entire ordeal. My aunt Sharon Sherrill and cousins Stacy and Scott Bennett and John and Kelly Sherrill and their little ones are more to me than their titles suggest; and I thank them for that. I love you all more than you will ever know.

David Kahle, Rice University, August 2011

Contents

Illustrations

Tables

List of Algorithms

Chapter 1

Introduction

Our primary concern in this work is estimation in special models for categorical data using minimum distance as the estimation paradigm. More specifically, we endeavor to illuminate the use of the Euclidean, chi-squared, and Neyman modified chi-squared distances for estimation in conditional independence models used to understand associations in finite state discrete multivariate experiments, with an emphasis on the Euclidean distance. As a benchmark for comparison, we also consider the more conventional maximum likelihood estimator. The models and methods described herein are entirely distribution free. Thus, they carry with them a clean and unimposing feel from the very beginning of the statistical analysis all the way through to the action taken.

The methods considered are applicable to virtually every kind of data set. From most general to most structured, the data hierarchy is generally considered to be

$$\text{Categorical (Nominal)} \quad < \quad \text{Ordinal} \quad < \quad \text{Interval} \quad < \quad \text{Ratio},$$

where $<$ is loosely thought of as "is less structured than." With ratio data, for instance distance measurements, observations are well ordered and differences and ratios have meaning. If these data are binned, so that the "new" data are counts of observations in certain intervals, we arrive at an example of ordinal data; that is, data which have a natural ordering but differences and ratios are either undefined or carry no intrinsic meaning. If we then disregard the ordering of the bins, we obtain categorical data.

Since any data set can be transformed into categorical data in a similar fashion, the methods discussed in this work can be seen to apply to any kind of data. That being said, they are not appropriate for every application. In the "forgetting" process we lose a considerable amount of information regarding the underlying experiment, and it is very often the case that this lost information is exactly the information we wish to understand better. Thus, the categorical nature of these methods makes them simultaneously very general and very restrictive.

The conditional independence models discussed in this work constitute a fairly large class of distributions with nice properties. They have very rich interpretations; indeed, their meaning goes to the very heart of understanding any association whatsoever. Until fairly recently, however, they have been somewhat overlooked in their natural state[1] by the statistical community. Instead, statisticians have sought to solve the same problems using a collection of very similar models called log-linear models which were born over the first half of the 20th century out of an analogy with R. A. Fisher's analysis of variance (ANOVA). The now prominent class of graphical models is a curious product of the conditional independence and log-linear models. Graphical models use the graphs of graph theory as a macrolanguage to concisely communicate associations; however, depending on their context, their intended interpretation can be either a series of conditional independence statements concerning the variables in the graph *or* a factorization of the joint density of the variables in the graph. The first interpretation is exactly what a conditional independence model is; the second is a slight generalization of a log-linear model. A formal link between these two interpretations was not made until the early 1970's when in an unpublished paper the now celebrated Hammersley-Clifford theorem was proved granting sufficient conditions for

[1] That is, their implicit mathematical definition.

the equality of log-linear models and conditional independence models conveyed by undirected graphs, models known as undirected graphical models or Markov random fields. Generally speaking, the exact probabilistic model communicated by a graph currently depends on the field of application. Researchers in the field of data mining, artificial intelligence, an applied statistics usually mean the factorization of the joint density while those in mathematics, information science, and mathematical statistics typically mean conditional independence models (using the proper definition and not a derived equivalent). In this work we adopt the tradition of the latter. Our focus is thus conditional independence models of which graphical models present the most widely used and successful example.

Currently almost all categorical theory is based on likelihood theory, with the one prominent exception being Pearson's famous chi-squared statistic used in goodness-of-fit testing. However, even Pearson's chi-squared goodness-of-fit test is only a partial exception because, as we see in Section 2.3, the statistic is founded only half-heartedly on the idea of minimum distance. The dual estimation paradigm of minimum chi-squared, which has a small but tangible canon, has been almost entirely forgotten from the average statistician's imagination, to whom the battle cry when attacking a problem seems to always be formulate a log-likelihood and then maximize it. In the wonderful dialogue of responses between many statistical giants following Berkson [1980], Efron quipped, "Maximum likelihood is the original 'jackknife', a dependable tool for almost any estimation purpose." And he was right. Maximum likelihood is a useful paradigm which typically produces good estimators. That being said, minimum distance is also a good paradigm which carries with it a number of useful advantages.

To summarize, this work considers a classical estimation problem for contingency

table-type data from various angles. For this kind of data, an estimation problem consists of two components – a model and a estimation paradigm. In practice, for this kind of data the overwhelmingly prevalent strategy is to use log-linear models and maximum likelihood. In this thesis we consider conditional independence models using the estimation principle of minimum distance, with emphasis on the Euclidean distance, as the estimation paradigm.

1.1 Contributions

To begin, it should be noted that many of the ideas in this thesis are not novel. From the perspective of a purely theoretical statistician, minimum distance estimation has existed in a formal setting since the 1950's; it has even been discussed in regards to categorical problems. The conditional independence models in consideration are not novel either. From the applied perspective, the perhaps unfamiliar Gröbner bases machinery used in fitting is neither new nor is its application to statistical problems new. The numerical methods used in this work are not a discovery of the current author either.

However, there are nevertheless many aspects of this work which are novel contributions, both to the theory and application of minimum distance estimation in conditional independence models. These contributions come in three varieties – theory, fitting, and implementation.[2] We summarize these contributions separately.

---

[2]The difference between the terms "fitting" and "implementation" is nuanced. As an analogy to highlight the difference, the standard regression problem yields the L_2 fitting solution $\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y}$. However, when the applied statistician wants to actually calculate $\widehat{\boldsymbol{\beta}}$ with actual values for $\boldsymbol{A}$ and $\boldsymbol{y}$, he requires (in addition to tricks from numerical linear algebra for efficiency and stability) a computing environment for automated arithmetic and instructions (code) to actually

Theoretical contributions

By theoretical developments we mean the justification of the methods suggested (minimum distance, primarily least squares) in the settings described (conditional independence models) by classical means. Classical means of justifying statistical procedures come in the form of statistical properties. For example, we may select one estimator over another because it exhibits the property called unbiasedness while the other does not, and we happen to prefer unbiased estimators. In practice, we are only able to calculate the distribution of a finite sample statistic in a *very* limited number of cases, many (if not most) of which are so ideal as to be fantastic and altogether incredible. Instead, we approximate the distribution of the statistic with the distribution it would achieve had we an infinite number of observations; this is the idea of asymptotics. Since this is typically easier to obtain than that of the finite sample statistic, we can compare statistics based on their asymptotic distributions rather than their exact, finite sample distributions.

Philosophically speaking, the quality of the approximation of the asymptotic distribution to the finite sample distribution is paramount. It is "the most important, the most difficult, and consequently the least answered [question concerning such approximations]," write Bishop et al. [2007] in their authoritative text on discrete multivariate analysis. However, practically speaking, such considerations are almost never made. "Analytic answers to [this question] are usually very difficult, and it is more common to see reported the result of a simulation or a few isolated numerical calculations rather than an exhaustive answer." Indeed, even when attempted, the asymptotic justifications are often not considered rigorously. In practice the asymp-

---

carry out the procedure. This latter part we label implementation.

totic results themselves are "rarely answered carefully, and [are] typically tossed aside by a remark of the form '...assuming that higher-order terms may be ignored...'." (Bishop et al. [2007], pp. 457-458)

In a categorical context (i.e., one where experiments yield a finite number $r \in \mathbb{N}$ of outcomes), the fundamental problem faced when dealing with conditional independence models is that they are not parametric in the typical sense of the term; rather, they are implicitly defined models. To be specific, categorical conditional independence models are naturally *implicitly* defined as solution sets of systems of polynomial equations, whereas virtually every statistical model is *parametrically* defined in terms of a "nice" parametric model. A statistical model $\mathcal{M}_{r-1} \subset \mathbb{R}^r$ is said to be represented implicitly if functions $h_1, \ldots, h_k$ are known so that $\mathcal{M}_{r-1} = \{\boldsymbol{\pi} \in \mathbb{R}^r : h_1(\boldsymbol{\pi}) = \cdots = h_k(\boldsymbol{\pi}) = 0\} \cap \Delta_{r-1}$, where Δ_{r-1} is the collection of all probability vectors on r outcomes. It is represented parametrically if functions $\boldsymbol{\pi} : \boldsymbol{\Theta} \subset \mathbb{R}^d \to \Delta_{r-1}$ are known[3] so that $\mathcal{M}_{r-1} = \text{Im}(\boldsymbol{\pi})$, the image of $\boldsymbol{\pi}$. If $h_1, \ldots, h_k$ are polynomials, $\mathcal{M}_{r-1}$ is called an algebraic statistical model. Algebraic statistical models, and therefore conditional independence models, are more properly understood and studied as the algebro-geometric objects known as affine varieties and are thus a primary topic of interest in the nascent field of algebraic statistics. The algebraic nature of the models opens up the entire estimation procedure to algebraic investigation which has statistical value in understanding the models, the asymptotic theory of estimators in the models, and actually calculating those estimators when presented with data.

Since all categorical problems are trivially parameterized by their probabilities, categorical conditional independence models are technically parametric and therefore

[3]This abuse of notation is common in the area of categorical data analysis.

every notion from parametric point estimation theory transfers. However, while properties such as consistency and asymptotic normality have meaning, theorems which are typically used to guarantee them such as Birch's theorem no longer apply since conditional independence models do not have nice parametric forms. Thus, after laying out a formal framework in which to address these problems this work presents theoretical justifications for the various estimators in conditional independence models based on standard asymptotic theory, including discussions concerning existence and uniqueness. Traditionally these problems are considered in the nice parametric $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ framework and the asymptotic distribution considered is that of $\widehat{\boldsymbol{\theta}}$. In this exposition, the primary target of interest is $\boldsymbol{\pi}$ itself, since no parameterization of the previous form is available, and the asymptotic distribution of interest is that of the estimator $\widehat{\boldsymbol{\pi}}$ for various such estimators. Since the connection between the asymptotic theory and the finite sample theory is unknown, we also provide a number of simulations which speak directly to the validity of the methods in finite samples. The method driving these simulations is here labelled "Dr. Pangloss' Method" because it attempts in a very concrete way to generate "all possible worlds," i.e. every possible probability distribution in the model.

Fitting contributions

Put plainly, an estimator is worthless if it cannot be computed. While minimum distance estimation has existed for some time, it never gained wide spread support, a consequence likely due to the difficulty of fitting the models; i.e., computing the estimators when presented with data (Böhning and Holling [1986] reference this problem for the minimum chi-squared statistic). The same is true for minimum distance estimation in contingency table models where, in those exceedingly few cases where

it is mentioned, it is only granted a few pages which introduce the motivating ideas of the technique as opposed to a more general theory which includes properties of the resulting estimators as well as how the models can be fit. On the topic of fitting, the practitioner finds lines such as "The procedures ... will vary in their difficulty according to the nature of the particular problem." (Neyman [1949], p. 254), "Even in the simplest cases (such as independence in a two-way contingency table), however, these equations [used to determine the minimum chi-squared estimator] can be difficult to solve" (Bishop et al. [2007], p. 349), and (speaking about a similar problem) "all that is needed is a general minimization routine to compute the estimate" (Read and Cressie [1988], p. 32). It seems that the only work which actually considers such a routine is the five page article Böhning and Holling [1986], and it only considers two-way tables and the minimum chi-squared paradigm.

A discussion of fitting is even more important in this work than in the works above due to the implicit nature of the models considered. Fitting is in many ways a model-specific endeavor. Good model representations, if they exist, can make fitting almost trivial; bad ones can make fitting effectively impossible. Similar to the estimation paradigm of minimum distance, the primary reason that conditional independence models are not ubiquitous is that they are unwieldy.

After a reformulation of the problem, the current work introduces two methods for resolving the fitting problem – the use of semidefinite programming techniques from numerical optimization and the use of Gröbner bases techniques from algebraic geometry. These methods, established outside of statistics, can be seen to be specially tailored for our current purposes and carry with them a number of advantages and disadvantages which can be considered at application time. Generally speaking, the methods exhibit the advantages and disadvantages typical of the symbolic and exact

vs. numerical and approximate divide – while the algebraic technique is more mathematically elegant and descriptive, the numerical technique is often the only resort in applications.

Implementation contributions

As mentioned above, some of the techniques are nonstandard in statistics and require know-how and careful coding to execute correctly and efficiently. The implementation of the methods alone is a current area of active research, both for the algebraic methods and the numerical methods. Issues of model specification, data structuring, programming language consistency, and computational efficiency are all particulars which must be considered when attempting any of the methods described in this thesis. To further complicate the problem there are many insidious details which must be ironed out in creating such an implementation.

However, while the contents of the work have a fairly steep learning curve, the motivating ideas are very simple. To resolve this asymmetry, for the more applied statistician we present a fully functioning, user-friendly implementation of the methods described in this work in the form of an R package called `catcim`. The implementation has both fitting and utility functions for dealing with contingency table (`array` type) data. Moreover, it is as general as the mathematical framework allows – in principle `catcim` can take in any conditional independence model for any contingency table and fit it using either the maximum likelihood or minimum distance paradigms (minimum chi-squared, minimum Neyman modified chi-squared, and least squares). It is of course limited by the computing resources available, but the limitations are quite reasonable for moderate problems.

In addition to `catcim`, a second novel package called `mpoly` is presented which

drives `catcim`. `mpoly` is an implementation of symbolic computing with polynomials in R which boasts a fairly wide array of features – from polynomial arithmetic to the computation of Gröbner bases. The package not only provides R users with the new capabilities of symbolic computing with polynomials but also lays a foundation for future work in algebraic statistics which, as of yet, R has little to offer.

Additional comments on contributions

Among these other things, this work is intended to be a contribution to algebraic statistics. Algebraic statistics is a very young discipline which can be generally thought of as any investigation into a statistical procedure using polynomial algebra. This work is thus algebraic in two ways. First, the models themselves have an algebraic interpretation as affine or projective varieties. Second, some of the methods used to fit the models are algebraic in nature. While the algebraic statistics community has considered similar models and methods since the mid-1990's, their discussion has been limited entirely to likelihood theory. Thus, to algebraic statisticians this work introduces a new principle of estimation which aligns beautifully with their algebraic and geometric predilections. Moreover, there is no reason to believe that the minimum distance estimators discussed here cannot be investigated more algebraically in much the same way as the maximum likelihood estimator, a labor which has already produced many fruits (for example Diaconis and Sturmfels [1998], Hosten et al. [2005], Geiger et al. [2006], and Drton and Sullivant [2007]).

It has been said that Euclid's primary contribution to mathematics was not proving geometric theorems, for many of these were already known; rather, his foremost contribution was a compilation and synthesis of the mathematical knowledge available in and around Greece in the ancient world. It is this author's belief that statistics

is the philosophy, science, and art of data analysis, including its synthetic interpretation in light of the rest of the human experience, and that therefore the statistician is charged with using any and all means available in order to most clearly illuminate the hidden mechanisms generating the observed phenomena. It is hoped that this thesis aids in that endeavor by providing a synthesis of ideas from many disparate areas of the mathematical sciences in order to solve a fundamentally statistical problem. For this reason, the work boasts a diverse array of sources, from algebraic geometry to asymptotic statistical theory to numerical optimization to statistical software. It is hoped that the reader will seek these out should he be interested in further study or should some of the ideas discussed be unclear.

1.2 Organization

The organization of the current work is as follows.

The current chapter has served to introduce the general gist of the thesis – the data type and estimation question of interest. It has also highlighted the contributions of the work.

Chapter 2 provides three concrete examples which showcase the methods to be presented. The methods in these examples are readily extended into higher dimensions and much more complex experiments and provide motivation for the heavier content of the thesis.

Conditional independence models are discussed at length in Chapter 3, which serves three purposes. First, it provides a more thorough description of the data amenable to the methods discussed in this thesis including notational conventions. Second, it presents categorical conditional independence models in their most general setting. Since conditional independence models have only come into interest in the

algebraic statistics community over the past fifteen years, the content here is non-standard and is only available in a few of other sources (e.g., Drton et al. [2009] or Pachter and Sturmfels [2005]). While largely available in these sources, the current exposition presents a fresh perspective by focusing entirely on the statistical problems at hand beginning with foundational motivating examples. Finally, a more general class of models which includes the conditional independence models called algebraic statistical models is described. Extensive literature references are included throughout.

Chapter 4 considers estimation in the setting of the conditional independence models introduced in Chapter 3. It begins with a brief overview of the estimation enterprise. The discussion then introduces the current means of estimation in conditional independence models, namely maximum likelihood, and presents various results concerning its description. The minimum distance paradigm is presented next and in a complementary manner to that of the likelihood paradigm in the sense that the geometric representation of the estimators is analogous to that of the likelihood estimators. Next existence and uniqueness of the estimators is considered followed by asymptotic results.

In Part II, application of the methods of Part I is considered in detail.

Chapter 5 describes the two novel means of fitting conditional independence models using the likelihood and L2E paradigms by breaking them into the general categories of numerical schemes in Section 5.2 and algebraic methods in Section 5.3. The former is devoted to describing the use of numerical schemes such as semidefinite programming methods to compute the estimators and the latter formalizes the Gröbner basis techniques seen in Chapter 2.

Chapter 6 presents a tour of the `catcim` and `mpoly` R packages, including the

underlying intuition used in building them. Concepts of data storing, sparsity, and model specification and syntax are stressed in an effort to create the most efficient and user-friendly implementations possible. It also provides a real world example and how it can be addressed using `catcim`.

Chapter 7, the conclusion of the work, presents large scale finite sample simulation results for a few common conditional independence models. These results are the product of an extensive and near exhaustive kind of simulation, here called Dr. Pangloss' method, and provide an indication of some of the properties of the maximum likelihood estimator and minimum distance estimators in various settings.

Chapter 2

Motivating Examples

Before moving into the details, a lot of benefit can be had by going over three examples in detail. One of the most elegant features of this work is that it can be seen for the most part as an extension and elaboration of one simple example – the binomial example – to far more complex models. The binomial example is concerned with the estimation of the probability of success, π, in a binomial model of a trinomial experiment. As we will see at length in this work, almost all of the ideas in the example are generalizable. Indeed, understanding the binomial example is the key to understanding the content of this thesis. The intent of the second example – the 2×2 contingency table example – is to focus the discussion more on probabilistic independence. While the same ideas used in the binomial example are employed in this second example, the latter more clearly showcases the natural implicit representation of independence and how the problem of estimation can be seamlessly and straightforwardly integrated into the framework built in the binomial example. The third and last example concerns the standard method of goodness-of-fit testing, that of Pearson, in light of the first two examples. It reveals that Pearson's chi-squared test is actually a strange half-breed of statistical paradigms which, although mechanically sound, is philosophically awkward. Since these examples motivate and form the foundation for the entire work, they are presented here before all else as a primer for what is to follow.

2.1 The binomial example

Suppose we know that a certain experiment X has three possible outcomes or combinations called 0, 1, and 2 so that $X \in \{0, 1, 2\}$, and we want to understand the probabilistic nature of the outcomes. As a notational convenience, we denote the probability of each outcome $P[X = 0] = \pi_0$, $P[X = 1] = \pi_1$, and $P[X = 2] = \pi_2$, where $\pi_k \geq 0$ for $k = 0, 1, 2$ and $\sum_{k=0}^{2} \pi_k = 1$.

We often describe the distribution of such an random variable X by its density $f_X : \mathbb{R} \to [0, 1]$ – here a probability mass function – defined

$$f_X(x) := \begin{cases} \pi_0 & x = 0 \\ \pi_1 & x = 1 \\ \pi_2 & x = 2 \\ 0 & otherwise \end{cases} . \tag{2.1}$$

We identify f_X with the vector $\boldsymbol{\pi} = [\pi_0 \ \pi_1 \ \pi_2]' \in \mathbb{R}^3$ in three dimensions. The set of all such vectors $\boldsymbol{\pi}$ in $\mathbb{R}^3$ is called the probability simplex

$$\Delta_2 = \left\{ \boldsymbol{\pi} \in \mathbb{R}^3 : \mathbf{1}_3' \boldsymbol{\pi} = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}_3 \right\}, \tag{2.2}$$

where the order symbol $\geq$ is interpreted element-wise. We recognize that $\pi_0 + \pi_1 + \pi_2 = \mathbf{1}_3' \boldsymbol{\pi} = 1$ is simply the equation of a plane in $\mathbb{R}^3$; and, incorporating the second condition, we realize that the plane is bounded by the coordinate axes to form the triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ corresponding to the degenerate distributions $X \equiv 0$, $X \equiv 1$, and $X \equiv 2$. The simplex Δ_2 can therefore be visualized in three dimensions; this plot is in Figure 2.1.

Any point $\boldsymbol{\pi}$ on the simplex represents a different probability distribution, a different density f_X. Therefore, a statistical model $\mathcal{M}$ – defined as a collection of

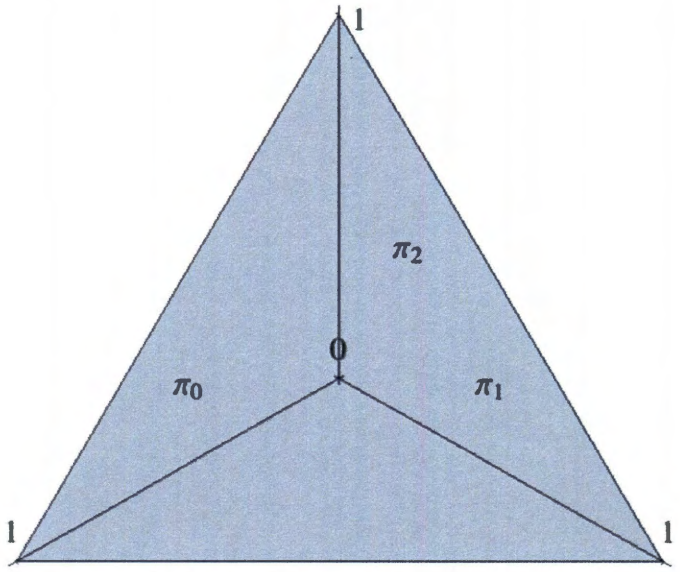

Figure 2.1 : The probability simplex Δ_2

probability distributions – is simply any subset $\mathcal{M} \subseteq \Delta_2$. One such model on three outcomes which could be used is the binomial model with size 2 and unknown probability of success $\pi \in [0, 1]$. For any random variable $X \sim \text{Bin}(2, \pi)$, X exhibits the density

$$f_X(x) := \begin{cases} (1-\pi)^2 & x = 0 \\ 2\pi(1-\pi) & x = 1 \\ \pi^2 & x = 2 \\ 0 & otherwise \end{cases} . \tag{2.3}$$

Thus, the binomial model $\mathcal{M}$ is simply the collection

$$\mathcal{M} = \text{Bin}(2, \pi) = \left\{ \boldsymbol{\pi} = \begin{bmatrix} (1-\pi)^2 \\ 2\pi(1-\pi) \\ \pi^2 \end{bmatrix} \in \Delta_2 : \pi \in [0, 1] \right\} \tag{2.4}$$

which is the image of the map $\boldsymbol{\pi} : [0,1] \to \Delta_2 \subset \mathbb{R}^3$ defined by

$$\pi \overset{\boldsymbol{\pi}}{\mapsto} \begin{bmatrix} (1-\pi)^2 \\ 2\pi(1-\pi) \\ \pi^2 \end{bmatrix}, \tag{2.5}$$

so that $\mathrm{Im}(\boldsymbol{\pi}) = \mathcal{M} \subset \Delta_2$. It is thus a parametric model with the parameterization presented above. In Chapter 4 we will refer to such models as $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ models.

Of course, models for X, being subsets of the simplex Δ_2, can also be considered geometrically. Figure 2.2 contains a plot displaying the binomial model $\mathcal{M}$ as representing only a small sliver of Lebesgue measure 0 in the larger simplex. In this most simple of examples we can actually *see* the statistical model $\mathcal{M}$. Indeed, the binomial model seems almost oppressively restrictive. The canonical experiment to which the binomial model is attributed is that of counting the total number of heads in a sequence of *independent* coin flips. It is this simple assumption of independence that collapses the would-be model of the entire simplex Δ_2 down to the curve $\mathcal{M}$.

Understanding the geometry of this example provides a great deal of insight into the statistical problems we encounter in this work. Thus, suppose we see $N = 20$ observations of N independent and identically distributed copies of X, $X_1 = x_1, X_2 = x_2, \ldots, X_{20} = x_{20}$. Specifically, suppose we see the data

$$1,\ 0,\ 1,\ 2,\ 0,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ 2,\ 1,\ 1. \tag{2.6}$$

To get a preliminary understanding of the data, we would then determine the counts of each of the outcomes $T_0 = t_0 = 7$, $T_1 = t_1 = 11$, $T_2 = t_2 = 2$, from which we would

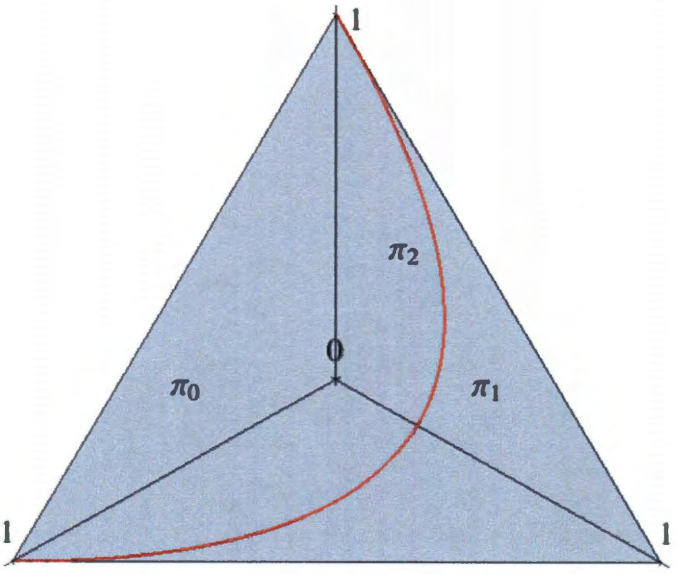

Figure 2.2 : The binomial model $\mathcal{M} = \text{Bin}(2, \pi)$ in probability simplex Δ_2

determine the empirical relative frequencies of each of the outcomes

$$\widehat{\boldsymbol{\pi}}_{EMP} = \begin{bmatrix} \widehat{\pi}_0^{EMP} \\ \widehat{\pi}_1^{EMP} \\ \widehat{\pi}_2^{EMP} \end{bmatrix} = \begin{bmatrix} T_0/N \\ T_1/N \\ T_2/N \end{bmatrix} = \boldsymbol{T}/N = \begin{bmatrix} 7/20 \\ 11/20 \\ 1/10 \end{bmatrix} \in \mathbb{R}^3. \tag{2.7}$$

Being a distribution itself, $\widehat{\boldsymbol{\pi}}_{EMP}$ is a point on the simplex which can be visualized. Not surprisingly, it is not on the curve – the empirical distribution is not in the binomial family $\mathcal{M} = \text{Bin}(2, \pi)$. $\widehat{\boldsymbol{\pi}}_{EMP}$ is included as a black point on the simplex in Figure 2.3.

That being said, the assumption of a model demands that the distribution estimate be in the model. So how do we estimate the distribution of X under the assumption of the model $\mathcal{M}$? Standard calculations demonstrate that the maximum likelihood estimator (MLE) for the binomial parameter π is $\widehat{\pi}_{MLE} = \overline{X}/n = \frac{3}{8}$, where $n = 2$ is

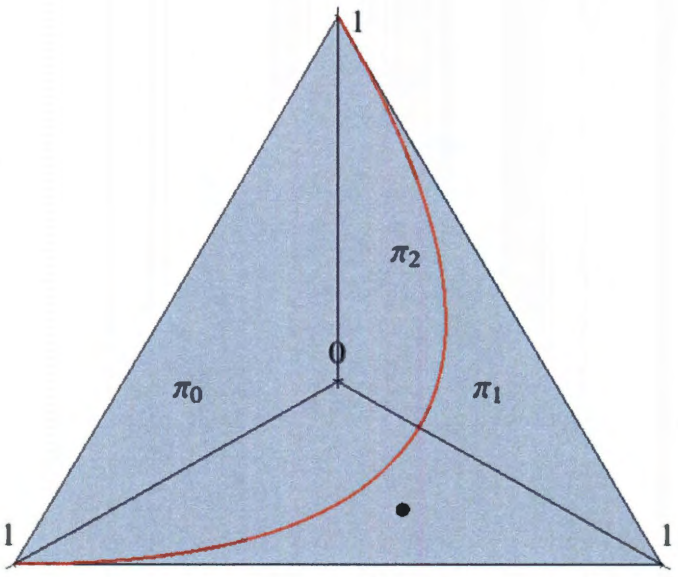

Figure 2.3 : The empirical relative frequencies $\widehat{\boldsymbol{\pi}}_{EMP}$ as a black point in probability simplex Δ_2

the binomial parameter. Excluding degenerate cases, the sum statistic $S = \sum_{k=1}^{N} X_k$ is complete and sufficient for π, and since $\widehat{\pi}_{MLE}$ (seen as a function of S) is unbiased for π, the estimator $\widehat{\pi}_{MLE}$ is also the uniformly minimum variance unbiased estimator (UMVUE) of π as a consequence of the Lehmann-Scheffé lemma. From (2.5), the estimated distribution is therefore

$$\widehat{\boldsymbol{\pi}}_{MLE} = \boldsymbol{\pi}(\widehat{\pi}_{MLE}) = \boldsymbol{\pi}(3/8) = \begin{bmatrix} (1-\frac{3}{8})^2 \\ 2\frac{3}{8}(1-\frac{3}{8}) \\ (\frac{3}{8})^2 \end{bmatrix} = \begin{bmatrix} 25/64 \\ 15/32 \\ 9/64 \end{bmatrix} \in \mathcal{M} \subset \Delta_2. \qquad (2.8)$$

Of course, $\widehat{\boldsymbol{\pi}}_{MLE} \in \mathcal{M}$ is a binomial population by design. It therefore can be seen as a point on the curve $\mathcal{M}$ in the simplex; this is the green point on the red curve $\mathcal{M}$ in Figure 2.4.

In light of its optimality, one might assume that the point $\widehat{\boldsymbol{\pi}}_{MLE}$ is the unique point

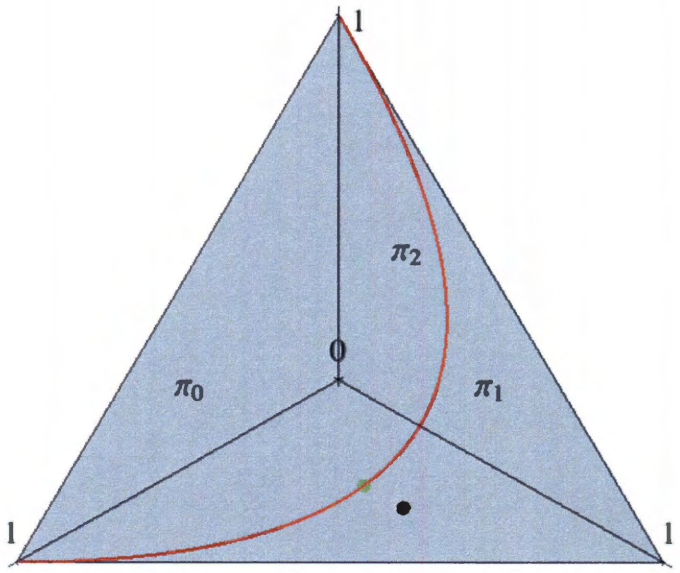

Figure 2.4 : The maximum likelihood estimate $\widehat{\boldsymbol{\pi}}_{MLE}$ as a green point on $\mathcal{M}$ in probability simplex Δ_2

on the curve $\mathcal{M}$ which is closest to the empirical relative frequencies $\widehat{\boldsymbol{\pi}}_{EMP}$; however, a quick calculation demonstrates that this is in fact not the case. By definition, the closest point in the Euclidean metric on the curve to the empirical relative frequencies is given by the parameter[1]

$$\widehat{\pi}_{L2E} = \arg\min_{\pi\in[0,1]} ||\boldsymbol{\pi}(\pi) - \widehat{\boldsymbol{\pi}}_{EMP}||_2 \tag{2.9}$$

To obtain this point, we simply take the derivative and set it equal to zero. The resulting polynomial has one real and two complex roots for π. The real root, approximately 0.397, corresponds to $\widehat{\pi}_{L2E}$, the estimator of π which yields the smallest Euclidean distance to the empirical relative frequencies $\widehat{\boldsymbol{\pi}}_{EMP}$ in the model. The

[1]Note that arg min denotes the argument value which minimizes the objective function as opposed to the value of the minimum itself.

distribution estimator is then

$$\widehat{\boldsymbol{\pi}}_{L2E} = \boldsymbol{\pi}(\widehat{\pi}_{L2E}) \approx \begin{bmatrix} 0.363 \\ 0.479 \\ 0.158 \end{bmatrix}. \tag{2.10}$$

A visual representation of $\widehat{\boldsymbol{\pi}}_{L2E}$ in the simplex can be seen in Figure 2.5.

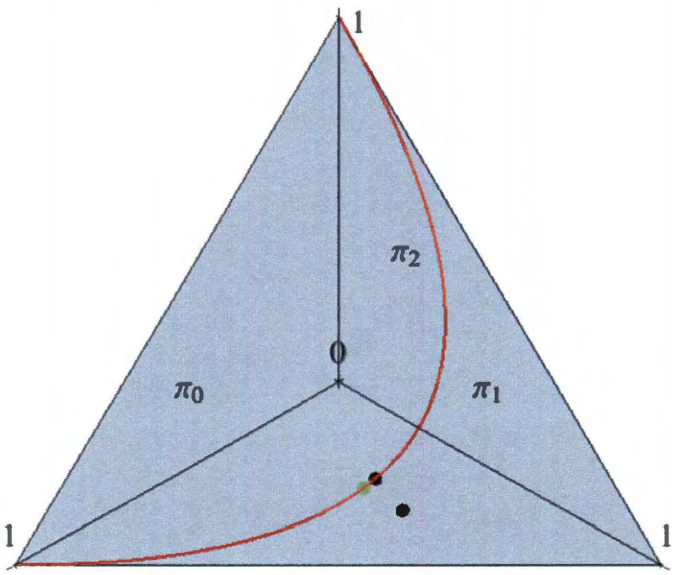

Figure 2.5 : The minimum distance estimator $\widehat{\boldsymbol{\pi}}_{L2E}$ as a black point on $\mathcal{M}$ in probability simplex Δ_2

The data presented in (2.6) was in fact a sample of size 20 from the distribution Bin$(2, 1/2)$. This is presented in Figure 2.6. Notice that the black point on the curve, $\widehat{\boldsymbol{\pi}}_{L2E}$, is closer than the green point, $\widehat{\boldsymbol{\pi}}_{MLE}$, to the true underlying distribution – both on the simplex Δ_2 and in $\mathcal{M}$ (in the sense that $\widehat{\pi}_{L2E}$ is closer to $1/2$ than the $\widehat{\pi}_{MLE}$ is). It is therefore reasonable to say that it provides a better estimate for this sample. However, it is by no means obvious that $\widehat{\boldsymbol{\pi}}_{L2E}$ is better than $\widehat{\boldsymbol{\pi}}_{MLE}$ for other possible samples. Estimation theory classically defines the "best" estimator as

one which minimizes a weighted average of losses over all possible outcomes for some portion of the model.[2] The purpose of this work is to investigate this relationship and extend these ideas to higher dimensions and richer models.

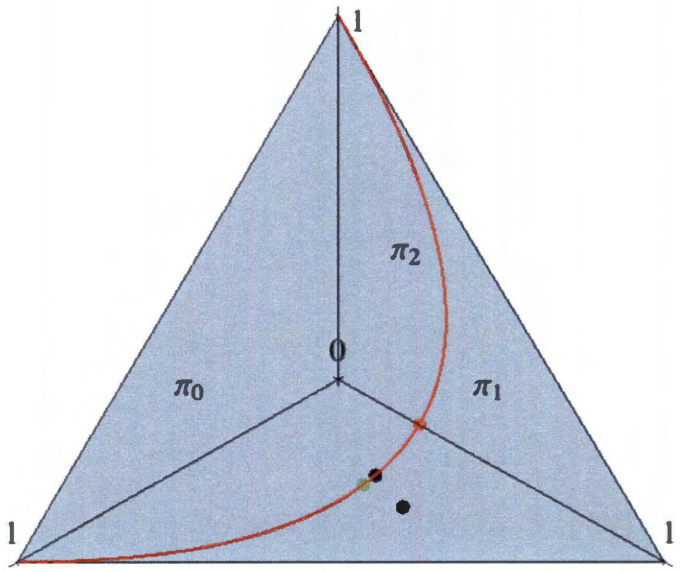

Figure 2.6 : The estimators $\widehat{\boldsymbol{\pi}}_{MLE}$ (green) and $\widehat{\boldsymbol{\pi}}_{L2E}$ (black) along with the true distribution Bin$(2, 1/2)$ (red) on $\mathcal{M}$ in probability simplex Δ_2

One intuitive note regarding the comparison between the MLE and L2E (minimum Euclidean distance estimator) is in order. We have a good feel for what the L2E is doing in terms of the simplex – it is the distribution in the model closest to the empirical relative frequencies in the Euclidean norm. On the other hand, the MLE has its own criterion – it maximizes the likelihood function. Alternatively, the MLE minimizes a different distance, the Kullback-Leibler (KL) divergence, seen in Chapter 4. Unlike the Euclidean norm, balls of constant distance in the KL divergence change shape depending on their position on the simplex, seen in Figure 2.7. One would

[2]This is the concept of risk and associated concepts minimax, Bayes, etc.

think then that the L2E would be more robust, that is to say, relatively insensitive to untrue model assumptions. It turns out that minimum distance estimators such as the L2E are known to have such properties in other settings (Parr and Schucany [1980], Parr [1981], Donoho and Liu [1988]). While not taken up in this work, this robustness is a promising avenue of future research for the L2E which is especially nice for conditional independence models as they make such strong experimental assumptions concerning association.

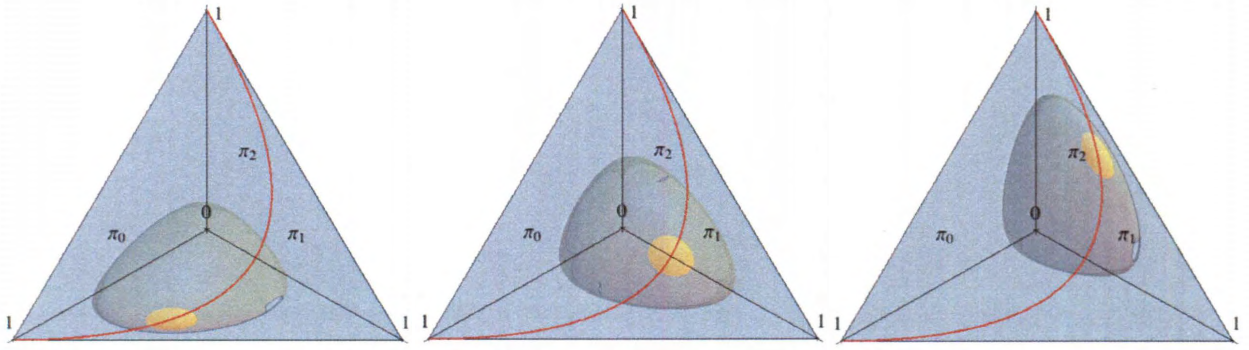

Figure 2.7 : Balls in the Kullback-Leibler divergence of "radius" $\frac{1}{10}$ centered at binomial distributions with $\pi = .25$, .5, and .75. The bright yellow regions where the balls intersect the simplex represent distributions less than $\frac{1}{100}$ from the respective distributions in the divergence.

Algebraic investigation of the binomial model

Much more can be said about the binomial model and minimum L_2 method of estimation. Statistical models, subsets of Δ_2, can be described in more than one way. If we ignore the degenerate cases, we recognize the well known fact that $\mathcal{M}$ is an exponential family (thus the complete and sufficient results of the counts mentioned before). The binomial model presented in (2.4) represents the most common way

statistical models are presented – as explicitly defined parametric models. That is to say, we can easily generate any distribution in the model as points on the simplex by simply applying a function to the parameter. In this case, the function is $\boldsymbol{\pi}$ in (2.5). However, models can also be defined implicitly. For example, it can be shown that the binomial model in (2.4) can be equivalently defined

$$\mathcal{M} = \left\{ \boldsymbol{\pi} = \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \end{bmatrix} \in \mathbb{R}^3 : \pi_1^2 - 4\pi_2 + 4\pi_1\pi_2 + 4\pi_2^2 = 0 \right\} \cap \Delta_2, \tag{2.11}$$

which makes no reference to parameters.[3] The special polynomial in (2.11) is

$$h(\pi_0, \pi_1, \pi_2) = \pi_1^2 - 4\pi_2 + 4\pi_1\pi_2 + 4\pi_2^2. \tag{2.12}$$

The zero set of the polynomial, denoted $V(h)$, is an example of an affine variety –

$$V(h) := \left\{ \boldsymbol{\pi} \in \mathbb{R}^3 : h(\boldsymbol{\pi}) = 0 \right\}. \tag{2.13}$$

Using this notation, we can rewrite (2.11) as

$$\mathcal{M} = V(h) \cap \Delta_2. \tag{2.14}$$

Notice three things. First, points on $V(h)$ need not be on the simplex Δ_2, so we intersect the variety with the simplex to make sure that the resulting elements are all valid probability distributions and therefore collectively a valid statistical model. Second, $V(h)$ is itself a geometric structure which can be visualized in three dimensions, and third, h is not unique for this purpose – there are many such polynomials h for which $\mathcal{M}$ can be written as (2.14). The notion of a "good" representation can be

[3] I.e., ones which could be used explicitly by applying a function as previously.

found in the concept of a Gröbner basis, and finding such a representation amounts to calculating such a basis which we demonstrate shortly.

Adding $V(h)$ to our plot to produce Figure 2.8 illustrates the concept of the variety intersecting the simplex to create the binomial model. These kinds of models which are defined as the intersection of a variety and the simplex are very special models since the polynomial nature of the model lends itself to algebraic investigation. As we will see in this work, tools from algebra, particularly computational algebraic geometry, can be used on these models to create elegant solutions to quite complex problems.

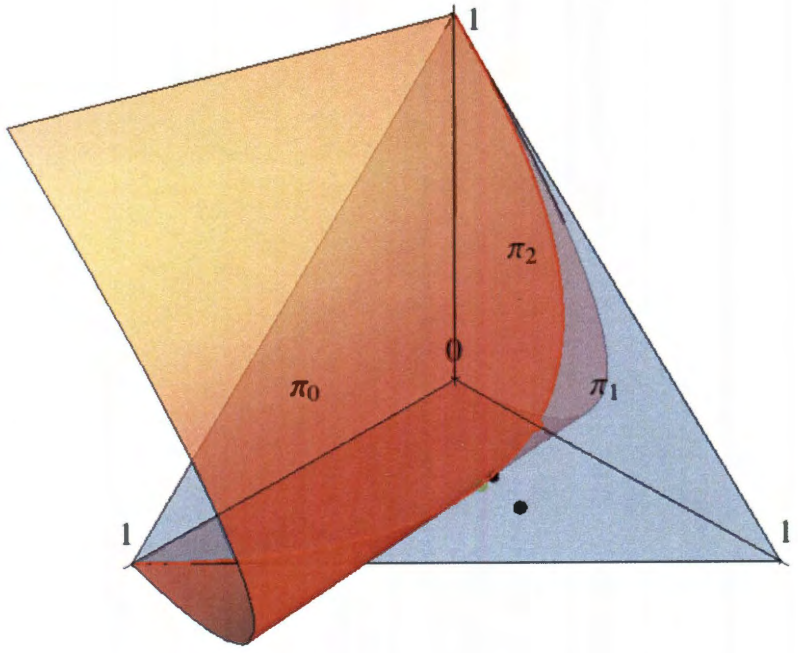

Figure 2.8 : The affine variety $V(h)$ from (2.12) and (2.13) along with the simplex Δ_2

A good sample of the algebraic flavor can be seen in this example. The field of algebraic geometry is concerned with the algebraic description and investigation of

geometric structures. While the varieties we discussed before are geometric structures (we can plot them, for example), they have algebraic analogues which allow them to be amenable to algebraic investigation.

The algebraic analogue of an affine variety is a polynomial ideal. A polynomial ideal is a collection of polynomials which are polynomial combinations of polynomials. Specifically, if $h_1, \ldots, h_k$ are polynomials in the r indeterminates $x_1, \ldots, x_r$, we define the ideal generated by $h_1, \ldots, h_k$ to be the collection of polynomials

$$\langle h_1, \ldots, h_k \rangle := \left\{ \sum_{i=1}^{k} f_i(\boldsymbol{x}) h_i(\boldsymbol{x}) : f_i(\boldsymbol{x}) \in \mathbb{R}[\boldsymbol{x}] \right\}, \tag{2.15}$$

where $\mathbb{R}[\boldsymbol{x}]$ denotes the ring of polynomials in $x_1, \ldots, x_r$.

There are two fundamental results which are needed to understand what follows. The first is that the variety of an ideal – the set $\boldsymbol{x} \in \mathbb{R}^r$ where every polynomial combination of the generating polynomials is zero – is equivalent to the variety of the polynomials which generate it, so that $V(\langle h_1, \ldots, h_k \rangle) =: V(h_1, \ldots, h_k) = \{\boldsymbol{x} \in \mathbb{R}^r : h_1(\boldsymbol{x}) = \cdots = h_k(\boldsymbol{x}) = 0\}$. The second key result is that the variety of an ideal is invariant with respect to the polynomials which generate it. In other words, if we could find polynomials $g_1, \ldots, g_m \in \mathbb{R}[\boldsymbol{x}]$ such that $\langle h_1, \ldots, h_k \rangle = \langle g_1, \ldots, g_m \rangle$, then their varieties would be equal, $V(\langle h_1, \ldots, h_k \rangle) = V(\langle g_1, \ldots, g_m \rangle)$.

One application of this theory is the systematic study of the intersection of surfaces. Specifically, suppose we are trying to determine the intersection of the surfaces $V(h_1), \ldots, V(h_k)$. It is evident that $V(h_1, \ldots, h_k) = \bigcap_{i=1}^{k} V(h_i)$ and therefore the above theorems tell us that if we can obtain another set of polynomials $g_1, \ldots, g_m \in \mathbb{R}[\boldsymbol{x}]$ such that $\langle h_1, \ldots, h_k \rangle = \langle g_1, \ldots, g_m \rangle$, then finding the intersection of $V(h_1), \ldots, V(h_k)$ is equivalent to finding the intersection of $V(g_1), \ldots, V(g_m)$. While this may seem like simply trading one difficult problem for another, we will soon see that for a carefully selected set of g_i's called a Gröbner basis, the tradeoff

vastly improves our ability to compute the intersection of the surfaces defined by the h_i's.

How can we exploit these ideas from algebraic geometry in a statistical context? From Figure 2.8 we know that the binomial model can be seen as the intersection of a surface and the probability simplex. It turns out that every polynomial parametric categorical statistical model[4] can be shown to be the intersection of an affine variety and the probability simplex; however, the converse is not true – not every intersection can be parameterized.

Figure 2.9 displays the culmination of our growing sequence of illustrations related to the binomial model. In addition to what has been seen, Figure 2.9 includes a carefully selected sphere centered around the empirical relative frequencies. The significance of the sphere is that of equidistance in the Euclidean norm, a different L_p distance would result in a different shape. The intersection of the probability simplex with the implicit binomial model with this special sphere is precisely one point – the minimum distance estimator $\widehat{\pi}_{L2E}$.

Using the tools of computational algebraic geometry we can obtain $\widehat{\pi}_{L2E}$ exactly as the point where these surfaces intersect. To do this, we identify the optimization problem

$$\widehat{\pi}_{L2E} = \arg\min_{\pi \in \mathcal{M}} ||\pi - \widehat{\pi}_{EMP}||_2 = \arg\min_{\pi \in V(h) \cap \Delta_2} ||\pi - \widehat{\pi}_{EMP}||_2^2 \qquad (2.16)$$

and solve the problem by introducing Lagrange multipliers, differentiating to obtain a system of polynomial equations, transforming the system into an easier system with the technique of Gröbner bases, and solving the system by systematically solving

[4]That is, one defined as before with a parameter $\theta \mapsto \pi(\theta)$ where θ is a parameter in $\mathbb{R}^k$ and π is polynomial. (Before we had $\theta = \pi$.)

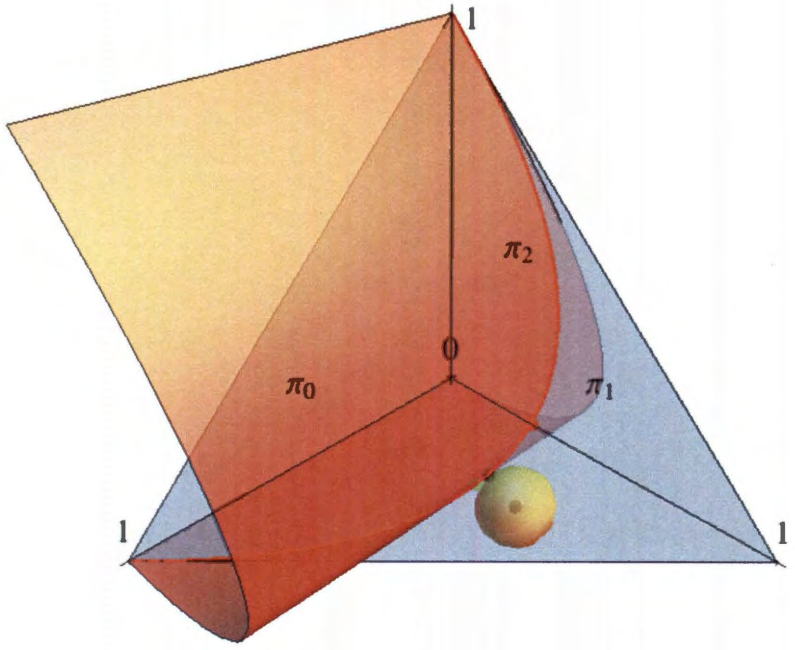

Figure 2.9 : $\widehat{\boldsymbol{\pi}}_{L2E}$ as the intersection of $V(h)$, Δ_2, and a properly chosen sphere centered at the empirical relative frequencies

univariate root finding problems and back substituting. Specifically, we define $\boldsymbol{h}^{\Delta} \in \mathbb{R}[\boldsymbol{\pi}]^{k+1}$ to be the vector of the $k+1$ generators (including the simplex generator[5]), which are regarded as constraints, and the Lagrange function Λ to be

$$\begin{aligned} \Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) &:= ||\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}_n||_2^2 + \boldsymbol{\lambda}' \boldsymbol{h}^{\Delta}(\boldsymbol{\pi}) && (2.17) \\ &= (\pi_0 - \widehat{\pi}_0^{EMP})^2 + (\pi_1 - \widehat{\pi}_1^{EMP})^2 + (\pi_2 - \widehat{\pi}_2^{EMP})^2 && (2.18) \\ &\quad + \lambda_1(\pi_1^2 - 4\pi_2 + 4\pi_1\pi_2 + 4\pi_2^2) \\ &\quad + \lambda_2(\pi_0 + \pi_1 + \pi_2 - 1). \end{aligned}$$

Following the theory of Lagrange multipliers, we then solve the equation $\nabla_{\boldsymbol{\pi},\boldsymbol{\lambda}} \Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) =$

[5]This method finds all possible roots. The estimator is then obtained by eliminating solutions not on the simplex to ensure the nonnegativity condition on the π_k's.

0, a system of nonlinear polynomial equations. In this case, the system is

$$\begin{aligned}
\lambda_2 + 2\left(\pi_0 - \frac{7}{20}\right) &= 0 && (2.19)\\
\lambda_1(2\pi_1 + 4\pi_2) + \lambda_2 + 2\left(\pi_1 - \frac{11}{20}\right) &= 0 && (2.20)\\
\lambda_1(4\pi_1 + 8\pi_2 - 4) + \lambda_2 + 2\left(\pi_2 - \frac{1}{10}\right) &= 0 && (2.21)\\
\pi_1^2 - 4\pi_2 + 4\pi_1\pi_2 + 4\pi_2^2 &= 0 && (2.22)\\
\pi_0 + \pi_1 + \pi_2 - 1 &= 0, && (2.23)
\end{aligned}$$

which looks difficult to solve. However, our geometric intuition compels us to believe there is a unique real solution. Recognizing that by the definition of $V(\cdot)$ the solutions to this system are precisely the points constituting $V(\nabla_{\boldsymbol{\pi},\boldsymbol{\lambda}}\Lambda(\boldsymbol{\pi},\boldsymbol{\lambda}))$, we can use the technique of Gröbner bases to reformulate this system *leaving the associated variety unchanged.* The result is the system

$$\begin{aligned}
14400\pi_2^3 - 5280\pi_2^2 + 4129\pi_2 - 576 &= 0 && (2.24)\\
291\pi_1 + 480\pi_2^2 - 46\pi_2 - 144 &= 0 && (2.25)\\
291\pi_0 - 480\pi_2^2 + 337\pi_2 - 147 &= 0 && (2.26)\\
2910\lambda_2 + 9600\pi_2^2 - 6740\pi_2 + 903 &= 0 && (2.27)\\
3880\lambda_1 - 9600\pi_2^2 + 920\pi_2 - 321 &= 0. && (2.28)
\end{aligned}$$

Now we reap the benefits of choosing the nice set of polynomials we previously labeled $g_1, \ldots, g_m$, a Gröbner basis. To obtain $\widehat{\boldsymbol{\pi}}_{L2E}$ we need but to solve the first three equations (2.24), (2.25), and (2.26), since they are the equations in the three unknowns of interest. This is actually an incredibly easy task since the first equation (2.24) is a cubic equation in π_2 and the second and third equations (2.25) and (2.26) are linear in π_1 and π_0, respectively, once π_2 is known.

To that end, (2.24) yields the one real and two complex roots, so we take $\widehat{\pi}_2^{L2E}$ (the π_2 component of $\boldsymbol{\pi}_{L2E}$) to be the real root. We can then plug this into (2.25), which is linear in π_1, and solve to obtain $\widehat{\pi}_1^{L2E}$. $\widehat{\pi}_0^{L2E}$ can then be obtained either by applying the same back substitution to (2.26) or by simply setting $\widehat{\pi}_0^{L2E} = 1 - \widehat{\pi}_2^{L2E} - \widehat{\pi}_1^{L2E}$. Of course, the latter reduces the problem even further, since in the end all that was required to determine $\widehat{\boldsymbol{\pi}}_{L2E}$ once the problem was reformulated was to solve the cubic equation (2.24) and then the linear equation (2.25). The problem is therefore solved, and the solution is precisely that determined originally in (2.10).

Discussion

The obvious question to ask is – why should we go through all of the algebraic machinery? Was not the first method of taking the derivative and setting it equal to zero much easier? Of course, the answer is yes. So why go through the algebra?

There are a few reasons why using the algebraic geometry is statistically relevant. First, a large class of complex statistical models – the so-called conditional independence models – are naturally implicitly defined and known to be unparameterizable in general (Drton et al. [2009]). These will be discussed in Chapter 3. Second, notice that the algebraic machinery allowed us to do something we were unable to do with conventional methods – compute sub-estimates of the estimator $\widehat{\boldsymbol{\pi}}_{L2E}$ without having to compute the entire estimator. In other words, suppose we are provided with a contingency table upon which we know a particular conditional independence model holds, and we are only interested in the probability of a certain cell. The algebraic machinery allows us to choose to solve only those equations which will lead us to candidates for the estimate of the cell probability of interest. The manner in which we "eliminated" certain variables and then back substituted them into other equa-

tions to arrive at solutions is a major topic in computational algebraic geometry and carries the name of "elimination theory" (Cox et al. [2007]). In the context of the binomial example, the fact that the first equation contained only π_2 was no mistake. This same kind of phenomenon happens in great generality. The catch however is the requirement of computing the Gröbner basis, which is a highly nontrivial task.

2.2 The independence example

The most basic and common example one encounters in the analysis of discrete multivariate data is the 2×2 contingency table. To begin, suppose we observe the gender and handedness of $N = 20$ individuals as seen in Table 2.1. From the perspective of

	right-handed	**left-handed**
male	8	3
female	8	1

Table 2.1 : Gender and handedness of 20 individuals

the experimenter, any given person will fall into exactly one of four possible categories : (male, right), (male, left), (female, right), or (female, left). For ease of notation, we define the random vector

$$\boldsymbol{X}(\boldsymbol{\omega}) = [X_1(\boldsymbol{\omega}), X_2(\boldsymbol{\omega})]' = \begin{cases} [0,0]' & \boldsymbol{\omega} = [\text{male, right}]' \\ [1,0]' & \boldsymbol{\omega} = [\text{male, left}]' \\ [0,1]' & \boldsymbol{\omega} = [\text{female, right}]' \\ [1,1]' & \boldsymbol{\omega} = [\text{female, left}]' \end{cases} . \tag{2.29}$$

Any probability distribution on these four possible outcomes consists of four numbers which we label π_{00}, π_{10}, π_{01}, and π_{11} where, for example, $\pi_{00} = P[\boldsymbol{X} = [0,0]']$ is the probability that an individual is a right-handed male. Taken together, the probabilities can be considered as a vector $\boldsymbol{\pi} = [\pi_{00}\ \pi_{10}\ \pi_{01}\ \pi_{11}]' \in \mathbb{R}^4$ satisfying the properties $\mathbf{1}_4'\boldsymbol{\pi} = 1$ and $\boldsymbol{\pi} \geq \mathbf{0}_4$. Collectively, the set of all possible probability distributions (i.e., vectors) on four outcomes is the probability simplex Δ_3, which is the four dimensional analogue of $\Delta_2 \subset \mathbb{R}^3$ referred to in equation (2.2). Specifically,

$$\Delta_3 = \left\{\boldsymbol{\pi} \in \mathbb{R}^4 : \mathbf{1}_4'\boldsymbol{\pi} = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}_4\right\}. \tag{2.30}$$

Unlike the binomial example in Section 2.1, we cannot directly visualize Δ_3 since it is in 4 dimensions; however, we can do the next best thing. The constraint that $\mathbf{1}_4'\boldsymbol{\pi} = 1$ implies that there are only 3 "free" parameters by which the probability distribution is completely determined. In other words, if we know π_{00}, π_{10}, and π_{01}, we also know π_{11} since it is completely determined by the equation $\mathbf{1}_4'\boldsymbol{\pi} = 1$, $\pi_{11} = 1 - \pi_{00} - \pi_{10} - \pi_{01}$. We can therefore identify each point $\boldsymbol{\pi}$ on the 4-dimensional simplex with its "projection" $\boldsymbol{\pi}^- = [\pi_{00}\ \pi_{10}\ \pi_{01}]' \in \mathbb{R}^3$, where the quotes are used because we are actually changing the ambient space from $\mathbb{R}^4$ to $\mathbb{R}^3$ so that the operation is not a projection in the usual sense.[6] The collection of these points,

$$\Delta_3^- = \left\{\boldsymbol{\pi}^- = [\pi_{00}\ \pi_{10}\ \pi_{01}]' \in \mathbb{R}^3 : \mathbf{1}_4'\boldsymbol{\pi} = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}_4\right\}, \tag{2.31}$$

is readily seen to form the tetrahedron defined by the unit vectors in three dimensions as seen in Figure 2.10.

Every point in the tetrahedron Δ_3^- represents a different probability distribution. Moreover, taken as a whole, the volume represents all possible probability distributions. Since a statistical model is simply a set of probability distributions, $\mathcal{M} \subset \Delta_3$,

[6] Alternatively, we could consider the problem in barycentric coordinates.

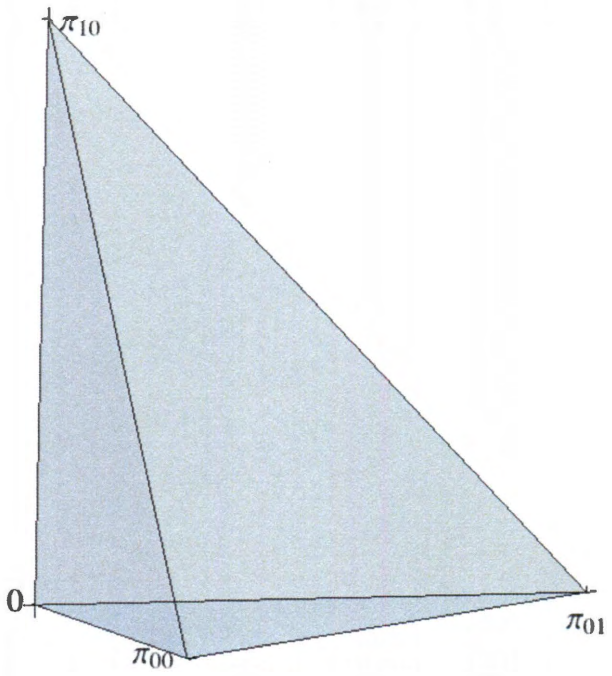

Figure 2.10 : The "projected" probability simplex Δ_3^-

any model can be projected likewise and visualized as a subset of the tetrahedron, $\mathcal{M}^- \subset \Delta_3^-$.

One very important model for 2×2 contingency tables is the so-called independence model. The independence model asserts the statistical independence of the two component random variables of $\boldsymbol{X}$, X_1 and X_2, which correspond to gender and handedness. The interpretation of independence is that knowing an individual's gender provides no information as to that individual's handedness (and vice versa), which is a very old and fundamental question in the analysis of 2×2 contingency tables. Mathematically, the independence model is defined by the statement

$$P[X_1 = x_1, X_2 = x_2] = P[X_1 = x_1] P[X_2 = x_2] \quad \text{for all } (x_1, x_2) \in \{0,1\}^2 . \quad (2.32)$$

Recalling that $P[X_1 = x_1] = \sum_{x_2} P[X_1 = x_1, X_2 = x_2]$, and denoting this probability π_{0+} for $x_1 = 0$ and π_{1+} for $x_1 = 1$ (and similarly for π_{+0} and π_{+1}), (2.32) is equivalently expressed

$$\pi_{x_1x_2} = \pi_{x_1+}\pi_{+x_2} \quad \text{for } (x_1, x_2) \in \{0,1\}^2, \tag{2.33}$$

a system of 4 equations. Thus, the model itself is simply defined

$$\mathcal{M} = \left\{\boldsymbol{\pi} \in \mathbb{R}^4 : \pi_{x_1x_2} = \pi_{x_1+}\pi_{+x_2} \quad \text{for } (x_1, x_2) \in \{0,1\}^2, \mathbf{1}_4'\boldsymbol{\pi} = 1, \boldsymbol{\pi} \geq \mathbf{0}_4\right\}. \tag{2.34}$$

The projected model, $\mathcal{M}^- = \left\{\boldsymbol{\pi}^- = [\pi_{00}\ \pi_{10}\ \pi_{01}]' \in \Delta_3^- : \boldsymbol{\pi} \in \mathcal{M}\right\}$, carves out a surface in the projected probability simplex which is seen from two angles in Figure 2.11.

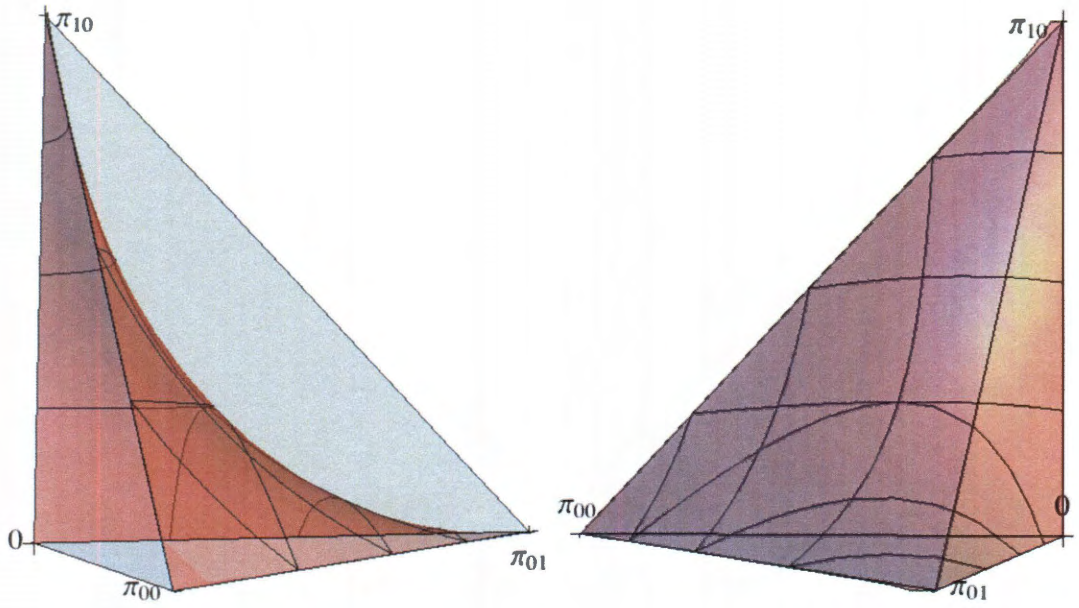

Figure 2.11 : The projected independence model $\mathcal{M}^-$

Now, assuming that the gender and handedness are independent, how do we select the distribution in the model which most closely resembles the data in Table 2.1? As

in the binomial example, the distribution given by the relative frequency of counts ($\widehat{\boldsymbol{\pi}}_{EMP}$) is a valid probability distribution and can therefore be seen as a point in the projected simplex; however, it does not lie on the independence surface. This can be seen in Figure 2.12.

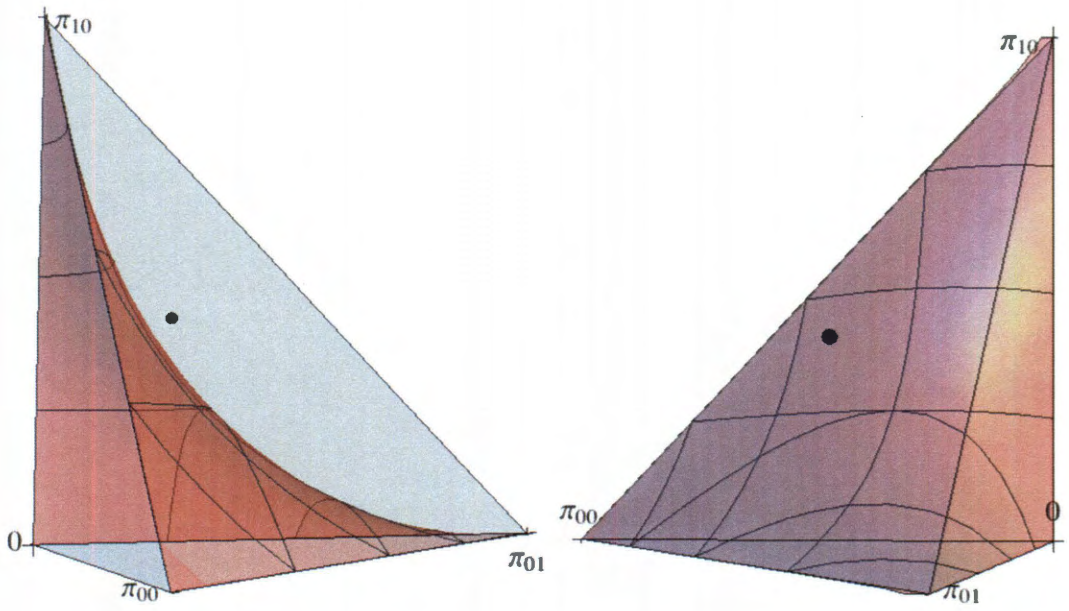

Figure 2.12 : The projected independence model $\mathcal{M}^{-}$ along with the projected empirical relative frequencies $\widehat{\boldsymbol{\pi}}^{-}_{EMP}$ in black

The estimator for $\boldsymbol{\pi}$ used in practice is invariably the maximum likelihood estimator (MLE). To calculate it, we use a simple and straightforward trick. The trick used to compute the MLE in closed form is to realize that independence is equivalent to each marginal random variable X_1 and X_2 being Bernoulli with probability of success π_1 and π_2 and the joint distribution being simply the product of these binary

distributions. That is to say, under independence

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} (1-\pi_1)(1-\pi_2) & \boldsymbol{x} = (0,0) \\ (1-\pi_1)\pi_2 & \boldsymbol{x} = (0,1) \\ \pi_1(1-\pi_2) & \boldsymbol{x} = (1,0) \\ \pi_1\pi_2 & \boldsymbol{x} = (1,1) \end{cases} \tag{2.35}$$

so that $\mathcal{M}$ in (2.34) is equivalently expressed

$$\mathcal{M} = \left\{ \boldsymbol{\pi} = \begin{bmatrix} (1-\pi_1)(1-\pi_2) \\ (1-\pi_1)\pi_2 \\ \pi_1(1-\pi_2) \\ \pi_1\pi_2 \end{bmatrix} \in \mathbb{R}^4 : [\pi_1 \ \pi_2]' \in [0,1]^2 \right\}. \tag{2.36}$$

In general, if $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_N = \boldsymbol{x}_N$ are N independent and identically distributed (iid) samples from the categorical distribution Multinom$_4(1, \boldsymbol{\pi})$ distribution, their sum gives the contingency table counts $T_{00} = t_{00}$, $T_{01} = t_{01}$, $T_{10} = t_{10}$, and $T_{11} = t_{11}$ with multinomial distribution Multinom$_4(N, \boldsymbol{\pi})$. The likelihood function is therefore

$$\begin{aligned} L(\pi_1, \pi_2) &= \binom{N}{t_{00}\ t_{01}\ t_{10}\ t_{11}} \pi_{00}^{t_{00}} \pi_{01}^{t_{01}} \pi_{10}^{t_{10}} \pi_{11}^{t_{11}} && (2.37) \\ &= \binom{N}{t_{00}\ t_{01}\ t_{10}\ t_{11}} \big((1-\pi_1)(1-\pi_2)\big)^{t_{00}} \big((1-\pi_1)\pi_2\big)^{t_{01}} \\ &\qquad \times \ \big(\pi_1(1-\pi_2)\big)^{t_{10}} (\pi_1\pi_2)^{t_{11}} && (2.38) \end{aligned}$$

which by standard techniques is shown to be maximized at

$$\widehat{\pi}_1^{MLE} = \frac{t_{1+}}{N} \qquad \text{and} \qquad \widehat{\pi}_2^{MLE} = \frac{t_{+1}}{N}, \tag{2.39}$$

where $\frac{t_{1+}}{N} = \frac{t_{10}+t_{11}}{N}$ is the marginal empirical relative frequency of observing $X_1 = 1$ and similarly with $\frac{t_{+1}}{N}$.

Now, these "marginal empirical relative frequencies" are simply the empirical relative frequencies of the marginals. Therefore, if we use the notation $\widehat{\boldsymbol{\pi}}_{EMP}$ to denote the empirical relative frequencies of counts as in the last example, we have that for the example at hand

$$\widehat{\boldsymbol{\pi}}^{(1)}_{EMP} = \begin{bmatrix} \widehat{\pi}^{(1)}_0 \\ \widehat{\pi}^{(1)}_1 \end{bmatrix} = \begin{bmatrix} t_{0+}/N \\ t_{1+}/N \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\pi}}^{(2)}_{EMP} = \begin{bmatrix} \widehat{\pi}^{(2)}_0 \\ \widehat{\pi}^{(2)}_1 \end{bmatrix} = \begin{bmatrix} t_{+0}/N \\ t_{+1}/N \end{bmatrix} \tag{2.40}$$

are the marginal empirical relative frequencies[7]

$$\widehat{\boldsymbol{\pi}}^{(1)}_{EMP} = \begin{bmatrix} 11/20 \\ 9/20 \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\pi}}^{(2)}_{EMP} = \begin{bmatrix} 16/20 \\ 4/20 \end{bmatrix}, \tag{2.41}$$

then the maximum likelihood estimator is

$$\widehat{\boldsymbol{\pi}}_{MLE} = \begin{bmatrix} \widehat{\pi}^{MLE}_{00} \\ \widehat{\pi}^{MLE}_{01} \\ \widehat{\pi}^{MLE}_{10} \\ \widehat{\pi}^{MLE}_{11} \end{bmatrix} = \begin{bmatrix} (1-\widehat{\pi}^{MLE}_1)(1-\widehat{\pi}^{MLE}_2) \\ (1-\widehat{\pi}^{MLE}_1)\widehat{\pi}^{MLE}_2 \\ \widehat{\pi}^{MLE}_1(1-\widehat{\pi}^{MLE}_2) \\ \widehat{\pi}^{MLE}_1\widehat{\pi}^{MLE}_2 \end{bmatrix} = \begin{bmatrix} \widehat{\pi}^{(1)}_0\widehat{\pi}^{(2)}_0 \\ \widehat{\pi}^{(1)}_0\widehat{\pi}^{(2)}_1 \\ \widehat{\pi}^{(1)}_1\widehat{\pi}^{(2)}_0 \\ \widehat{\pi}^{(1)}_1\widehat{\pi}^{(2)}_1 \end{bmatrix} = \begin{bmatrix} .44 \\ .11 \\ .36 \\ .09 \end{bmatrix}. \tag{2.42}$$

This is a probability distribution in the independence model by design, and therefore can be visualized (once projected) as not only a point in the tetrahedron but also a point on the independence surface. This point is included in Figure 2.13 in green.

What about the minimum distance estimator? Recall that in the binomial example the minimum distance estimator was first calculated by simply formulating the

[7]Here the superscripts reference the variable label. In this work we label the variable whose categories constitute the rows with 1, and the variable whose categories constitute the columns with 2. So in the current example X_1 refers to gender, and $\widehat{\pi}^{(1)}_{EMP}$ is the marginal empirical relative frequencies of gender.

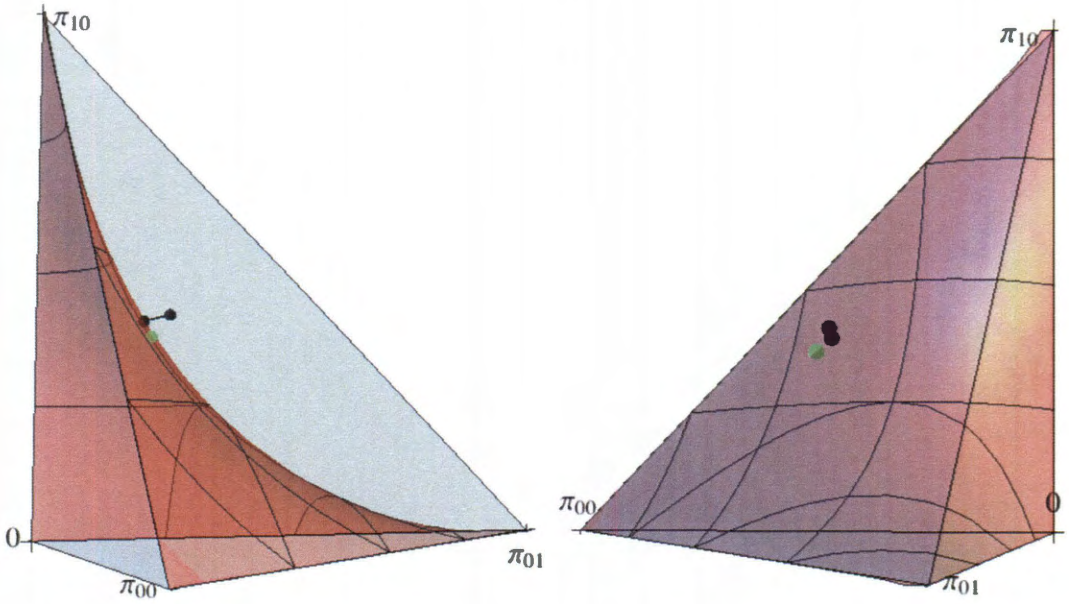

Figure 2.13 : The projected independence model $\mathcal{M}^-$ along with the projected maximum likelihood estimator $\widehat{\boldsymbol{\pi}}^-_{MLE}$ (green) and the projected minimum distance estimator $\widehat{\boldsymbol{\pi}}^-_{L2E}$ (black)

distance from any point in the model to the empirical relative frequencies, differentiating, setting equal to zero, and solving. This method is also available here. The problem we wish to solve is

$$\arg\min_{\boldsymbol{\pi}\in\mathcal{M}} ||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}||_2 \,, \tag{2.43}$$

which we reformulate as

$$\begin{aligned}\arg\min_{(\pi_1,\pi_2)\in[0,1]^2} ||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}||_2^2 \;=\; & \left(\widehat{\pi}_{00}^{EMP} - (1-\pi_1)(1-\pi_2)\right)^2 + \left(\widehat{\pi}_{01}^{EMP} - (1-\pi_1)\pi_2\right)^2 \\ & + \left(\widehat{\pi}_{10}^{EMP} - \pi_1(1-\pi_2)\right)^2 + \left(\widehat{\pi}_{11}^{EMP} - \pi_1\pi_2\right)^2 . \end{aligned} \tag{2.44}$$

Unfortunately, there are no known closed form solutions to this optimization problem

in general. However, the problem can be readily solved numerically to yield

$$\widehat{\pi}_{L2E} = \begin{bmatrix} .411 \\ .104 \\ .387 \\ .098 \end{bmatrix}. \tag{2.45}$$

This is illustrated in Figure 2.13.

The obvious question is simply this – which is preferable? What properties does each exhibit? These are the questions which are posed and considered in this work. In considering these questions we turn to comparison by classical properties of estimation : bias, efficiency, and asymptotic behavior.

Algebraic investigation of the independence model

While we were able to calculate the MLE and L2E for the independence model in the 2×2 contingency table case, our methods – while common to statisticians – are actually not native to the problem at hand; we simply happened to have a parameterization of the model. In general we have no such parameterization. The oddity comes in the model description itself.

Recall that in the binomial example the original description of the binomial model in (2.4) was explicit in nature and only afterwards reformulated to the implicit description of (2.11). The former description is the one with which we are most familiar while the latter description, which was simply presented without explanation of how it was obtained, carried with it a contrived, unnatural feeling.

The case with the independence model is exactly the opposite. The first description of the model, (2.34), although implicit is the natural, primitive description of the model. The second description, provided in (2.36), is the derived, secondary descrip-

tion which were it not for years of training to the contrary would carry with it the same contrived feeling as the implicit description of the binomial model in (2.11).

The raw definition of the independence model is that in (2.34). Independence, and therefore the independence model, is an implicitly defined condition. Similarly, more complex independence structures such as the conditional independence models are naturally implicit. Fortunately, for this independence model we are able to exactly translate the implicit description in (2.34) into the explicit one in (2.36). By contrast, more complicated independence models are not always so amenable to translation, and for most there is simply no such translation. This is one of the reasons why it is important to make an effort to understand implicit descriptions in general, starting with this most simple of independence models. While translations are helpful, for a thorough understanding we must increase our understanding of implicit models; it is this endeavor which brings us to algebra.

The nascent field of algebraic statistics has identified and prioritized issues concerning implicitly defined models and has purposed to solve and understand their associated likelihood theory, beginning with the 2×2 independence model and moving into more complicated independence models.[8] One goal of the present work is to obtain a similar understanding of the minimum distance theory associated with such models which, until now, has been wholly overlooked.

Recall from (2.43) that the minimum distance problem at hand has the form of

[8]Of course, the fruits of the algebraic investigation of categorical problems do not end there. Interesting and important contributions have been made in the field of data disclosure limitation where algebraic statisticians have made advances by, inter alia, considering reconstructing joint distributions from marginal and conditional information (Fienberg and Slavkovic [2005], Slavkovic [2004], Sullivant [2006]).

the optimization problem

$$\widehat{\boldsymbol{\pi}}_{L2E} = \arg\min_{\boldsymbol{\pi}\in\mathcal{M}} ||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}||_2 \,. \tag{2.46}$$

More explicitly, moving all of the conditions to one side the above problem is[9]

$$\begin{aligned} \text{arg-minimize} \quad & ||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}||_2^2 \\ \text{subject to} \quad & \pi_{ij} - \pi_{i+}\pi_{+j} = 0, \qquad i,j = 0,1 \\ & \pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} - 1 = 0. \end{aligned} \tag{2.47}$$

The first four constraint equations in (2.47) can be seen to be analogous to (2.12) for the binomial model in the sense that they really define the model itself – every such $\boldsymbol{\pi}$ satisfying those equations which is also on the simplex is a member of the model. To be as specific as possible, define the four polynomials

$$h_1(\boldsymbol{\pi}) := \pi_{00} - \pi_{0+}\pi_{+0} \tag{2.48}$$

$$h_2(\boldsymbol{\pi}) := \pi_{01} - \pi_{0+}\pi_{+1} \tag{2.49}$$

$$h_3(\boldsymbol{\pi}) := \pi_{10} - \pi_{1+}\pi_{+0} \tag{2.50}$$

$$h_4(\boldsymbol{\pi}) := \pi_{11} - \pi_{1+}\pi_{+1}. \tag{2.51}$$

As before, we define the affine variety of a collection of polynomials to be the intersection of their varieties –

$$V(h_1,h_2,h_3,h_4) := \left\{\boldsymbol{\pi}\in\mathbb{R}^4 : h_1(\boldsymbol{\pi}) = h_2(\boldsymbol{\pi}) = h_3(\boldsymbol{\pi}) = h_4(\boldsymbol{\pi}) = 0\right\}. \tag{2.52}$$

With this collection in mind, we can describe the independence model $\mathcal{M}$ just as in (2.14),

$$\mathcal{M} = V(h_1,h_2,h_3,h_4) \cap \Delta_3, \tag{2.53}$$

[9]Note the lack of positivity. Using this method positivity is dealt with a posteriori, i.e. after the candidate solutions are obtained, as in the previous example.

so that the model itself is the intersection of a variety and the probability simplex.

To solve the optimization problem, we again introduce Lagrange multipliers, one for each equality constraint, so that the Lagrangian is

$$\begin{aligned}
\Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) &:= \left|\left|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}^{EMP}\right|\right|_2^2 + \boldsymbol{\lambda}' \boldsymbol{h}^{\Delta}(\boldsymbol{\pi}) \qquad (2.54)\\
&= (\pi_{00} - \widehat{\pi}_{00}^{EMP})^2 + (\pi_{01} - \widehat{\pi}_{01}^{EMP})^2 + (\pi_{10} - \widehat{\pi}_{10}^{EMP})^2 + (\pi_{11} - \widehat{\pi}_{00}^{EMP})^2 \\
&\quad + \lambda_1(\pi_{00} - \pi_{0+}\pi_{+0}) + \lambda_2(\pi_{01} - \pi_{0+}\pi_{+1}) \\
&\quad + \lambda_3(\pi_{10} - \pi_{1+}\pi_{+0}) + \lambda_4(\pi_{11} - \pi_{1+}\pi_{+1}) \\
&\quad + \lambda_5(\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} - 1).
\end{aligned}$$

Again, the theory of Lagrange multipliers requires that we solve the system of equations $\nabla_{\boldsymbol{\pi},\boldsymbol{\lambda}} \Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \mathbf{0}$, which is

$$\begin{aligned}
\lambda_1(-2\pi_{00} - \pi_{01} - \pi_{10} + 1) - \lambda_2(\pi_{01} + \pi_{11}) - \lambda_3(\pi_{10} + \pi_{11}) + \lambda_5 + 2\left(\pi_{00} - \frac{2}{5}\right) &= 0 \qquad (2.55)\\
-\lambda_1(\pi_{00} + \pi_{10}) + \lambda_2(-\pi_{00} - 2\pi_{01} - \pi_{11} + 1) - \lambda_4(\pi_{10} + \pi_{11}) + \lambda_5 + 2\left(\pi_{01} - \frac{3}{20}\right) &= 0 \qquad (2.56)\\
-\lambda_1(\pi_{00} + \pi_{01}) + \lambda_3(-\pi_{00} - 2\pi_{10} - \pi_{11} + 1) - \lambda_4(\pi_{01} + \pi_{11}) + \lambda_5 + 2\left(\pi_{10} - \frac{2}{5}\right) &= 0 \qquad (2.57)\\
-\lambda_2(\pi_{00} + \pi_{01}) - \lambda_3(\pi_{00} + \pi_{10}) + \lambda_4(-\pi_{01} - \pi_{10} - 2\pi_{11} + 1) + \lambda_5 + 2\left(\pi_{11} - \frac{1}{20}\right) &= 0 \qquad (2.58)\\
\pi_{00} - (\pi_{00} + \pi_{01})(\pi_{00} + \pi_{10}) &= 0 \qquad (2.59)\\
\pi_{01} - (\pi_{00} + \pi_{01})(\pi_{01} + \pi_{11}) &= 0 \qquad (2.60)\\
\pi_{10} - (\pi_{00} + \pi_{10})(\pi_{10} + \pi_{11}) &= 0 \qquad (2.61)\\
\pi_{11} - (\pi_{01} + \pi_{11})(\pi_{10} + \pi_{11}) &= 0 \qquad (2.62)\\
\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} - 1 &= 0 \qquad (2.63)
\end{aligned}$$

This system is analogous to that which we determined in the binomial example. From here, we could try to numerically solve this system by reformulating it into a sum of squares, a numerical technique discussed in Chapter 5. For now, we can use the technique of Gröbner bases discussed in the binomial example to reformulate the

system as

$$25600\pi_{00}^5 - 33920\pi_{00}^4 + 17136\pi_{00}^3 - 7674\pi_{00}^2 + 3733\pi_{00} - 760 = 0 \quad (2.64)$$

$$8268800\pi_{00}^4 - 6181760\pi_{00}^3 + 2039248\pi_{00}^2 - 1350658\pi_{00} + 99140\pi_{01} + 393585 = 0 \quad (2.65)$$

$$-2329600\pi_{00}^4 + 1790720\pi_{00}^3 - 541376\pi_{00}^2 + 387079\pi_{00} + 24785\pi_{10} - 135088 = 0 \quad (2.66)$$

$$1049600\pi_{00}^4 - 981120\pi_{00}^3 + 126256\pi_{00}^2 - 98518\pi_{00} + 99140\pi_{11} + 47627 = 0 \quad (2.67)$$

(There are, in addition to these, two other much larger equations which involve the λ's which have been omitted as they are irrelevant to the π's.) Thus, we have obtained for the 2×2 contingency table independence model the simplified minimum distance equations, analogous to (2.24)–(2.26) for the binomial example. Since the first equation is a quintic in π_{00}, we cannot in general provide a closed form solution (in terms of radicals) for $\widehat{\boldsymbol{\pi}}_{L2E}$; however, we are able to solve numerically the quintic and even much higher degree univariate polynomials quickly and accurately using numerical root finders (Newton, Jenkins-Traub, etc.). Using a root finder to determine the solution $\widehat{\pi}_{00}^{L2E}$, we then simply substitute its value to solve the other three equations, each linear in its respective variable.[10] The result is precisely the same estimate in (2.45).

This solution can also be visualized in three dimensions as the "projection" of an intersection of surfaces in four dimensions. The four dimensional surfaces involved are the probability simplex, the varieties (all the same on the simplex), and an appropriately chosen sphere centered at the empirical relative frequencies $\widehat{\boldsymbol{\pi}}_{EMP}$. This illustration is provided in Figure 2.14.

[10]If there is more than one root in [0,1], then we repeat the process for each root.

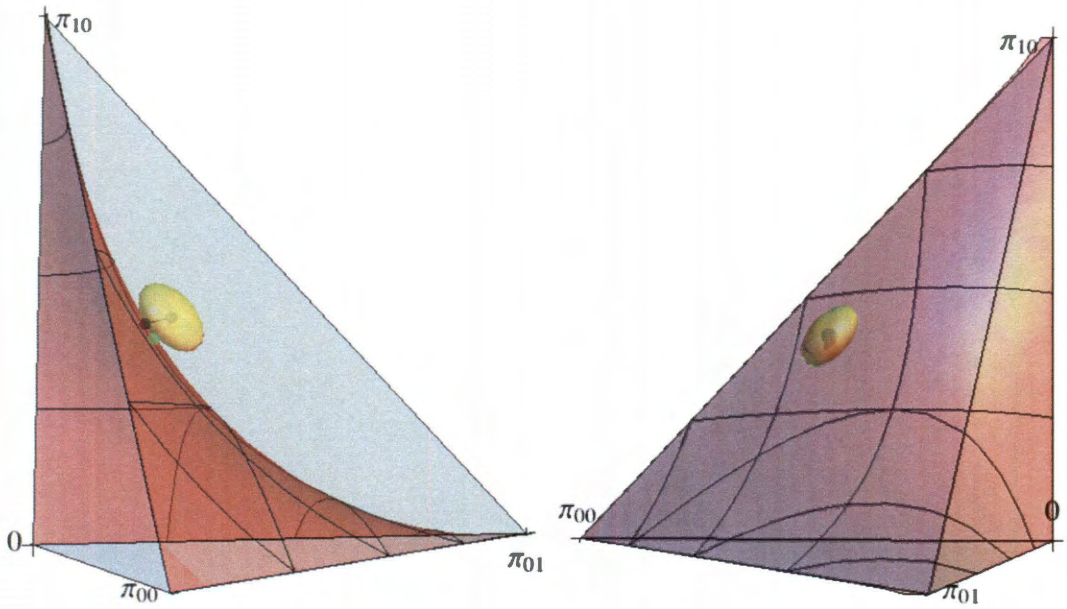

Figure 2.14 : The determination of $\widehat{\boldsymbol{\pi}}_{L2E}$ as the intersection of surfaces projected

2.3 Pearson's X^2 goodness-of-fit test

This last example views the same problem from the last example from the perspective of hypothesis testing. While in the past two examples we have been interested in estimation problems, in this example we are concerned with an inferential procedure well known to statisticians and non-statisticians alike – Pearson's chi-squared (X^2) goodness-of-fit test. Since this thesis is devoted to estimation, this section serves to 1. introduce Pearson's chi-squared statistic and distance, 2. show that the classical test is only half-heartedly based on minimum distance, and 3. motivate future work in testing for conditional independence models using the minimum distance paradigm and in particular the asymptotic theory in Chapter 4.

Proposed originally by Karl Pearson (Pearson [1900]), the X^2 test can now be seen to be one of the early achievements of asymptotic theory in statistics. The original

idea of the test was to be a goodness-of-fit test for categorical data; that is, a test designed to determine whether or not a set of observed data came from a particular probability distribution.

In its modern form, the setup is as follows. Suppose that $X_1, \ldots, X_N$ are N iid copies of some discrete random variable with a finite number of outcomes with probabilities $P[X = k] = \pi_k$, $k = 1, \ldots, r$. Pearson showed that if $T_1, \ldots, T_r$ are the counts of each outcome with $\sum_{k=1}^{r} T_k = N$ known and each of the probabilities is positive, then

$$X^2 := \sum_{k=1}^{r} \frac{(T_k - N\pi_k)^2}{N\pi_k} \xrightarrow{d} \chi^2_{r-1}, \tag{2.68}$$

that is, X^2 converges in distribution to a random variable which follows a chi-square distribution with $r - 1$ degrees of freedom. Using this result, we can test whether the samples $X_1, \ldots, X_N$ can reasonably be said to have been from the distribution $\boldsymbol{\pi} = [\pi_1 \ \cdots \ \pi_r]'$ by calculating X^2 using $\boldsymbol{\pi}$ and then comparing it to, say, the 95th percentile of the chi-squared distribution with $r - 1$ degrees of freedom. If X^2 exceeds this percentile, we conclude that the sample did not in fact come from the distribution $\boldsymbol{\pi}$, since X^2 can only be large if the counts, T_k, are significantly different from what we would expect them to be under the null hypothesis. Thus, X^2 is seen as a measure of disparity between the empirical frequencies $\frac{T_1}{N}, \ldots, \frac{T_r}{N}$ and the proposed probability distribution $\boldsymbol{\pi} = [\pi_1 \ \cdots \ \pi_r]'$.

The test has since been adapted to testing for independence in contingency tables. Without getting into too much detail, if we assume that the categories labelled $1, \ldots, r$ in the above setup correspond to 00, 01, 10, and 11 in the 2×2 contingency table example, then we can test whether or not the data which constitute the contingency

table come from a distribution $\boldsymbol{\pi} = [\pi_{00}\ \pi_{01}\ \pi_{10}\ \pi_{11}]'$ by applying (2.68) –

$$X^2 = \sum_{j=0}^{1}\sum_{k=0}^{1} \frac{(T_{jk} - N\pi_{jk})^2}{N\pi_{jk}} \xrightarrow{d} \chi^2_{4-1} = \chi^2_3. \tag{2.69}$$

Of course, when presented with data we never know which $\boldsymbol{\pi}$ is the true distribution from which the data came, so we cannot use either of the results above in a test for independence unless we want to assume outright a specific distribution which exhibits independence. If we reject this option, we must first select a $\boldsymbol{\pi}$ against which to test the counts to make inference.

To overcome this obstacle, virtually every textbook on elementary statistics suggests the following procedure popularized by R. A. Fisher (see, e.g., Peck et al. [2008], Moore and McCabe [1998], or even Agresti and Franklin [2007]). Since we do not know what the correct $\boldsymbol{\pi}$ is, we estimate it with the distribution provided by the product of the empirical marginal distributions, a distribution which we have already seen to be equivalent to $\widehat{\boldsymbol{\pi}}_{MLE}$ in (2.42). Thus, we have

$$X^2_{MLE} = \sum_{j=0}^{1}\sum_{k=0}^{1} \frac{(T_{jk} - N\widehat{\pi}^{MLE}_{jk})^2}{N\widehat{\pi}^{MLE}_{jk}}, \tag{2.70}$$

which Fisher correctly demonstrated has an asymptotic distribution which is χ^2_1 under the same conditions as Pearson's result (not χ^2_3, contrary to Pearson's belief) if the hypothesis of independence is in fact correct. Since each of the quantities in (2.70) is known, we can therefore calculate X^2_{MLE} and draw inference based on the 95th percentile of the χ^2_1 distribution.

Upon further inspection, the careful observer will notice that something peculiar occurs in the inferential procedure just described. In particular, *two different metrics are being used in the same problem.* First, *the chi-squared metric* from (2.68) is used to determine the discrepancy between the empirical distribution and a probability

distribution. However, to select the probability distribution which best approximates the observed distribution out of all those which exhibit independence, we then use *the likelihood discrepancy* (the negative log-likelihood). One consequence of this "ruler inconsistency" is that there are in fact different estimators for $\boldsymbol{\pi}$ which still exhibit the property of independence and which yield a smaller X^2 value. Since the hypothesis test is designed to reject if X^2 is too large, this switching of rulers can lead to different inferential conclusions; specifically, it will inadvertently increase the probability of type one error (simultaneously increasing power, of course) by causing the test to reject more often.

Using (2.68) as the discrepancy upon which to draw inference, the obvious choice for the estimator of $\boldsymbol{\pi}$ which one should use in (2.69) is that which minimizes X^2 itself since it would be the least likely distribution in the model to reject the null hypothesis of independence. Such an estimator is called a minimum chi-squared (X^2) estimator $\widehat{\boldsymbol{\pi}}_{X^2}$ and has a small but palpable and now dated literature. It is one of the estimators considered in this thesis as a minimum distance estimator. In the 2×2 case presented in the previous example, we have

$$\widehat{\boldsymbol{\pi}}_{X^2} = \begin{bmatrix} .430 \\ .116 \\ .358 \\ .096 \end{bmatrix}. \tag{2.71}$$

If we calculate X^2 using (2.42), (2.45), and (2.71), we obtain

$$X^2_{MLE} = .808 \qquad X^2_{L2E} = .892 \qquad X^2_{X^2} = .789. \tag{2.72}$$

Now, since the 95th percentile of the χ^2_1 distribution is 3.84, none of these test statistics is at risk for rejecting the independence hypothesis. However, the question remains

as to which estimator to use both for the estimation procedure as well as the testing procedure.

If we use $\widehat{\pi}_{X^2}$ for the estimator of π in (2.70), we obtain a ruler consistent test for independence. It has already been noted that G^2 is a ruler consistent test since it uses the likelihood paradigm for both estimation and testing. Similarly, we can imagine another ruler consistent test based on the minimum Euclidean distance paradigm for estimation. Unfortunately, there is no ruler consistent test using the minimum Euclidean distance metric.

One last point should be made before the conclusion of this example. It might be argued that the X^2 statistic usually used for the independence test (2.70) is not selected on the grounds that it uses the MLE, but rather on the grounds that it is, by construction, a distribution which satisfies the three properties of interest – it incorporates the data, it exhibits independence, and it has an asymptotic normal distribution. The fact that this estimator is also the MLE being coincidental, there is no reason to investigate the nature of the test statistic resulting from the use of other distributional estimators or even other "ruler consistent" estimators. This view is not entirely without merit. However, while the argument is true, practically convenient, easily understandable, and has many other nice properties, it is, at its heart, not a statistical criticism. Statistics is not only concerned with the analysis of data, but the optimal means by which one might do so. Thus, statistics is concerned both with the applicability of the procedure, as with the current state of affairs *and* the more theoretical consideration.

Part I

Theory

Chapter 3

Conditional Independence Models

The intent of this chapter is to give an account of the conditional independence models which form the basis of this work. The motivating characteristic of these models is that they ideally capture the special experimental relationships of relevance and irrelevance among the variables in the contingency table whose probabilistic nature they describe. The terms relevance and irrelevance are the experimental analogues of the statistical (read mathematical) notions of conditional dependence and conditional independence. A thorough understanding of conditional independence models is essential to understanding this work.

As discussed in Chapter 1, we are concerned exclusively with discrete multivariate models with finite sample spaces. Such data are typically presented as contingency tables. For this reason, Section 3.2 of this chapter is dedicated to outlining contingency tables in their most general form, taking us from the simple 2×2 contingency table of elementary applied statistics to the most general multi-way table. In Section 3.3, we move the discussion into conditional independence and properly define the conditional independence models. We close the chapter with a brief discussion of the more general algebraic statistical models. However, before approaching these we begin with a brief discussion of the fundamental statistical framework we assume in order to situate our understanding of how this work fits in to the overarching statistical endeavor.

3.1 Experimental and statistical models

The statistician always has in mind the foundational concept of an experiment. The totality of possible outcomes of the experiment, denoted Ω, he calls the outcome or sample space. Associated with this outcome space he has in mind a probability space $(\Omega, \mathcal{B}, P)$ which completely describes the random mechanism which produces the data he sees, data typically being observations of a random vector $\boldsymbol{X}$ (or any measurable map) defined on the probability space. P, the probability measure associated with the experiment, is referred to as the population and is unknown. The game of the statistician is to determine P or some aspect thereof in light of the data he observes.

Without any further assumptions it is difficult to ascertain even very simple characterizations of P. So to make the task easier, in conjunction with expert collaborators the statistician imposes what is called an experimental model, a set of assumptions regarding the nature of the underlying phenomena of the experiment. Since not all probability measures exhibit behavior consistent with the experimental model, in selecting an experimental model he tacitly narrows the search for P by looking only among the subset $\mathcal{M}$ of all possible probability measures Δ_P which conform to the experimental model. Consequently, the subset $\mathcal{M}$ is referred to as a statistical model since it is the mathematical manifestation of the assumptions made regarding the experiment.[1] While both are typically infinite, the subset $\mathcal{M}$ is almost always much "smaller" than Δ_P itself.

[1]While this is the classical view of statistical modeling, it is very often the case in modern applications that the game is played in reverse. That is, oftentimes a statistical model $\mathcal{M}$ is first posited as a subset of the set of all possible probability measures Δ, perhaps for mathematical convenience or even tractability, and then the mathematical assumptions are interpreted into an experimental model and the pair of models are accepted or rejected on reasonable grounds.

The models in this chapter are characterized by exhibiting the statistical property of conditional independence, but as we will see in the coming sections conditional independence itself is not the fundamental motivation for the models. Association is. The experimental models to which the conditional independence models correspond naturally precede the conditional independence models in terms of the scientific process; they consist of experimental beliefs about the conditional association and non-association of variables in the experiment. In other literature, these experimental notions of conditional association and non-association are called relevance and irrelevance (resp.) and so that terminology is used here as well (Pearl [1988], Pearl [2000]). Before we can explore this concept fully, however, we need to understand a bit more about the setup of a discrete multivariate problem.

3.2 Contingency tables and the problem setup

The primary consideration of this thesis is conditional independence models for cross-classified data, that is, contingency tables. In particular, we are interested in estimating the distribution of multi-way tables under the hypothesis of a conditional independence model.[2] In addition to the interesting statistical problems which come from this sort of endeavor, estimation in this setting has the potential to produce much richer interpretations concerning the underlying experimental phenomena than do most estimation problems because of the experimental implications of relevance (or irrelevance) and causality associated with conditional independence, and more specif-

[2]Since this work is focused primarily on the issue of estimation, by "hypothesis" we mean the single hypothesis of the model which simultaneously represents each of the conditional independence statements. The notion of testing each hypothesis individually (which would introduce a multiple testing problem) is also meaningful and left for future work.

ically graphical, models. While not emphasized in the current work, philosophically this manner of reasoning is the primary motivation for conditional independence models in practice and therefore should be on the reader's mind throughout this chapter in addition to the more mathematical and statistical discussion presented. For a more thorough explication of the more philosophical aspects of conditional independence models, see Pearl [1988] or Pearl [2000].

A vast portion of the statistical canon is devoted to the analysis of contingency tables. This is not surprising, considering that contingency tables are the fundamental tools used to analyze associations in any problem involving discrete multivariate data and even many involving continuous data by means of a binning procedure. The volume of literature is far too large for a general discussion in any work not solely dedicated to that purpose, so here we will present a more focused bibliography. A broad high level introduction to contingency tables is typically provided in a doctoral level text on mathematical statistics. Excellent examples include Lehmann and Casella [2003], Lehmann and Romano [2005], and Shao [2003]. A more direct and yet still general introduction to contingency tables is provided in, for example, Agresti [2002]. For our discussion, the online short-text Lauritzen [2002] provides an excellent and free alternative. Another nice overview with a more statistical flavor is provided in Darroch et al. [1980].

As Lauritzen [2002] and Edwards [2000] note, the primary difficulty with the generalization from 2-way contingency tables ("$R \times C$ tables") to multi-way tables is notational. The problem is a struggle between a notation which is explicit enough to readily interpret at a glance, such as P[red hair, blue eyes, American], and a notation which is concise and general enough to be applicable to a wide class of similar problems, such as π_{421}. For this reason we begin with an introduction to a simple and

systematic notation used for contingency tables and illustrate it with several examples. While the notation for contingency tables may feel cumbersome at first, after a little use it becomes second nature.

A contingency table is an array of counts of cross-classified data. The cross-classified data are themselves built from a sequence of outcomes $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N$ of an experiment. Each individual outcome $\boldsymbol{\omega}_k = [\omega_1^{(k)} \cdots \omega_p^{(k)}]'$ consists of p measurements on the same subject.[3] The p measurements are the observed values of p different factors or classification criteria for the kth subject. The set of classification criteria we denote $\mathcal{K}$. It is often the interactive effects of the criteria in $\mathcal{K}$ which is of interest to the experimenter. Since there are p such criteria, $|\mathcal{K}| = p$.

An assumption made throughout this work is that each classification criterion has a finite number of possible classes called levels. Mathematically, for each $j \in [p]$, Ω_j is used to denote the set of levels for the jth classification criterion, and it contains $r_j := |\Omega_j| < \infty$ elements. Using the vector notation, $\boldsymbol{\omega}_k \in \boldsymbol{\Omega} = \prod_{j=1}^{p} \Omega_j$. Since the individual outcome sets Ω_j are finite, the collection of all possible outcomes $\boldsymbol{\Omega}$ is finite; it contains $r := \prod_{j=1}^{p} r_j$ possible classes or combinations called cells.[4]

EXAMPLE 3.1 *A physician is interested in determining whether or not a patient's gender (male, female), ethnicity (caucasian, african-american, hispanic, asian, or*

[3]The prime notation $'$ continues to denote transpose. The labeling of the elements of a vector $\boldsymbol{\omega}_k$ which already has a subscript can be confusing; the convention followed in this thesis is that the subscript of the vector (e.g. k) becomes the superscript, allowing for the "new" subscript to denote the element index. If the superscript is a number, it is put in parentheses to distinguish from exponentiation.

[4]Some cells may have probability zero. For example, if "male" is in the set Ω_1, and "pregnant" is in the set Ω_2, then clearly the outcome [male, pregnant]' cannot occur; nevertheless, to retain the rectangular structure of $\boldsymbol{\Omega}$ we retain this possibility.

other), and blood type (A, B, AB, or O) has an effect on whether a patient will have a positive, negative, or no response to a treatment. The classification criteria are therefore $\mathcal{K} = \{$*gender, ethnicity, blood type, response*$\}$*. The levels of gender are* $\Omega_1 = \{$*male, female*$\}$*, those of blood type are* $\Omega_3 = \{A,\ B,\ AB,\ O\}$*, and so on, with sizes* $r_1 = 2$*,* $r_2 = 5$*,* $r_3 = 4$*, and* $r_4 = 3$*. A typical sample of five patients would look like*

$$\begin{aligned} \boldsymbol{\omega}_1 &= [\textit{male, hispanic, AB, no response}]' \\ \boldsymbol{\omega}_2 &= [\textit{male, african-american, B, negative}]' \\ \boldsymbol{\omega}_3 &= [\textit{female, asian, AB, positive}]' \\ \boldsymbol{\omega}_4 &= [\textit{male, asian, AB, negative}]' \\ \boldsymbol{\omega}_5 &= [\textit{female, caucasian, O, positive}]' \end{aligned} \tag{3.1}$$

The set of cells is therefore $\boldsymbol{\Omega} = \{[\textit{male, caucasian, A, positive}]', \ldots,$ $[\textit{female, other, O, no response}]'\}$*. Obviously, each of the observations* $\boldsymbol{\omega}_1$*,* $\boldsymbol{\omega}_2$*,* $\boldsymbol{\omega}_3$*,* $\boldsymbol{\omega}_4$*, and* $\boldsymbol{\omega}_5$ *are in* $\boldsymbol{\Omega}$*, and the total number of cells is* $r = \prod_{j=1}^{4} r_j = 120$*.*

In practice, it is highly inconvenient to work with the $\boldsymbol{\omega}_k$'s themselves because their labels are so long. For this reason as well as abstraction, we relabel the outcomes $\boldsymbol{\omega}$ with a random vector $\boldsymbol{X}(\boldsymbol{\omega})$, often referred to simply as $\boldsymbol{X}$. Specifically, we have

$$\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{\omega}) = \begin{bmatrix} X_1(\boldsymbol{\omega}) \\ \vdots \\ X_p(\boldsymbol{\omega}) \end{bmatrix} = \begin{bmatrix} X_1(\omega_1) \\ \vdots \\ X_p(\omega_p) \end{bmatrix}, \tag{3.2}$$

where the X_j's take on values in $[r_j]$ according to the level the associated ω_j assumes. In other words, in addition to labeling the classification criteria $\mathcal{K}$ with numbers, we also label the levels of each criterion with numbers, and since there are r_j levels of the jth criterion, we conveniently choose $[r_j]$ for the labels. We call the relabeled

sets of levels $\mathcal{X}_j$, so that $X_j \in \mathcal{X}_j = [r_j] = \text{Im}(\Omega_j)$ is the image of Ω_j under the component transformation X_j. As a consequence, the totality of possibilities for $\boldsymbol{X}$ is therefore $\boldsymbol{\mathcal{X}} := \prod_{j=1}^{p}[r_j]$ and corresponds to $\boldsymbol{\Omega}$ as its image under the transformation $\boldsymbol{X}$. Of course, since we are only relabeling the outcomes, the number of outcomes $|\boldsymbol{\Omega}| = |\boldsymbol{\mathcal{X}}| = r$ remains unchanged.

EXAMPLE 3.2 *Using the content of Example 3.1, we reformulate the problem using the labeling of a random vector. Thus, the observations* $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, $\boldsymbol{\omega}_3$, $\boldsymbol{\omega}_4$, *and* $\boldsymbol{\omega}_5$ *are converted to the observations called* $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, *and* $\boldsymbol{x}_5$,

$$\begin{aligned} \boldsymbol{x}_1 &= [1,\ 3,\ 3,\ 3]' \\ \boldsymbol{x}_2 &= [1,\ 2,\ 2,\ 2]' \\ \boldsymbol{x}_3 &= [2,\ 4,\ 3,\ 1]' \\ \boldsymbol{x}_4 &= [1,\ 4,\ 3,\ 2]' \\ \boldsymbol{x}_5 &= [2,\ 1,\ 4,\ 2]' \end{aligned} \tag{3.3}$$

Notice that the labels given correspond to the order in which they were first presented. For example, the third element of $\boldsymbol{x}_2$, *written* $x_3^{(2)}$, *is 2 because*

$$X_3(\boldsymbol{\omega}) = X_3(\omega_3) = \begin{cases} 1 & \omega_3 = A \\ 2 & \omega_3 = B \\ 3 & \omega_3 = AB \\ 4 & \omega_3 = O \end{cases} \tag{3.4}$$

and $\omega_3^{(2)} = B$. *Again, if not provided explicitly as in (3.4), the component function (labeling process) is assumed to be respective to the presentation of the levels.*

While this kind of labeling is useful, we will find two more kinds of labeling helpful as well, particularly for estimation theory. Focusing on the fact that there

are $r = \prod_{j=1}^{p} r_j = |\Omega| = |\boldsymbol{\mathcal{X}}|$ cells, the second labeling relabels the outcomes from $\boldsymbol{\mathcal{X}} \in \mathbb{Z}_{\geq 0}^{p}$ with a standard basis vector of $\mathbb{R}^r$ indicating which cell the observation belongs to. We call this the "$\boldsymbol{Y}$" label. For example,

$$\boldsymbol{x}_1 = \boldsymbol{X}(\omega_1) = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 3 \end{bmatrix} \overset{\boldsymbol{Y}}{\mapsto} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{y}_1 \in \mathbb{Z}^r. \tag{3.5}$$

The collection of basis vectors of $\mathbb{R}^r$ we call $\boldsymbol{\mathcal{Y}}$, which is the image of $\boldsymbol{\mathcal{X}}$ under the transformation $\boldsymbol{Y}$. The position of the 1 in the $\boldsymbol{y}$ representation is the rank of $\boldsymbol{x}$ in the ordering of $\boldsymbol{\mathcal{X}}$. This rank we give the label "Z". The whole process can be kept consistent with the use of a total order. For instance, using the lexicographic order so that

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} < \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \end{bmatrix} < \begin{bmatrix} 1 \\ 1 \\ 1 \\ 3 \end{bmatrix} < \begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \end{bmatrix} < \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix} < \cdots < \begin{bmatrix} 1 \\ 3 \\ 3 \\ 2 \end{bmatrix} < \begin{bmatrix} 1 \\ 3 \\ 3 \\ 3 \end{bmatrix} < \cdots < \begin{bmatrix} 2 \\ 5 \\ 4 \\ 2 \end{bmatrix} < \begin{bmatrix} 2 \\ 5 \\ 4 \\ 3 \end{bmatrix}, \tag{3.6}$$

the 1 in (3.5) is the 33rd element and so receives the Z label 33.

A last note on the labeling process is very useful in practice. It is very common to see practitioners condense the labels even further. This was in fact done without mention in the 2×2 contingency table example in Chapter 1. *If the meaning is un-*

ambiguous[5] then the labels can be condensed further by removing the vector notation altogether and simply running the numbers together. This is what we do in written language, where "one" is a condensation of the character vector [o, n, e]'. We call this the "*C*" representation. This convention is almost always used respective of the $\boldsymbol{X}$ labeling convention as opposed to the $\boldsymbol{Y}$ convention because $\boldsymbol{X}$ requires only p digits while $\boldsymbol{Y}$ requires r (notice that by design $r >> p$).

EXAMPLE 3.3 *Using the content of Example 3.2, we can further reduce* $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, *and* $\boldsymbol{x}_5$ *to*

$$\boldsymbol{x}_1^l = 1333 \qquad \boldsymbol{x}_2^l = 1222 \qquad \boldsymbol{x}_3^l = 2431 \qquad \boldsymbol{x}_4^l = 1432 \qquad \boldsymbol{x}_5^l = 2142. \tag{3.7}$$

In summary, Table 3.1 is a helpful at-a-glance guide for remembering the notational conventions.

Now that we can easily and efficiently refer to cells (and trace them back to their original experimental meaning), we can begin discussing probability measures for the kinds of experiments we are considering. Since there are r possible outcomes, the joint distribution of $\boldsymbol{X}$ or any of its equivalents is completely described by the r numbers

$$\pi_{\boldsymbol{x}} = \pi_{x_1 \cdots x_p} := P\left[X_1 = x_1, \ldots, X_p = x_p\right] = P\left[\boldsymbol{X} = \boldsymbol{x}\right], \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \tag{3.8}$$

for each $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \boldsymbol{\mathcal{X}}$. Viewed collectively, $\boldsymbol{\pi} = \{\pi_{\boldsymbol{x}}\}_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}}$ is sometimes referred to as the probability table, since we can visualize it in a multi-dimensional array format. Using the Z representation, we can also stack up (or "melt") the elements of the array into one big vector, $\boldsymbol{\pi} = [\pi_z]_{z \in [r]}$. Since these are simply different

[5]Usually meaning if no criterion has more than ten categories. This simply depends on the alphabet used for the labels, which here is the integers 0-9. If we used Roman characters, we could represent up to 26 categories.

ω	$\boldsymbol{X}$	Z	$\boldsymbol{Y}$	C
$\begin{bmatrix} \text{male} \\ \text{hispanic} \\ \text{AB} \\ \text{no response} \end{bmatrix}$	$\begin{bmatrix} 1 \\ 3 \\ 3 \\ 3 \end{bmatrix}$	33	$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$	1333

Table 3.1 : Labeling conventions

representations of the same information, we will think of $\boldsymbol{\pi}$ both ways. Of course, the laws of probability require that $\boldsymbol{\pi} \geq \mathbf{0}_r$ (component-wise) and $\mathbf{1}_r'\boldsymbol{\pi} = 1$. Now, the second condition implies that there are $r-1$ "free" parameters in determining the distribution of $\boldsymbol{X}$; this is the most general case conceivable. The collection of all such $\boldsymbol{\pi}$'s, that is, all probability distributions on r outcomes, is known collectively as the $(r-1)$ probability simplex, $\Delta_{r-1} \subset \mathbb{R}^r$.

The goal of contingency table analysis is to understand various aspects of the unknown distribution of $\boldsymbol{X}$, $\boldsymbol{\pi} = \{\pi_{\boldsymbol{x}}\}_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}$, in light of data. Therefore, suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ are N independent and identically distributed copies of $\boldsymbol{X}$. As a consequence of the independence assumptions, the order in which the data are observed is immaterial, so we typically reduce these raw data into the collection of counts of each of the r possible outcomes. To that end, let $T_{\boldsymbol{x}} = \sum_{k=1}^{N} 1_{[\boldsymbol{X}_k=\boldsymbol{x}]}$ denote the number of observations of each $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ in the collection of data. The collection of the $T_{\boldsymbol{x}}$'s, denoted $\boldsymbol{T}$, is therefore the random array of counts which we call the

contingency table. Just as $\boldsymbol{X}$ has r possible outcomes, $\boldsymbol{T}$ has r entries (cells). The number of components of $\boldsymbol{X}$, $|\mathcal{K}| = p$, is known as the dimension of the table. Just as the probability table can be lined up in a vector, so can the contingency table $\boldsymbol{T}$. Like $\boldsymbol{\pi}$, we will use the bold-faced $\boldsymbol{T}$ in both ways, both as the array and as the vector of counts, the meaning being clear from the context.

REMARK 3.1 Before we move into a few examples, a remark on sampling schemes is in order. In discussions of contingency tables one often talks of sampling schemes. A sampling scheme of a study or contingency table is a method by which the data, and more specifically the $T_{\boldsymbol{x}}$'s, are collected. In some studies, the number N is not selected beforehand and the experimenters simply count the number of respondents. In such cases it is reasonable to put a distribution on the number of observations seen in the table such as the Poisson distribution; this is known as Poisson sampling. In this work however we will always assume that the sample size $N = n$ is known a priori so that the distribution of the contingency table is multinomial with size parameter equal to the known sample size; this is called multinomial sampling. Thus we will always assume multinomial sampling and that, consequently, the full model for the table, also called the saturated model, is the multinomial model with sample size N and unrestricted probability vector $\boldsymbol{\pi}$ which for obvious reasons is associated with the probability simplex Δ_{r-1}. The conditional independence models are a submodel of this full model in the sense that only certain probability vectors $\boldsymbol{\pi}$ are allowed.

EXAMPLE 3.4 *Snee [1974] presents the data in Table 3.2 regarding the hair color of 592 subjects collected by students at the University of Delaware. In this example,* $N = 592$, $\mathcal{K} = \{$*hair color*$\}$, $p = 1$, $r = r_1 = 4$, $\boldsymbol{\Omega} = \Omega_1 = \{$*black, brunette, red, blonde*$\}$, *the* ω_k*'s are individual subjects for* $k \in [592]$, *and the* $T_{\boldsymbol{x}}$ *are the counts of each element*

of $\boldsymbol{\mathcal{X}} = [r_1] = [4]$. $\boldsymbol{T}$ *is therefore simply the numbers in Table 3.2.*

black	108
brunette	286
red	71
blonde	127

Table 3.2 : Hair color of 592 subjects from Snee [1974]

EXAMPLE 3.5 *The same data set was expanded in Snee [1974] to also contain eye color. This data set is presented in Table 3.3. In this example,* N *is still 592, but* $p = 2$ *so that the observations are observations of a random vector with* $p = 2$ *components,* $\boldsymbol{X} = [X_1,\ X_2]'$. $\mathcal{K} = \{$*hair color, eye color*$\}$. *For every* $k \in [592]$, ω_1 *takes on values in* $\Omega_1 = \{$*black, brunette, red, blonde*$\}$ *with labels* $X_1 = X_1(\omega_1)$ *in* $\mathcal{X}_1 = [4]$ *so that* $r_1 = 4$. *Similarly,* ω_2 *takes on values in* $\Omega_2 = \{$*brown, blue, hazel, green*$\}$ *with labels* $X_2 = X_2(\omega_2)$ *in* $\mathcal{X}_2 = [4]$ *so that* $r_2 = 4$ *also. Therefore the random vectors* $\boldsymbol{X}_k$ *take on values which are pairs of these two, that is,* $\boldsymbol{\Omega} = \{$(*black, brown*), (*black, blue*), $\ldots$, (*blonde, green*)$\}$ *labeled* $\boldsymbol{\mathcal{X}} = \{[1,1]', [1,2]', \ldots, [4,4]'\}$. *There are* $r = r_1 r_2 = 16$ *such pairs.*

Now, for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $T_{\boldsymbol{x}}$ *is the number of that type of individual in the data set. For example,* $T_{(red,blue)} = T_{(3,2)} = T_{32}$ *is the number of individuals found with red hair and blue eyes. Viewed collectively, if* $N = n$ *is known then without any assumptions* $\boldsymbol{T}$ *is multinomially distributed* $\boldsymbol{T} \sim Multinom_{16}(n = 592, \boldsymbol{\pi})$, *this is the saturated model. This distribution has* $16 - 1 = 15$ *free parameters, since* $\mathbf{1}_{16}'\boldsymbol{\pi} = \sum_{\boldsymbol{x}} \pi_{\boldsymbol{x}} = 1$. *Having* $p = 2$ *dimensions, the contingency table* $\boldsymbol{T}$ *is called a two-way table and can be seen to be a generalization of the* 2×2 *contingency table example in Chapter 1.*

	brown	blue	hazel	green
black	68	20	15	5
brunette	119	84	54	29
red	26	17	14	14
blonde	7	94	10	16

Table 3.3 : Hair and eye color of 592 subjects from Snee [1974]

Marginalization

It is easy to look at Tables 3.2 and 3.3 from Examples 3.4 and 3.5 and realize that Table 3.2 can be determined from Table 3.3 by simply summing the rows and then throwing away everything but the sums. The result is just the information on the subjects' hair color. This constitutes a loss of information; we have irreversibly lost the information about the subjects' eye color. However, the result is a smaller table. This process of neglecting certain criteria variables by summing over them is called marginalization, and the resulting table is known as the marginal table. As a convention, if $\mathcal{A} \subseteq [p]$ is a subset of the classification criteria labels, we define $\boldsymbol{x}_{\mathcal{A}} \in \prod_{k \in \mathcal{A}} \mathcal{X}_k$ to be the restriction of the vector $\boldsymbol{x}$ to the classification criteria referenced in $\mathcal{A}$, and similarly with $\boldsymbol{\mathcal{X}}_{\mathcal{A}}$. We define the marginal table $\boldsymbol{T}_{\mathcal{A}}$ to be the array $\{T_{\boldsymbol{x}_{\mathcal{A}}}\}_{\boldsymbol{x}_{\mathcal{A}} \in \boldsymbol{\mathcal{X}}_{\mathcal{A}}}$ where each count $T_{\boldsymbol{x}_{\mathcal{A}}}$ is defined

$$T_{\boldsymbol{x}_{\mathcal{A}}} := \sum_{\boldsymbol{j} \,:\, \boldsymbol{j}_{\mathcal{A}} = \boldsymbol{x}_{\mathcal{A}}} T_{\boldsymbol{j}}. \tag{3.9}$$

We move from the probabilities in the original table $\boldsymbol{\pi} = \{\pi_{\boldsymbol{x}}\}_{\boldsymbol{\mathcal{X}}}$, which represent the probability of $\boldsymbol{X}$ being classified to the $\boldsymbol{x}$th cell, to the probabilities in the marginal table the same way by summing over the probabilities on the cells assigned

to variables we are forgetting. Thus, the marginal probability of an individual $\boldsymbol{X}$ being classified to the $\boldsymbol{i}_{\mathcal{A}}$th category is

$$\pi_{\boldsymbol{x}_{\mathcal{A}}} = \sum_{\boldsymbol{j} \,:\, \boldsymbol{j}_{\mathcal{A}} = \boldsymbol{x}_{\mathcal{A}}} \pi_{\boldsymbol{j}}, \tag{3.10}$$

and the marginal distribution is $\boldsymbol{\pi}_{\mathcal{A}}$.

The run-together C notation is particularly helpful for the marginalization operation. The operations in (3.9) and (3.10) can be equivalently written with the + sign in the $[p] \setminus \mathcal{A}$ digits. For example, if $p = 5$ and $\mathcal{A} = \{1, 2, 4\}$, then the marginal distribution of $[X_1, X_2, X_4]'$ can be written either $\boldsymbol{\pi}_{\{1,2,4\}}$ or $\boldsymbol{\pi}_{x_1 x_2 + x_4 +}$ (cell probabilities are written the same way without the bold face). In general, the + notation is used as a short hand denoting summation over that index. A similar notation, the $\bullet$ notation, is used as a short hand denoting averaging over that index; it is typically used in the context of log-linear models due to its similarities with the analysis of variance.

EXAMPLE 3.6 *Considering Table 3.3 in Example 3.5, we can write out the associated probability table as in Table 3.4. If $\mathcal{A} = 2$, then the marginal distribution corresponds to the second classification criterion, i.e., eye color.*

	brown	**blue**	**hazel**	**green**
black	π_{11}	π_{12}	π_{13}	π_{14}
brunette	π_{21}	π_{22}	π_{23}	π_{24}
red	π_{31}	π_{32}	π_{33}	π_{34}
blonde	π_{41}	π_{42}	π_{43}	π_{44}

Table 3.4 : Concisely referenced probability table of Table 3.4 Snee [1974]

From (3.10) we have that

$$\pi_{+1} = \sum_{j=1}^{4} \pi_{j1} = \sum_{\boldsymbol{x}} \pi_{\boldsymbol{x}_{\{2\}}} = \pi_{11} + \pi_{21} + \pi_{31} + \pi_{41}, \tag{3.11}$$

and the marginal distribution itself is

$$\boldsymbol{\pi}_{\{2\}} = \boldsymbol{\pi}_{+x_2} = \begin{bmatrix} \pi_{+1} \\ \pi_{+2} \\ \pi_{+3} \\ \pi_{+4} \end{bmatrix} = \begin{bmatrix} \pi_{11} + \pi_{21} + \pi_{31} + \pi_{41} \\ \pi_{12} + \pi_{22} + \pi_{32} + \pi_{42} \\ \pi_{13} + \pi_{23} + \pi_{33} + \pi_{43} \\ \pi_{14} + \pi_{24} + \pi_{34} + \pi_{44} \end{bmatrix} \in \Delta_{4-1} \subset \mathbb{R}^4 = \mathbb{R}^{r_2}. \tag{3.12}$$

3.3 Conditional independence models

The experimental model assumptions of relevance and irrelevance concerning the underlying experiment summarized by a contingency table often induce statistical model assumptions in the form of a conditional independence structure. Such models are called conditional independence models which we will explore in detail in this section.

Since the sample spaces in this work are finite, every random vector exhibits a joint density with respect to the counting measure, an object called a probability mass function (pmf) in elementary statistics. This means that for every random vector $\boldsymbol{X}$ we discuss, $\boldsymbol{X}$ admits a density $f_{\boldsymbol{X}}$ with respect to the counting measure so that $P[\boldsymbol{X} \in \mathcal{A}] = \sum_{\boldsymbol{x} \in \mathcal{A}} f_{\boldsymbol{X}}(\boldsymbol{x})$.

Having a basic understanding of conditional independence is essential for understanding conditional independence models. In the classical sense, the motivating idea behind the independence of two random variables X and Y is that in knowing one we gain no information pertaining to the other. Experimentally speaking, we say X is irrelevant in predicting Y. Probabilistically, this intuition is formally encoded

$$P[X \in \mathcal{A} \mid Y \in \mathcal{B}] = P[X \in \mathcal{A}], \tag{3.13}$$

or equivalently

$$P[X \in \mathcal{A}, Y \in \mathcal{B}] = P[X \in \mathcal{A}]\, P[Y \in \mathcal{B}] \tag{3.14}$$

for all events $\mathcal{A}$ and $\mathcal{B}$. This can be shown to be equivalent to the factorization of the joint density into the marginal densities so that

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{3.15}$$

for all x and y, or similarly[6]

$$f_{X|Y}(x|y) = f_X(x). \tag{3.16}$$

The motivating idea of conditional independence is almost identical – given that we observe $Z = z$, learning one of X or Y provides no additional knowledge concerning the other. While the idea is just as simple, the mathematics for general measures is much more involved; fortunately, we can keep it simple since the heavy duty mathematics are not required for the discrete case.

DEFINITION 3.1 For three random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, we say that $\boldsymbol{X}$ is conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, denoted $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, if and only if

$$f_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z}) f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}), \tag{3.17}$$

where for any two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ the conditional density is defined

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{x}_2) := \frac{f_{\boldsymbol{X}_1,\boldsymbol{X}_2}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)} \tag{3.18}$$

for all $\boldsymbol{x}_2$ such that $f_{\boldsymbol{X}_2}(\boldsymbol{x}_2) > 0$.

---

[6]We typically think of independence mathematically as (3.14) or (3.15), since they don't require any conditions concerning the existence of conditional distributions. In the cases where the conditional distributions exist, however, (3.13) and (3.16) more accurately represent the interpretation of no association, an interpretation which is made regardless of which definition is used.

The same conditional independence statement can be equivalently expressed in a variety of ways (for proofs, see the references below). The following expressions, although seemingly more general, are in fact equivalent to definition provided in (3.17) for some measurable h and g (not necessarily densities) :

$$f_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z}) = h(\boldsymbol{x},\boldsymbol{z})g(\boldsymbol{y},\boldsymbol{z}), \tag{3.19}$$

$$f_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z}), \tag{3.20}$$

$$f_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = h(\boldsymbol{x},\boldsymbol{z}). \tag{3.21}$$

In particular, (3.20) really captures the experimental notion of irrelevance – if we know that $\boldsymbol{Z} = \boldsymbol{z}$, then we acquire no additional information regarding $\boldsymbol{X}$ by learning $\boldsymbol{Y} = \boldsymbol{y}$.

A good low level review of these statistical ideas can be found in Casella and Berger [2002]. In his landmark paper *Conditional Independence in Statistical Theory*, Dawid [1979] provides an elegant and well written in-depth mathematical exposition of conditional independence which provides a unified perspective of many seemingly disparate statistical notions dating back to Fisher's landmark introduction of sufficiency, efficiency, etc. in Fisher [1922]. Still more detailed treatments can be found in Billingsley [1995], Resnick [1999], and especially Dawid [1980]. For more insight into why conditional independence is used in association with graphs and relevance, including a detailed philosophical theory of relevance and irrelevance, see the excellent and now canonical text Pearl [1988]. In addition to *Graphical Models* (Lauritzen [1996]), *Probabilistic Reasoning in Intelligent Systems* (Pearl [1988]) is an authoritative text which should be consulted for any further information concerning the interpretations of conditional independence.

Conditional independence in contingency tables

Conditional independence can, of course, be formulated using the contingency table friendly notation introduced in Section 3.2 as well, and this will be our primary means of precisely communicating such relationships. Let's start with a basic example.

EXAMPLE 3.7 *Consider the three-way table presented in Table 3.5 altered from Agresti [2002]. The table cross-classifies 100 patients on three different criteria – the treatment they received (X_1), the success of the treatment (X_2), and the clinic they attended (X_3).*

CLINIC 1

	success	**failure**
treatment A	18	12
treatment B	12	8

CLINIC 2

	success	**failure**
treatment A	9	5
treatment B	3	8

Table 3.5 : Data concerning two different treatments at two different clinics altered from Agresti [2002]

A priori it is not unreasonable to assume that since in an ideal situation treatment A at clinic one is identical to treatment A at clinic two, the outcome of the treatment is independent of the clinic at which treatment was administered provided we know what the treatment is.[7] *This is an experimental model of irrelevance which asserts that*

[7]This assumption may of course be false, just like a statistical model may be false. Even if that

once we know which treatment is administered, we learn no additional information towards the patients outcome by learning the clinic at which they receive that treatment. Mathematically, the statistical model corresponding to such an experimental model is

$$X_2 \perp\!\!\!\perp X_3 \mid X_1. \tag{3.22}$$

Note that this is different from stating simply that the clinic a patient visits is immaterial in predicting the patient's outcome, something that seems far less likely to be true. The statistical model corresponding to that experimental model is the unconditional (marginal) independence statement

$$X_2 \perp\!\!\!\perp X_3, \tag{3.23}$$

which is also considered to be a conditional independence model just with a null conditioning set. Of course, we can *impose either model; nothing stops us from making the assumption and cranking through the methods for an answer. However, it is clear in this case that the first model is preferable to the second model.*

How can we describe the conditional independence model in (3.22) using the definition of conditional independence and the contingency table notation? Using the notation in Definition 3.1, the entire joint distribution of $\boldsymbol{X}$ is equivalently written

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \pi_{\boldsymbol{x}} \qquad \text{for any } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}. \tag{3.24}$$

Using this together with the marginalization C notation (i.e., the $+$ notation), we

---

is the case, however, we are reminded of the old statistics adage "all models are wrong, some are useful."

have from (3.17) and (3.18)

$$\begin{array}{ccc} f_{X_2,X_3|X_1}(x_2,x_3|x_1) & = & f_{X_2|X_1}(x_2|x_1)f_{X_3|X_1}(x_3|x_1) \\ || & & || \\ \dfrac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{X_1}(x_1)} & & \dfrac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}\dfrac{f_{X_1,X_3}(x_1,x_3)}{f_{X_1}(x_1)} \\ || & & || \\ \dfrac{\pi_{x_1x_2x_3}}{\pi_{x_1++}} & & \dfrac{\pi_{x_1x_2+}}{\pi_{x_1++}}\dfrac{\pi_{x_1+x_3}}{\pi_{x_1++}} \end{array} \tag{3.25}$$

The above equation is one of the fundamental equations of this work. When we clear denominators, we have

$$\pi_{x_1x_2x_3}\pi_{x_1++} = \pi_{x_1x_2+}\pi_{x_1+x_3}, \tag{3.26}$$

which may be taken to be the definition of conditional independence in the context of multivariate categorical data as opposed to being somewhat derivative. Dividing through, we obtain what is called a cross product ratio (CPR) or odds ratio (OR) :

$$\frac{\pi_{x_1x_2x_3}\pi_{x_1++}}{\pi_{x_1x_2+}\pi_{x_1+x_3}} = 1. \tag{3.27}$$

The modern study of contingency analysis and indeed a large portion of categorical data analysis focuses heavily on such ratios; for more information, see Agresti [2002] or Bishop et al. [2007]. Another rearrangement of (3.26) that is gaining popularity among some circles is the cross product difference (CPD), with which the condition of conditional independence is

$$\pi_{x_1x_2x_3}\pi_{x_1++} - \pi_{x_1x_2+}\pi_{x_1+x_3} = 0. \tag{3.28}$$

This work will focus on this representation. For other works using cross product differences, see Geiger et al. [2006] or Drton et al. [2009].

Note that (3.26), (3.27), or (3.28) must hold for all x_1, x_2, and x_3, so there are actually $r_1r_2r_3 = 8$ different equations implied by each. Interestingly, these are

for this conditional independence example what equations (2.33) are for marginal independence in the 2×2 contingency table example in Chapter 1.

For an arbitrary random vector $\boldsymbol{X}$, conditional independence statements come in the form

$$\boldsymbol{X}_{\mathcal{A}} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{B}} \mid \boldsymbol{X}_{\mathcal{C}}, \tag{3.29}$$

where $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are disjoint subsets of $[r]$ and $\mathcal{C}$ is allowed to be empty. For example,

$$(X_1, X_3) \perp\!\!\!\perp X_5 \mid X_4 \tag{3.30}$$

and

$$(X_1, X_3, X_7) \perp\!\!\!\perp (X_5, X_8, X_{14}, X_{16}) \mid (X_3, X_{15}). \tag{3.31}$$

A conditional independence model is a collection of such statements. Consequently, each of these has a description in terms of the $\pi_{\boldsymbol{x}}$'s analogous to (3.26); however, there is currently no notation to generalize the simple conditional independence model described by (3.26). To remedy this, the following is a simple thought process used by the author to convert conditional independence statements into systems of equations.

Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ be a partition of $[r]$. Then (3.29) will result in a system of equations defined through a single equation having to hold for every possible combination of $\boldsymbol{x}$ analogous to (2.33) and (3.26). That equation we write as having four factors, as in (3.26), which we denote

$$\pi_{\boldsymbol{x}}^{(1)} \pi_{\boldsymbol{x}}^{(2)} = \pi_{\boldsymbol{x}}^{(3)} \pi_{\boldsymbol{x}}^{(4)}, \tag{3.32}$$

with equivalent odds ratio condition (as long as everything is positive)

$$\frac{\pi_{\boldsymbol{x}}^{(1)} \pi_{\boldsymbol{x}}^{(2)}}{\pi_{\boldsymbol{x}}^{(3)} \pi_{\boldsymbol{x}}^{(4)}} = 1, \tag{3.33}$$

or cross product difference condition (without restriction)

$$\pi_{\boldsymbol{x}}^{(1)}\pi_{\boldsymbol{x}}^{(2)} - \pi_{\boldsymbol{x}}^{(3)}\pi_{\boldsymbol{x}}^{(4)} = 0. \tag{3.34}$$

Now, each of $\pi_{\boldsymbol{x}}^{(1)}$, $\pi_{\boldsymbol{x}}^{(2)}$, $\pi_{\boldsymbol{x}}^{(3)}$, and $\pi_{\boldsymbol{x}}^{(4)}$ will have +'s in the indices included in $\mathcal{D}$ – these correspond to variables not referenced in the conditional independence statement yet are nevertheless present in $\boldsymbol{X}$. They are, as far as the relation is concerned, extraneous and therefore must be marginalized. $\pi_{\boldsymbol{x}}^{(1)}$ will have no +'s other than those of $\mathcal{D}$. $\pi_{\boldsymbol{x}}^{(2)}$ will have +'s in the indices provided by $\mathcal{A}$ and $\mathcal{B}$ in addition to those of $\mathcal{D}$. $\pi_{\boldsymbol{x}}^{(3)}$ will have +'s in the $\mathcal{B}$ and $\mathcal{D}$ slots, while $\pi_{\boldsymbol{x}}^{(3)}$ will have +'s in the $\mathcal{A}$ and $\mathcal{D}$ slots. Every factor has the indices in $\mathcal{C}$; this corresponds to the independence relation holding for each configuration of the conditioning variables.

Every conditional independence statement will induce a set of such equations. Collectively, these equations define the conditional independence model $\mathcal{M}$ as

$$\mathcal{M} = \{\boldsymbol{\pi} \in \mathbb{R}^r : \boldsymbol{\pi} \text{ satisfies (3.29)}\} \cap \Delta_{r-1} = \{\boldsymbol{\pi} \in \mathbb{R}^r : \boldsymbol{\pi} \text{ satisfies (3.34)}\} \cap \Delta_{r-1}. \tag{3.35}$$

For any specific conditional independence model provided by a series of conditional independence statements, the model $\mathcal{M}$ is constructed as the intersection of such models.

Comments on conditional independence models

Several general observations can be made concerning the conditional independence models just described. First, the systems of equations in (2.33), (3.26), and most generally (3.32) are systems of nonlinear *polynomial* equations. The cross product differences in (2.47), (3.28), and (3.33) are also systems of polynomial equations. We will call these systems conditional independence systems, the polynomials conditional

independence polynomials denoted by $\boldsymbol{h}(\boldsymbol{\pi}) \subset \mathbb{R}[\boldsymbol{\pi}]^k$, and their varieties (when intersected with the simplex) the conditional independence models themselves. The ideal $\langle \mathbf{1}_r'\boldsymbol{\pi} - 1, \boldsymbol{h}(\boldsymbol{\pi})\rangle = \left\langle \sum_{j=1}^r \pi_j - 1, h_1(\boldsymbol{\pi}), \ldots, h_k(\boldsymbol{\pi}) \right\rangle$ is called the conditional independence ideal. Like Geiger et al. [2006], we prefer using cross product differences to odds ratios. Second, the polynomials in the conditional independence systems are always quadratic and often homogeneous (the total degree of any monomial is equal to 2, disregarding the simplex condition). The varieties are therefore quadrics and can be thought of as being in affine or projective space by homogenizing when necessary. Third, *conditional independence models* $\mathcal{M} \subset \Delta_{r-1} \subset \mathbb{R}^r$ *are implicitly defined models which are fundamentally algebraic in nature.* As described in the 2×2 contingency table example in Chapter 1, for the most part statisticians are not used to working with implicitly defined models. Partly for this reason, historically statisticians have opted instead to use the more wieldy parametric loglinear models as surrogates. While many conditional independence models are log-linear, many are not and thus the need for proper methods for conditional independence models (Drton et al. [2009], Christensen [1997], Edwards [2000]).

3.4 Algebraic statistical models

The conditional independence models can be further generalized to a larger class of models which includes them. Notice that they are both defined as the intersection of an affine variety with the probability simplex. This leads to the more general notion of an algebraic statistical model whose description here follows Sullivant [2006].

DEFINITION 3.2 (ALGEBRAIC STATISTICAL MODEL) Let $\boldsymbol{h} = [h_1, \ldots, h_k]' \in \mathbb{R}[\boldsymbol{\pi}]^k$ be k elements of the polynomial ring in the r indeterminates $\pi_1, \ldots, \pi_r$ with real

coefficients. Then the algebraic statistical model generated by $\boldsymbol{h}$ is the model

$$\mathcal{M} = V(\boldsymbol{h}) \cap \Delta_{r-1} = V(h_1, \ldots, h_k) \cap \Delta_{r-1}. \tag{3.36}$$

We sometimes write this as

$$\mathcal{M} = V\left(\boldsymbol{h}^{\Delta}\right) \cap \mathbb{R}^r_{\geq 0}, \tag{3.37}$$

where $\boldsymbol{h}^{\Delta}$ is $\boldsymbol{h}$ along with the simplex polynomial $\mathbf{1}_r'\boldsymbol{\pi} - 1$ ($V(\mathbf{1}_r'\boldsymbol{\pi} - 1) = \Delta_{r-1}$ on the nonnegative orthant).

The notion of an algebraic statistical model is one which still has not found a consensus among algebraic statisticians. For example, Drton and Sullivant [2007] provide a more general description than the one above. However, the definition used here was one of the first successful attempts at a statistical framework amenable to notions from commutative algebra and algebraic geometry, see Pachter and Sturmfels [2005]. It is still the most common definition provided for such a model and, since it also suits our purposes here adequately, we assume it here. For a more general definition the reader is encouraged to see Drton and Sullivant [2007].

Before moving into more interesting aspects of conditional independence and algebraic statistical models, note the following fact.

PROPOSITION 3.1 *Conditional independence models, defined as series of statements of the form (3.29) with defining equations provided by (3.34), are algebraic statistical models.*

3.5 Results for conditional independence and algebraic statistical models

In this section we discuss various results which are useful in the study of conditional independence models. Many of these hold for arbitrary algebraic statistical models and are therefore inherited by conditional independence models.

We begin with a result from real algebraic geometry which appears as Proposition 2.1.3 in Bochnak et al. [1998].

PROPOSITION 3.2 *Given an affine variety* $V(\boldsymbol{h}) \subset \mathbb{R}^r$ *with* $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{x}]^k$, *there exists a single* $h \in \mathbb{R}[\boldsymbol{\pi}]$ *such that* $V = V(h)$.

The proof is constructive, simply set $h = \boldsymbol{h}'\boldsymbol{h}$ as the sums of squares of the h_k's (or any finite basis of $I(V) = \{f \in \mathbb{R}[\boldsymbol{\pi}] : f(\boldsymbol{\pi}) = 0 \; for \; \boldsymbol{\pi} \in V\}$, which exists by the Hilbert basis theorem).

3.5.1 Feasibility – checking for nonempty models

Let $\mathcal{M} = V(\boldsymbol{h}) \cap \Delta_{r-1}$ where $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$ be an algebraic statistical model. One problem which might be of interest is whether or not the model $\mathcal{M}$ is empty. For a conditional independence model, $\mathcal{M}$ being nonempty is an obvious condition for any minimum distance estimator to exist, but more importantly it is an essential condition for the experimental model in question. If $\mathcal{M} = \emptyset$ then the collection of conditional independence statements is inconsistent – the experimental assumptions are intrinsically impossible.

In optimization theory checking if $\mathcal{M} = \emptyset$ amounts to a feasibility problem, also known as a satisfiability or decision problem. In fact, the problem can be formulated in a very algebraic way as a consequence of the following theorem which appears as

Theorem 7.5 of Sturmfels [2002] (p. 94), originally proven in Stengle [1973]. It is a very powerful result which allows a number of the statistical considerations in this work to be formulated algebraically. A useful resource for understanding the result can be found in Laurent [2009].

THEOREM 3.1 (REAL NULLSTELLENSATZ)

The system of polynomial equations and inequalities

$$f_1(\boldsymbol{x}) = 0,\ f_2(\boldsymbol{x}) = 0,\ \ldots,\ f_r(\boldsymbol{x}) = 0, \tag{3.38}$$

$$g_1(\boldsymbol{x}) \geq 0,\ g_2(\boldsymbol{x}) \geq 0,\ \ldots,\ g_s(\boldsymbol{x}) \geq 0, \tag{3.39}$$

$$h_1(\boldsymbol{x}) > 0,\ h_2(\boldsymbol{x}) > 0,\ \ldots,\ h_t(\boldsymbol{x}) > 0, \tag{3.40}$$

either has a solution in $\mathbb{R}^n$ *or their exists a polynomial identity*

$$\sum_{i=1}^{r} a_i f_i + \sum_{\nu \in \{0,1\}^s} \left(\sum_j b_{j\nu}^2 \right) g_1^{\nu_1} \cdots g_s^{\nu_s} \tag{3.41}$$

$$+ \sum_{\nu \in \{0,1\}^t} \left(\sum_j c_{j\nu}^2 \right) h_1^{\nu_1} \cdots h_t^{\nu_t} + \sum_k d_k^2 + \prod_{l=1}^{t} h_l^{u_l} \ = \ 0, \tag{3.42}$$

where $u_j \in \mathbb{N}$ *and* a_i, $b_{j\nu}$, $c_{j\nu}$, d_k *are polynomials.*[8]

Applying this result to algebraic statistical models automatically allows us to reformulate the statistically relevant feasibility question – whether or not an experimental model is possible – as an algebraic problem involving polynomials.

PROPOSITION 3.3 (KAHLE) *Let* $\mathcal{M} = V\left(\boldsymbol{h}^{\Delta}\right) \cap \mathbb{R}^r_{\geq 0}$ *be an algebraic statistical model and define* $h = (\boldsymbol{h}^{\Delta})'(\boldsymbol{h}^{\Delta})$. *Then the following statements are equivalent.*

1. $\mathcal{M}$ *is nonempty and (consequently) the experimental assumptions are possible.*

2. *There exists a* $\boldsymbol{x} \in \mathbb{R}^r_{\geq 0}$ *with* $\boldsymbol{h}^{\Delta}(\boldsymbol{x}) = \boldsymbol{0}$.

[8]The last product is equal to one when empty.

3. *There do not exist sums of squares of polynomials* $g_\nu \in \sum(\mathbb{R}[\boldsymbol{x}])^2$ *and* $a \in \mathbb{R}[\boldsymbol{x}]$, *such that*

$$ah + \sum_{\nu \in \{0,1\}^r} g_\nu \boldsymbol{x}^\nu + 1 \equiv 0. \tag{3.43}$$

Moreover, a necessary condition for any of the above is that there exist no polynomials $a_0, a_1, \ldots, a_s \in \mathbb{R}[\boldsymbol{x}]$ *such that*

$$a_0 h + \sum_{k=1}^{s} a_k^2 + 1 = 0. \tag{3.44}$$

PROOF 3.1 The first and second conditions are clearly equivalent. The third follows from Proposition 3.2 and Theorem 3.1 by setting $f_1 = h$ and $g_i = x_i$ for $i = 1, \ldots, r$ to ensure the positive orthant condition. The g_ν's correspond to the sums of the $b_{j\nu}$'s in Theorem 3.1. The necessary condition follows from the same theorem by dropping that condition.

□

Chapter 4

Estimation

In this chapter we turn to point estimation theory. In Section 3.1 we noted that the game of the statistician is to make inferences concerning the unknown distribution $P \in \Delta_P$ from which the data are drawn. Incorporating experimental assumptions, we reduce our search for P by assuming it lies in a statistical model $\mathcal{M}_P \subset \Delta_P$. The task of estimation is therefore to select a distribution $\widehat{P} \in \mathcal{M}_P$ in the model from which the data most reasonably came. As with the rest of this thesis, we will be working entirely in the multivariate categorical setting assuming the sample size N is known. Thus, following Remark 3.1 the samples themselves will be assumed to follow a multinomial distribution with probability vector $\boldsymbol{\pi}$ constrained to a subset $\mathcal{M}_{r-1}$ of the simplex, and we use the $_{r-1}$ notation to emphasize this fact.

The field of discrete multivariate analysis typically considers a statistical model $\mathcal{M}_{r-1}$ to be parametric if there is a known function $\boldsymbol{\pi} : \Theta \to \Delta_{r-1}$ whose image is the model $\mathcal{M}_{r-1} = \mathrm{Im}(\boldsymbol{\pi})$ and whose domain, called the parameter space, is a subset of a Euclidean space $\Theta \subset \mathbb{R}^s$. The unfortunate abuse of notation caused by using $\boldsymbol{\pi}$ as both the parameterization (a function) as well as a generic element of the model (a vector) can be confusing, but it is standard. If the model is correctly specified, the true unknown distribution $\mathrm{Multinom}_r(1, \boldsymbol{\pi}^\star)$ corresponds to an unknown parameter $\boldsymbol{\theta}^\star \in \Theta$. Given a random sample $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$, the goal of estimation is therefore to select a plausible value $\boldsymbol{S}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}$ based on the data for the true unknown value $\boldsymbol{\theta}^\star$, where $\boldsymbol{S}$ represents some function known when $\mathcal{D}$ is known, a statistic. If $\widehat{\boldsymbol{\theta}} \in \widetilde{\Theta}$,

some roughly equivalent superset of Θ (e.g., its closure), $\boldsymbol{S}$ is properly called an estimator (Shao [2003], p. 122).

While not a technical specification, in both theory and practice parametric models are assumed to have a "nice" parameter space Θ. The field of categorical statistics presents a unique oddity in this way – all categorical models can naturally be parameterized by the probabilities of the configurations of the categories themselves by setting $\pi(\theta) = \theta$ and $\Theta = \mathcal{M}_{r-1}$. In other words, in categorical statistics the models themselves can act as the parameter space. While this recognition is interesting, it is of limited practical use. In virtually every applied situation the parameter space Θ is supposed to be contained in a lower dimensional Euclidean space $s < r$ and is a product of intervals. This is the context of loglinear models, for example.

The fundamental complication with conditional independence models is that while they are naturally technically parametric, they do not satisfy the conditions of classical theorems associated with parametric models. The same holds true for algebraic statistical models. For example, Birch's theorem (presented as Theorem 4.1) assumes a parametric model in which Θ contains an open set. However, the parameter space of the trivial parameterization of an algebraic statistical model (or conditional independence model) does not contain an open set in $\mathbb{R}^r$ since it is a subset of the $(r-1)$-dimensional simplex; that is, the dimensions do not align. Even when we eliminate the last variable, $\mathcal{M}_{r-1}^{-}$ does not contain an open set in $\mathbb{R}^{r-1}$ since it is an affine variety and therefore only contains an open set if it contains the entire affine space, i.e. the entire projected simplex. Thus, the models we deal with here are not amenable to the classical theorems which are used to demonstrate their asymptotic theory (for example).

We call parametric models with "nice" parameter spaces $\theta \mapsto \pi(\theta)$ models. While

the precise definition of "nice" is not important, the important thing to note is that conditional independence models are generally not $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ models. Rather, they are implicit statistical models, i.e. collections of probability vectors $\boldsymbol{\pi}$ satisfying a list of constraints. In some cases it is possible to find a parameterization $\boldsymbol{\pi}(\boldsymbol{\theta})$ of an implicit model $\mathcal{M}_{r-1}$ where $\boldsymbol{\Theta}$ does contain a nice open set in some lower dimensional space $\mathbb{R}^s$, and in those cases typically much of classical theory including that of exponential families applies. One example is provided by the introductory 2×2 independence model where $\mathcal{M}_{r-1}$ is, by definition, the collection of vectors in $\mathbb{R}^4$ which satisfy the independence condition in (2.33). It was seen that that collection constitutes a surface in $\mathbb{R}^4$ which can be recognized as the image of a map $\boldsymbol{\pi}$ from $\boldsymbol{\Theta} = [0,1]^2 \subset \mathbb{R}^2$ into $\mathbb{R}^4$ given by $[\theta_1\ \theta_2]' \overset{\boldsymbol{\pi}}{\mapsto} [(1-\theta_1)(1-\theta_2)\ (1-\theta_1)\theta_2\ \theta_1(1-\theta_2)\ \theta_1\theta_2]'$. Unfortunately, no such parameterization is available in general for either algebraic statistical models or conditional independence models.

On the other hand, since conditional independence models are technically parametric notions from parametric statistics are still well defined. For example an estimator is still a statistic $\boldsymbol{S}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}$ lying in the parameter space, but since the parameter space is the model itself $\widehat{\boldsymbol{\theta}}$ is more properly written $\widehat{\boldsymbol{\pi}}$. Moreover, the distinction between a statistic and an estimator is even more evident – an estimator is a statistic which attempts to approximate the true probability vector $\boldsymbol{\pi}^\star = \boldsymbol{\pi}(\boldsymbol{\theta}^\star) = \boldsymbol{\theta}^\star$ while confined to the statistical model, i.e. an estimate of the joint distribution which adheres to experimental assumptions. Thus, while $\widehat{\boldsymbol{\pi}}_{EMP}$ is an estimator for $\boldsymbol{\pi}^\star$ in the full model Δ_{r-1}, it is not an estimator for $\boldsymbol{\pi}^\star$ in any other model $\mathcal{M}_{r-1}$ as it does not conform to the experimental assumptions.

Of course, not just any estimator $\boldsymbol{S}$ will do. Ideally the goal is to find "optimal" estimators, estimators which exhibit certain intuitive refinements on what it means

for an estimator to be plausible. These refinements are called properties, an example of which is unbiasedness. However, properties are used primarily to compare estimators and as a consequence are only useful when estimators are already known. For example, since $\widehat{\boldsymbol{\pi}}_{EMP}$ has already been seen not to be an estimator in the model $\mathcal{M}_{r-1}$, no estimator of $\boldsymbol{\pi}^\star$ is known and thus the ability to compare estimators is not useful – comparison is only useful when a selection is available. Therefore when an estimation problem is presented the statistician first uses principles of estimation to discover estimators. Matching sample moments to theoretical moments, maximizing the likelihood function, and minimizing distances to the empirical distribution are all estimation procedures used to construct estimators. In this thesis we consider the estimator resulting from maximizing the likelihood of the data (which is in fact equivalent to minimizing a special distance) and those resulting from minimizing the distance from the empirical to the model for three statistically relevant distances. In particular, this thesis understands the term "minimum distance" as referring to this procedure using the following three distances familiar to the statistical community – the Euclidean distance (L_2), Pearson's chi-squared distance (X^2), and Neyman's modified chi-squared distance (X^2_N). While the properties of each of these have been discussed in the parametric $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ setting, no consensus has been reached as to which is preferable (Berkson [1980]). Our goal in this chapter is to demonstrate the asymptotic theorems concerning consistency and asymptotic normality for parametric $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ case still hold in a natural way for arbitrary algebraic statistical models and therefore also conditional independence models provided certain fairly minimal conditions.

The following is a brief outline of the current chapter. We begin in Section 4.1 with formal definitions of the estimators we will be considering – the maximum likelihood

estimator and the minimum distance estimators. In Section 4.2, we state various existence and uniqueness results for the estimators for conditional independence models; many of which are demonstrated for arbitrary algebraic statistical models which then hold for conditional independence models as a special case. In Section 4.3, we provide a discussion of the asymptotic theory regarding the various estimators.

4.1 Formal description of estimators considered

In this section we properly define and demonstrate equivalent formulations of the maximum likelihood estimators and minimum distance estimators which are the study of this work.

4.1.1 Maximum likelihood estimators

Stepping out of the multivariate categorical context for the moment, generally speaking the classical likelihood procedure concerns models of densities $\mathcal{M}_f$ indexed by some parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. In particular, the likelihood approach maintains that the ideal estimator for the parameter $\boldsymbol{\theta}$ is one which maximizes the likelihood function

$$L(\boldsymbol{\theta}) := f_{\boldsymbol{\theta}}(\mathcal{D}), \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{4.1}$$

where $\mathcal{D}$ again represents a sample of data. For example, if $\mathcal{D}$ is a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the density $f_{\boldsymbol{\theta}^\star}$, we have that

$$L(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\mathcal{D}) = f_{\boldsymbol{\theta}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \prod_{k=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{X}_k), \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{4.2}$$

This brings us to the formal definition of a maximum likelihood estimator.

DEFINITION 4.1 (MAXIMUM LIKELIHOOD ESTIMATOR) Let $\mathcal{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ be N independent and identically distributed copies of $\boldsymbol{X} \sim f_{\boldsymbol{\theta}^\star}$ and $\mathcal{M}_f$ a statistical

model of densities of $\boldsymbol{X}$ parameterized by a real vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. A maximum likelihood estimator (MLE) $\boldsymbol{S}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}_{MLE}$ for the parameter $\boldsymbol{\theta}$ is a statistic

$$\boldsymbol{S}(\mathcal{D}) \;=\; \widehat{\boldsymbol{\theta}}_{MLE} \;:=\; \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}) \;=\; \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\mathcal{D}). \tag{4.3}$$

Of course, such an estimator need not exist. The topic of the existence of the MLE in categorical models is considered in more detail in the next section.

Recall from Section 3.2 that the full model refers to the collection of multinomial distributions with no restrictions on the probability vector $\boldsymbol{\pi}$. In the multivariate categorical setting, the $\boldsymbol{Y}$ indicator notation of Section 3.2 is preferable to other representations since it fits into the conventional multinomial framework, and consequently we prefer dealing with the simplex representation of the full model (and other statistical models as well) instead of the density representation. In this context we can think of the MLE as maximizing a monomial over a subset of the probability simplex as follows. Suppose $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ are N independent and identically distributed copies of $\boldsymbol{Y} \sim \text{Multinom}_r(1, \boldsymbol{\pi}^\star)$. Then any statistic satisfying

$$\boldsymbol{S}(\mathcal{D}) \;=\; \widehat{\boldsymbol{\pi}}_{MLE} \;=\; \arg\max_{\boldsymbol{\pi}\in\mathcal{M}_{r-1}} \boldsymbol{\pi}^{\boldsymbol{T}} \tag{4.4}$$

is a maximum likelihood estimator for $\boldsymbol{\pi}^\star$ in the model $\mathcal{M}_{r-1}$, where $\boldsymbol{T} = \sum_{k=1}^n \boldsymbol{Y}_k$ and $\boldsymbol{\pi}^{\boldsymbol{T}} = \pi_1^{T_1} \cdots \pi_r^{T_r}$ is the multidegree notation.

We close this subsection by presenting the fact that the MLE, viewed from the geometric perspective, can be seen as minimizing a specific distance on the interior of the simplex between the empirical distribution and the model. It appears to have first been written about in Neyman [1949] following the underlying geometric perspective in Hotelling [1930] formalized by Doob [1934].

PROPOSITION 4.1 *Suppose* $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ *are* N *independent and identically distributed copies of* $\boldsymbol{Y} \sim \text{Multinom}_r(1, \boldsymbol{\pi}^\star)$, *and let* $\mathcal{M}_{r-1}$ *be a statistical model for*

$\pi^\star$. *If* $\mathcal{M}_{r-1} \subset int(\Delta_{r-1})$ *and* $\widehat{\pi}_{EMP} = \boldsymbol{T}/N$, *the maximum likelihood estimator* $\boldsymbol{S}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}_{MLE}$ *is equivalently expressed*

$$\boldsymbol{S}(\mathcal{D}) = \widehat{\pi}_{MLE} = \arg\max_{\pi \in \mathcal{M}_{r-1}} \pi^{\boldsymbol{T}} = \arg\min_{\pi \in \mathcal{M}_{r-1}} \sum_{k=1}^{r} \widehat{\pi}_k^{EMP} \log \frac{\widehat{\pi}_k^{EMP}}{\pi_k}, \tag{4.5}$$

where we can interpret $0 \log 0 = 0$ *and* $0 \log \frac{0}{0} = 0$.

Proposition 4.1 is a key realization. The point is that in categorical settings we can deal exclusively with the geometry of the model and the empirical relative frequencies to describe meaningful statistical quantities. In particular, for two distributions $\pi_1, \pi_2 \in \text{int}\,(\Delta_{r-1})$, we have that the MLE distance[1]

$$\delta_{MLE}(\pi_1, \pi_2) = \sum_{k=1}^{r} \pi_k^{(1)} \log \frac{\pi_k^{(1)}}{\pi_k^{(2)}} \tag{4.6}$$

is a statistically meaningful distance measure from the vector π_1 to the vector π_2. Note that it is not defined for $\pi_2 \in \partial\Delta_{r-1}$. $\delta_{MLE}(\pi_1, \pi_2)$ is commonly referred to as the Kullback-Leibler divergence in statistics and information science and is often written $KL(\pi_1 || \pi_2)$ (Kullback and Leibler [1951]). Of course, since it is not symmetric, it is not properly a metric (it is however always nonnegative). The above observation suggests the question – is $\delta_{MLE}(\pi_1, \pi_2)$ the only reasonable distance metric to use when selecting an estimator? The answer is of course no, which brings us to minimum distance estimators.

4.1.2 Minimum distance estimators

Minimum distance estimators were first thought of in the same categorical context as this thesis – that is, models for categorical experiments – in a remark in Neyman

[1] As in Section 3.2, when referring to elements of vectors which already have subscripts, the subscript moves to a superscript in parentheses.

[1949]. While the topic of the paper refers to the inferential side of statistics, the major thrust of the paper is finding necessary and sufficient conditions for estimators for $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ models to exhibit the same asymptotic properties of the maximum likelihood estimator, a class of estimators Neyman labeled best asymptotically normal (BAN) estimators. Neyman's view was that while maximum likelihood estimators have excellent asymptotic properties they are too difficult to compute (even in the $\boldsymbol{\theta} \mapsto \boldsymbol{\pi}(\boldsymbol{\theta})$ case). Inter alia, he found that while the MLE, minimum Pearson's chi-squared, and minimum Neyman modified chi-squared (defined there) are BAN, the L2E is not because while it is consistent and asymptotically normal, it fails to have minimum asymptotic variance. We now present the general framework for minimum distance estimators as presented in Bishop et al. [2007], pp. 502–508.

Let $\Delta_{r-1} \subset \mathbb{R}^r$ be the $(r-1)$–dimensional probability simplex, and let $\delta : \text{int}\,(\Delta_{r-1})^2 \to \mathbb{R}_{\geq 0}$ be a function such that

1. $\delta(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \geq 0$ and (4.7)
2. $\delta(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = 0$ if and only if $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ (4.8)

While δ is not formally a metric, we think of δ as the distance from $\boldsymbol{\pi}_1$ to $\boldsymbol{\pi}_2$. This immediately leads us the the definition of a minimum distance estimator.

DEFINITION 4.2 (MINIMUM DISTANCE ESTIMATOR) Let $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ be N independent and identically distributed copies of $\boldsymbol{Y} \sim \text{Multinom}_r(1, \boldsymbol{\pi}^\star)$, $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k$, and $\widehat{\boldsymbol{\pi}}_{EMP} = \boldsymbol{T}/N$; and let $\mathcal{M}_{r-1}$ be a statistical model. A minimum distance estimator $\boldsymbol{S}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}_{MDE}$ for $\boldsymbol{\pi}^\star$ is a statistic

$$\boldsymbol{S}(\mathcal{D}) \;=\; \widehat{\boldsymbol{\theta}}_{MDE} \;:=\; \arg \min_{\boldsymbol{\pi} \in \mathcal{M}_{r-1}} \delta\big(\widehat{\boldsymbol{\pi}}_{EMP}, \boldsymbol{\pi}\big). \tag{4.9}$$

We identify three distances for minimum distance estimation –

1. The L2E $(\widehat{\pi}_{L2E})^2$ –

$$\delta_{L2E}(\pi_1, \pi_2) = (\pi_1 - \pi_2)'(\pi_1 - \pi_2). \tag{4.10}$$

2. The minimum chi-squared (MCS, $\widehat{\pi}_{X^2}$) –

$$\delta_{X^2}(\pi_1, \pi_2) = (\pi_1 - \pi_2)'\boldsymbol{D}^{-1}(\pi_1 - \pi_2) \tag{4.11}$$

with $\boldsymbol{D} = \text{diag}(\pi_2)$.

3. The minimum Neyman-modified chi-squared (NCS, $\widehat{\pi}_{X^2_N}$) –

$$\delta_{X^2_N}(\pi_1, \pi_2) = (\pi_1 - \pi_2)'\boldsymbol{D}^{-1}(\pi_1 - \pi_2) \tag{4.12}$$

with $\boldsymbol{D} = \text{diag}(\pi_1)$.

Each of the above distances satisfies (4.7) and (4.8) for any π_1, $\pi_2 \in \text{int}\,(\Delta_{r-1})$; however, care must be taken on the boundary of the simplex. Pearson's chi-squared distance δ_{X^2} requires π_2 be positive in the same way that δ_{MLE} does. Experimentally this corresponds to all cells being possible; the distances are simply undefined on $\partial\Delta_{r-1}$. Neyman's modified chi-squared distance $\delta_{X^2_N}$ requires the first argument (i.e., π_1) to be positive which means that every cell must be observed ($\widehat{\pi}_{EMP} > \mathbf{0}_r$). By contrast, δ_{L2E} suffers none of these drawbacks.

On the other hand the statistical interpretation of the distances should be considered. While each of the distances is reasonable on the interior of the simplex, this is not the case on the boundary. Since samples cannot be observed in a distribution which attributes zero probability to them, it is unreasonable to assign an observed cell zero probability; however, none of the above distances demand this condition.

[2]The term "L2E" seems to be due to Scott [2001].

Both δ_{L2E} and $\delta_{X_N^2}$ include as a possibility estimators where there exists a $\pi_k^{(1)} > 0$ with $\pi_k^{(2)} = 0$, i.e. there is an observed cell which is attributed zero probability. Note that these do not occur for δ_{X^2} or δ_{MLE} because the distances are simply undefined on the "model side" of the simplex (i.e., in the model probability argument $\boldsymbol{\pi}_2$). In this work these two strange cases are assumed to be dealt with after the estimators are computed. `catcim`, the R package proposed to calculate these estimators in real-world problems, issues the user a warning when there are estimated cell probabilities which are zero for which the contingency table is nonzero.

4.2 Existence and uniqueness

In this section we discuss general results for the estimators for conditional independence models.

PROPOSITION 4.2 (EXISTENCE, KAHLE) *Let* $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ *be* N *independent and identically distributed copies of* $\boldsymbol{Y} \sim Multinom_r(1, \boldsymbol{\pi}^\star)$, $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k$, $\widehat{\boldsymbol{\pi}}_{EMP} = \boldsymbol{T}/N$, *and* $\mathcal{M}_{r-1}$ *be a statistical model compact in the usual metric topology of* $\mathbb{R}^r$. *Then if* $\delta(\widehat{\boldsymbol{\pi}}_{EMP}, \bullet)$ *is continuous on* $\mathcal{M}_{r-1}$, *a minimum distance estimator* $\widehat{\boldsymbol{\pi}}_\delta$ *exists.*

PROOF 4.1 The result is a direct consequence of the extreme value theorem.

□

COROLLARY 4.1 (EXISTENCE, KAHLE) $\widehat{\boldsymbol{\pi}}_{L2E}$ *exists for any nonempty algebraic statistical model* $\mathcal{M}_{r-1}$. $\widehat{\boldsymbol{\pi}}_{X_N^2}$ *exists for any nonempty algebraic statistical model when all cells are observed.* $\widehat{\boldsymbol{\pi}}_{X^2}$ *and* $\widehat{\boldsymbol{\pi}}_{MLE}$ *exist in any nonempty reduced algebraic statistical model* $\mathcal{M}_{r-1}^\epsilon = \{\boldsymbol{\pi} \in \mathcal{M}_{r-1} : \boldsymbol{\pi} \geq \epsilon \mathbf{1}_r\}$ *with* $\epsilon > 0$. *The results also hold for conditional independence models.*

PROOF 4.2 The key recognition is that algebraic statistical models are closed in the usual metric topology of $\mathbb{R}^r$. To see this, note that for any polynomial $h \in \mathbb{R}[\boldsymbol{\pi}]$, $V = V(h)$ is closed since $V = \{\boldsymbol{\pi} : \boldsymbol{\pi} \in h^{-1}(\{0\})\}$ is the inverse image of a closed set which is closed since h is continuous. Now, since algebraic statistical models are the intersection of such varieties and the probability simplex, they are a finite intersection of closed sets and are therefore closed. Being bounded on account of being subsets of the simplex, the Heine-Borel theorem implies that algebraic statistical models are compact. Thus, Proposition 4.2 applies since δ_{L2E} and $\delta_{X^2_N}$ are continuous for $\widehat{\boldsymbol{\pi}}_{EMP}$ fixed ($\widehat{\boldsymbol{\pi}}_{X^2_N}$ requiring in addition $\widehat{\boldsymbol{\pi}}_{EMP} > \mathbf{0}_r$).

For $\widehat{\boldsymbol{\pi}}_{X^2}$ and $\widehat{\boldsymbol{\pi}}_{MLE}$, note that the complications arise on the boundary of the simplex since the distances δ_{MLE} and δ_{X^2} are not defined there. If $\mathcal{M}_{r-1} \cap \partial\Delta_{r-1} = \emptyset$, existence follows from the previous argument. If the intersection is nonempty, the estimators obviously need not exist. However, if the model is reduced by requiring the probabilities be bounded away from the boundary of the simplex the preceding arguments apply to guarantee the existence of the estimators.

□

The previous theorem and corollary give sufficient conditions for the existence of a maximum likelihood estimator and minimum distance estimator which are of interest. Clearly such estimators need not be unique. The following extreme example demonstrates how this can happen for an algebraic statistical model.

EXAMPLE 4.1 *For a* 2×2 *contingency table consider the "spherical" algebraic statistical model defined by* $h(\boldsymbol{\pi}) = \left(\boldsymbol{\pi} - \mathbf{1}_4/4\right)'\left(\boldsymbol{\pi} - \mathbf{1}_4/4\right) - r^2$ *with* $0 < r < 1$, $\mathcal{M}_3 = V(h) \cap \Delta_3$. *Clearly if all cells are observed with equal counts then every* $\boldsymbol{\pi} \in \mathcal{M}_3$ *is a* $\widehat{\boldsymbol{\pi}}_{L2E}$ *and a* $\widehat{\boldsymbol{\pi}}_{X^2_N}$. *More generally, notice that* h *is symmetric in*

the π_k's. Thus, any distance δ which is also symmetric in the π_k's will not be unique when all cells are observed with equal counts. Since $\widehat{\boldsymbol{\pi}}_{MLE}$ and $\widehat{\boldsymbol{\pi}}_{X^2}$ are both symmetric in the π_k's, they are not unique; this fact is illustrated in Figure 4.1 for $\widehat{\boldsymbol{\pi}}_{X^2}$ with $r = \frac{1}{5}$, which can assume any permutation of the elements of $\left[\frac{1}{20}\left(5+2\sqrt{3}\right) \quad \frac{1}{60}\left(15-2\sqrt{3}\right) \quad \frac{1}{60}\left(15-2\sqrt{3}\right) \quad \frac{1}{60}\left(15-2\sqrt{3}\right)\right]'$.

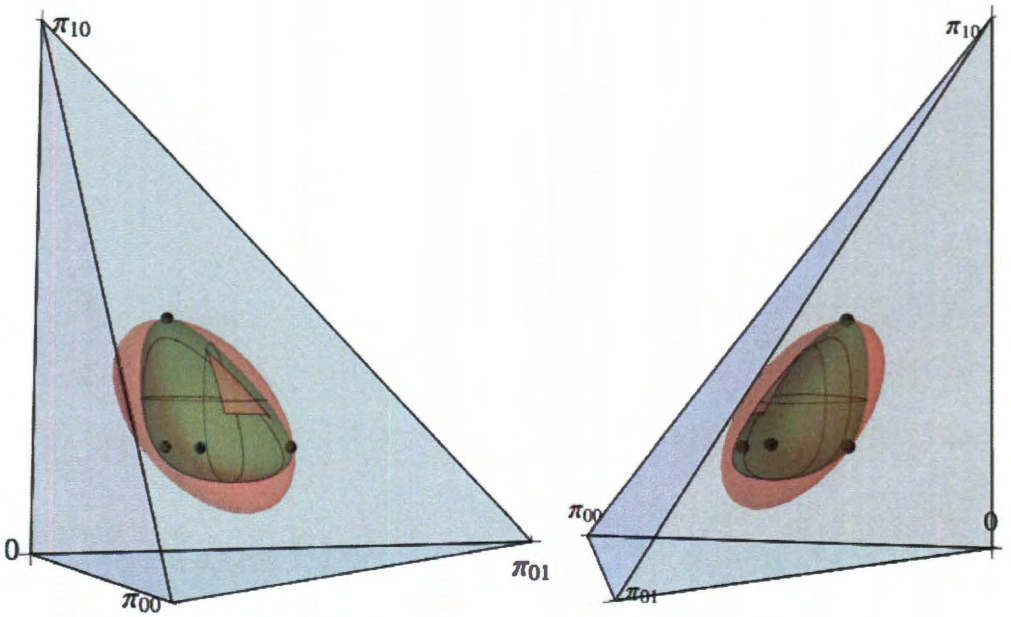

Figure 4.1 : The projected spherical 2×2 model $\mathcal{M}_3^-$ (red) along with the X^2 ball (green) of minimum distance from $\widehat{\boldsymbol{\pi}}_{EMP} = [.25\ .25\ .25\ .25]'$ (not shown) to the model. The black points represent the four possible values of $\widehat{\boldsymbol{\pi}}_{X^2}$. The triangular region in the middle of the green ball is an artifact of the plotting mechanism.

Thus, uniqueness is not generally obtained in algebraic statistical models. Demonstrating the uniqueness or non-uniqueness of the minimum distance estimators in arbitrary categorical conditional independence models seems to be a fairly difficult task. The following conjecture speaks to the next best thing – as sample sizes increase without bound, the estimators eventually become unique. A proof is left for future

work.

CONJECTURE 4.1 (UNIQUENESS, KAHLE) *Let* $\mathcal{D} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ *be* N *independent and identically distributed copies of* $\boldsymbol{Y} \sim Multinom_r(1, \boldsymbol{\pi}^\star)$ *and* $\mathcal{M}_{r-1}$ *a nonempty algebraic statistical model for* $\boldsymbol{Y}$ *with associated defining variety* $V(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$. *If* $\boldsymbol{\pi}^\star > \mathbf{0}_r$ *is a regular point*[3] *of* $V(\boldsymbol{h})$, *then* $\widehat{\boldsymbol{\pi}}_{L2E}$, $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, *and* $\widehat{\boldsymbol{\pi}}_{X^2_N}$ *are each unique with probability converging to one,*

$$\lim_{N\to\infty} P\left[\widehat{\boldsymbol{\pi}}_\delta \text{ is unique}\right] = 1. \tag{4.13}$$

4.3 Asymptotic results

4.3.1 Basic facts

We begin with a review of basic and fundamental results known for the multinomial distribution which serve as the theoretical basis for this section. For what follows we assume that $\boldsymbol{\pi}^\star \in \Delta_{r-1}$ is the true distribution of a single observation.

LEMMA 4.1 (MEAN AND VARIANCE OF MULTINOM$_r(1, \boldsymbol{\pi}^\star)$)
Let $\boldsymbol{Y} \sim Multinom_r(1, \boldsymbol{\pi}^\star)$. *Then* $\boldsymbol{\mu}_{\boldsymbol{Y}} = \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\pi}^\star$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = Var[\boldsymbol{Y}] = diag(\boldsymbol{\pi}^\star) - \boldsymbol{\pi}^\star(\boldsymbol{\pi}^\star)'$, *where*

$$diag(\boldsymbol{\pi}^\star) - \boldsymbol{\pi}^\star(\boldsymbol{\pi}^\star)' = \begin{bmatrix} \pi_1^\star(1-\pi_1^\star) & -\pi_1^\star\pi_2^\star & \cdots & -\pi_1^\star\pi_r^\star \\ -\pi_2^\star\pi_1^\star & \pi_2^\star(1-\pi_2^\star) & \cdots & -\pi_2^\star\pi_r^\star \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_r^\star\pi_1^\star & -\pi_r^\star\pi_2^\star & \cdots & \pi_r^\star(1-\pi_r^\star) \end{bmatrix}. \tag{4.14}$$

[3] A point $\boldsymbol{\pi} \subset \mathcal{M}_{r-1}$ is regular if the dimension of the tangent plane at $\boldsymbol{\pi}$ is the same as the dimension of the model $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}$ (see Section 4.3), $d = \dim T_{\boldsymbol{\pi}}(V) = \dim_{\boldsymbol{\pi}}(V)$.

Note that $\Sigma_{\boldsymbol{Y}}$ is singular since $\Sigma_{\boldsymbol{Y}} \mathbf{1}_r = \mathbf{0}_r$ (on account of $\mathbf{1}_r' \pi^\star = 1$). Since the variance-covariance matrix $\Sigma_{\boldsymbol{Y}}$ is so ubiquitous, we omit the subscript $_{\boldsymbol{Y}}$ which we use for other random vectors and simply refer to this variance-covariance matrix as Σ. Building on this result, we have the following generalization to an arbitrary finite sum of such random vectors.

LEMMA 4.2 (MEAN AND VARIANCE OF MULTINOM$_r(N, \pi^\star)$)
Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ be N independent and identically distributed copies of the categorical random vector $\boldsymbol{Y} \sim$ Multinom$_r(1, \pi^\star)$. Then the vector $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k \sim$ Multinom$_r(N, \pi^\star)$ by definition and has mean $\mu_{\boldsymbol{T}} = \mathbb{E}[\boldsymbol{T}] = N\pi^\star$ and variance-covariance matrix $\Sigma_{\boldsymbol{T}} = Var[\boldsymbol{T}] = N\Sigma$.

This applies almost trivially to yield the mean and variance of the empirical relative frequencies.

COROLLARY 4.2 (MEAN AND VARIANCE OF THE $\widehat{\pi}_{EMP}$) *Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ be N independent and identically distributed copies of $\boldsymbol{Y} \sim$ Multinom$_r(1, \pi^\star)$, $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k$, and $\widehat{\pi}_{EMP} = \boldsymbol{T}/N$. Then $\mathbb{E}[\widehat{\pi}_{EMP}] = \mu_{\widehat{\pi}_{EMP}} = \pi^\star$ and $Var[\widehat{\pi}_{EMP}] = \Sigma_{\widehat{\pi}_{EMP}} = \Sigma/N$.*

4.3.2 Laws of large numbers

In what follows we assume the general setting where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ are N independent and identically distributed copies of $\boldsymbol{Y} \sim$ Multinom$_r(1, \pi^\star)$, $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k$, $\widehat{\pi}_{EMP} = \boldsymbol{T}/N$, and each of $\widehat{\pi}_{L2E}$, $\widehat{\pi}_{X^2}$, $\widehat{\pi}_{X^2_N}$, and $\widehat{\pi}_{MLE}$ are defined as minimizing the distances in the previous subsection.

PROPOSITION 4.3 (STRONG CONSISTENCY OF $\widehat{\boldsymbol{\pi}}_{EMP}$) $\widehat{\boldsymbol{\pi}}_{EMP} \xrightarrow{a.s.} \boldsymbol{\pi}^\star$ *so that the relative frequencies converge almost surely to the true distribution. In other words,* $||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||_2 \to 0$ *as* $N \to \infty$ *with probability 1.*

PROOF 4.3 The result follows trivially from the strong law of large numbers since clearly $\mathbb{E}\left[||\boldsymbol{Y}||_2\right] = 1 < \infty$.

□

The same strong consistency is true for the L2E. While this result is almost surely known, no mention or proof of the fact could be found, so it is demonstrated here.

COROLLARY 4.3 (STRONG CONSISTENCY OF $\widehat{\boldsymbol{\pi}}_{L2E}$, KAHLE) *Let* $\mathcal{M}_{r-1} \subset \Delta_{r-1}$ *be a statistical model and suppose that* $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1} \subset \Delta_{r-1}$. *Then* $\widehat{\boldsymbol{\pi}}_{L2E} \xrightarrow{a.s.} \boldsymbol{\pi}^\star$ *so that* $\widehat{\boldsymbol{\pi}}_{L2E}$ *is strongly consistent for* $\boldsymbol{\pi}^\star$.

PROOF 4.4 By the triangle inequality,

$$||\widehat{\boldsymbol{\pi}}_{L2E} - \boldsymbol{\pi}^\star||_2 = ||\widehat{\boldsymbol{\pi}}_{L2E} - \widehat{\boldsymbol{\pi}}_{EMP} + \widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||_2 \tag{4.15}$$

$$\leq ||\widehat{\boldsymbol{\pi}}_{L2E} - \widehat{\boldsymbol{\pi}}_{EMP}||_2 + ||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||_2 . \tag{4.16}$$

Now, since $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$, $||\widehat{\boldsymbol{\pi}}_{L2E} - \widehat{\boldsymbol{\pi}}_{EMP}||_2 \leq ||\boldsymbol{\pi}^\star - \widehat{\boldsymbol{\pi}}_{EMP}||_2$ by the definition of $\widehat{\boldsymbol{\pi}}_{L2E}$ so that

$$||\widehat{\boldsymbol{\pi}}_{L2E} - \boldsymbol{\pi}^\star||_2 \leq 2\,||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||_2 \tag{4.17}$$

and the claim follows from Proposition 5.1.

□

Similar facts are known about the other estimators, but the demonstration is a bit more nuanced since the estimators do not exist in the same generality as the L2E. The following result appears in Rao [1965], pp. 291-294.

PROPOSITION 4.4 (STRONG CONSISTENCY OF $\widehat{\boldsymbol{\pi}}_{MLE}$) *Let $\mathcal{M}_{r-1} \subset \Delta_{r-1}$ be a statistical model and suppose that $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1} \subset \Delta_{r-1}$ with $\boldsymbol{\pi}^\star > \mathbf{0}_r$. For $\epsilon > 0$, define $A_\epsilon = \{\boldsymbol{\pi} \in \mathcal{M}_{r-1} : \delta_{MLE}(\boldsymbol{\pi}^\star, \boldsymbol{\pi}) \leq \epsilon\}$. Then if $A_\epsilon \in int(\mathcal{M}_{r-1})$ for some sufficiently small ϵ, as $N \to \infty$ $\widehat{\boldsymbol{\pi}}_{MLE}$ exists with probability one and $\widehat{\boldsymbol{\pi}}_{MLE} \xrightarrow{a.s.} \boldsymbol{\pi}^\star$ so that $\widehat{\boldsymbol{\pi}}_{MLE}$ is strongly consistent for $\boldsymbol{\pi}^\star$.*

We take for granted the same result for the minimum chi-squared distance and minimum Neyman modified chi-squared estimators suggested by Rao.

PROPOSITION 4.5 (STRONG CONSISTENCY OF $\widehat{\boldsymbol{\pi}}_{X^2}$, $\widehat{\boldsymbol{\pi}}_{X^2_N}$) *$\widehat{\boldsymbol{\pi}}_{X^2} \xrightarrow{a.s.} \boldsymbol{\pi}^\star$ and $\widehat{\boldsymbol{\pi}}_{X^2_N} \xrightarrow{a.s.} \boldsymbol{\pi}^\star$ provided the conditions of Proposition 4.4 replaced by δ_{X^2} and $\delta_{X^2_N}$.*

4.3.3 Central limit theorems

Since multinomial random vectors with parameter N are by definition the sum of N independent and identically distributed categorical random vectors, Lemma 4.2 is sufficient for the multivariate central limit theorem to apply from which we obtain the following results.

LEMMA 4.3 (ASYMPTOTIC NORMALITY OF MULTINOM$_r(N, \boldsymbol{\pi}^\star)$ AS $N \to \infty$)
Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ be N independent and identically distributed copies of the categorical random vector $\boldsymbol{Y} \sim$ Multinom$_r(1, \boldsymbol{\pi}^\star)$. Then the vector $\boldsymbol{T} = \sum_{k=1}^N \boldsymbol{Y}_k \sim$ Multinom$_r(N, \boldsymbol{\pi}^\star)$ exhibits the property that

$$N^{-1/2}(\boldsymbol{T} - N\boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r(\mathbf{0}_r, \boldsymbol{\Sigma}), \tag{4.18}$$

and consequently

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r(\mathbf{0}_r, \boldsymbol{\Sigma}). \tag{4.19}$$

The previous expression can be immediately applied to construct a hypothesis test for any statistical model. The intuition is that since $\widehat{\boldsymbol{\pi}}_{EMP}$ is asymptotically multivariate normal with mean $\boldsymbol{\pi}^\star$, an approximate confidence set for $\boldsymbol{\pi}^\star$ can be constructed as an ellipsoidal ball centered at $\widehat{\boldsymbol{\pi}}_{EMP}$ which can then be inverted to be used as a hypothesis test.

PROPOSITION 4.6 (TEST FOR STATISTICAL MODELS) *Let $\mathcal{M}_{r-1} \subset \Delta_{r-1}$ be any statistical model. For any $\boldsymbol{\pi} \in \mathcal{M}_{r-1}$, define*

$$\mathcal{C}_{\boldsymbol{\pi}^-} = \left\{ \boldsymbol{x} \in \mathbb{R}^{r-1} : (\boldsymbol{x} - \boldsymbol{\pi}^-)'(\boldsymbol{\Sigma}_{\boldsymbol{\pi}}^-/N)^{-1}(\boldsymbol{x} - \boldsymbol{\pi}^-) \leq c_{\alpha,\boldsymbol{\pi}^-} \right\}, \tag{4.20}$$

Then

$$T_N = \begin{cases} 1 & \mathcal{M}_{r-1}^- \cap \mathcal{C}_{\widehat{\boldsymbol{\pi}}_{EMP}^-} = \emptyset \\ 0 & otherwise \end{cases} \tag{4.21}$$

is an asymptotic level α test for the hypothesis $H_0 : \boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$ where the superscript $^-$ denotes removal of the rth element of a vector (vectors in a set, row and column of $\boldsymbol{\Sigma}_{\boldsymbol{\pi}}$) and $c_{\alpha,\boldsymbol{\pi}^-}$ is chosen such that $P\left[\boldsymbol{X} \in \mathcal{C}_{\widehat{\boldsymbol{\pi}}_{EMP}^-}\right] = 1 - \alpha$, where

$$\boldsymbol{X} \sim \mathcal{N}_{r-1}\left(\widehat{\boldsymbol{\pi}}_{EMP}^-, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\pi}}_{EMP}^-}/N\right). \tag{4.22}$$

Geometrically, the set $\mathcal{M}_{r-1}^- \cap \mathcal{C}_{\widehat{\boldsymbol{\pi}}_{EMP}^-}$ corresponds to the intersection of the model and an approximate level $1 - \alpha$ confidence set for $\boldsymbol{\pi}^\star$ (an ellipsoidal ball).

When the statistical model is an algebraic statistical model, this result can be characterized algebraically. We label such hypothesis tests – which reject/accept H_0 on the basis of an existing/nonexisting polynomial identity – algebraic hypothesis tests.

PROPOSITION 4.7 (TEST FOR ALGEBRAIC STATISTICAL MODELS, KAHLE)

Let $\mathcal{M}_{r-1} = V(\boldsymbol{h}) \cap \Delta_{r-1} = V\left(\boldsymbol{h}^{\Delta}\right) \cap \mathbb{R}^{r}_{\geq 0}$ *be an algebraic statistical model, set* $h(\boldsymbol{x}) = \boldsymbol{h}^{\Delta}(\boldsymbol{x})'\boldsymbol{h}^{\Delta}(\boldsymbol{x})$, *and let* T_N *be a function which assumes the value one if there exists a polynomial* $a \in \mathbb{R}[\boldsymbol{x}]$ *and sums of squares of polynomials* $g_{\nu} \in \sum(\mathbb{R}[\boldsymbol{x}])^2$ *such that*

$$ah + \sum_{\nu \in \{0,1\}^{r+1}} g_{\nu} x_1^{\nu_1} \cdots x_r^{\nu_r} e^{\nu_{r+1}} + 1 = 0 \tag{4.23}$$

where

$$e = c_{\alpha,\pi^-} - (\boldsymbol{x} - \widehat{\boldsymbol{\pi}}^{-}_{EMP})'(\boldsymbol{\Sigma}^{-}_{\widehat{\boldsymbol{\pi}}_{EMP}}/N)^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\pi}}^{-}_{EMP}), \tag{4.24}$$

and zero otherwise. Then T_N *is an asymptotic level* α *test for the hypothesis* H_0 : $\boldsymbol{\pi}^{\star} \in \mathcal{M}_{r-1}$.

PROOF 4.5 This follows directly from Propositions 3.2, 4.7, and the real Nullstellensatz (Theorem 3.1).

□

Before we proceed into the asymptotic theory of the estimators, it is helpful to take a moment to consider tangent spaces of affine varieties. Consider for a moment the unit sphere in $\mathbb{R}^3$, $V_1 = V(h_1(x,y,z)) = V(x^2 + y^2 + z^2 - 1)$. The tangent space of V_1 at a point $\boldsymbol{x}_0 = [x_0\ y_0\ z_0]'$ on V_1 is simply the plane running tangent to V_1 in $\mathbb{R}^3$ which is given implicitly as

$$T_{\boldsymbol{x}_0}(V_1) = \left\{\boldsymbol{x} \in \mathbb{R}^3 : \nabla h_1(\boldsymbol{x}_0)'(\boldsymbol{x} - \boldsymbol{x}_0) = 0\right\}, \tag{4.25}$$

which is the variety $V(\nabla h_1(\boldsymbol{x}_0)'(\boldsymbol{x} - \boldsymbol{x}_0))$. For example, at

$$\boldsymbol{x}_0 = \left[\sqrt{1/3}\ \sqrt{1/3}\ \sqrt{1/3}\right]' \in V_1, \tag{4.26}$$

we have

$$T_{\boldsymbol{x}_0}(V_1) = \left\{[x\ y\ z]' \in \mathbb{R}^3 : \frac{2}{\sqrt{3}}\left(x - \frac{1}{\sqrt{3}}\right) + \frac{2}{\sqrt{3}}\left(y - \frac{1}{\sqrt{3}}\right) + \frac{2}{\sqrt{3}}\left(z - \frac{1}{\sqrt{3}}\right) = 0\right\}, \tag{4.27}$$

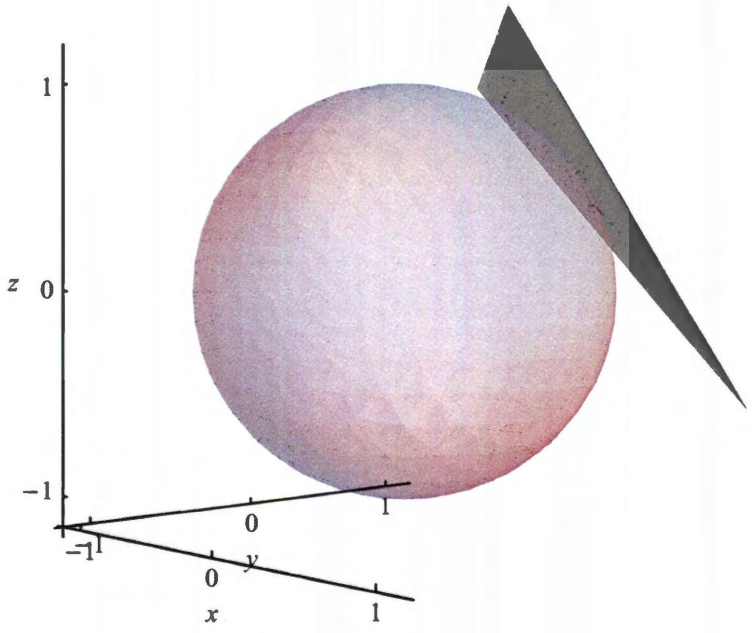

Figure 4.2 : V_1 (pink) and its tangent space at $\boldsymbol{x}_0$ (green)

and the mental image is Figure 4.2.

Now, suppose that we have a second variety $V_2 = V\left(h_2(x, y, z)\right) = V\left(x^2 + y^2 - z\right)$. Clearly we can compute the tangent space of V_2 at a point on V_2 in the same way as before, but consider the tangent space of the intersection $V = V_1 \cap V_2$ at a point $\boldsymbol{x}_0$. Obviously V is nonempty. In this case, the tangent space is

$$T_{\boldsymbol{x}_0}(V) = V\left(\boldsymbol{h}'(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)\right), \tag{4.28}$$

where $\boldsymbol{h}'(\boldsymbol{x}_0)$ is the Jacobian matrix evaluated at $\boldsymbol{x}_0$

$$\boldsymbol{h}'(\boldsymbol{x}_0) = \begin{bmatrix} - & \nabla h_1(\boldsymbol{x}) & - \\ - & \nabla h_2(\boldsymbol{x}) & - \end{bmatrix}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0}. \tag{4.29}$$

There are a number of ways to think about $T_{\boldsymbol{x}_0}(V)$. For example, it is every vector which is orthogonal to any vector normal to each of the varieties which intersect to

create V (based at $\boldsymbol{x}_0$). Equivalently it is any vector which when translated by $\boldsymbol{x}_0$ is in the null space of the Jacobian of $\boldsymbol{h} = [h_1\ h_2]'$ at $\boldsymbol{x}_0$. One useful way to think about $T_{\boldsymbol{x}_0}(V)$ is that it corresponds to the intersection of the individual tangent spaces of V at $\boldsymbol{x}_0$, namely $T_{\boldsymbol{x}_0}(V_1)$ and $T_{\boldsymbol{x}_0}(V_2)$. As an example, it can be checked that the point

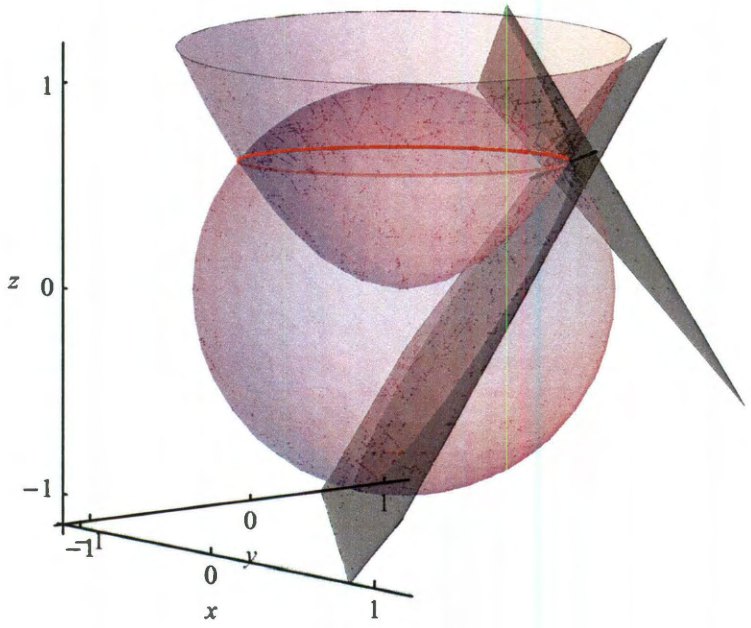

Figure 4.3 : $V = V_1 \cap V_2$ (in red) and its tangent space at $\boldsymbol{x}_0$ as the intersection of the individual tangent spaces (the black line)

$$\boldsymbol{x}_0 = \left[\frac{1}{2}\ \ \frac{\sqrt{2\sqrt{5}-3}}{2}\ \ \frac{\sqrt{5}-1}{2}\right]' \tag{4.30}$$

is in both V_1 and V_2. Thus, we calculate the tangent plane at the point $\boldsymbol{x}_0$ for both V_1 and V_2 and intersect these to determine the tangent space $T_{\boldsymbol{x}_0}(V)$. The mental image is Figure 4.3.

Thus, determining the tangent space of a variety at a point with nonzero Jacobian rows amounts to determining the intersection of planes, which is precisely the basic

question of interest in linear algebra. Suppose that we have still another variety defined by the polynomial $h_3(x, y, z) = -2h_2(x, y, z)$. Clearly $V(h_2) = V(h_3) = V_3$, and thus no additional information or reduction takes place : $V_1 \cap V_2 = V_1 \cap V_2 \cap V_3$. Therefore the tangent space of V_3 at a point $\boldsymbol{x}_0 \in V_3$ is precisely that V_2 at the same point, the row rank of $\boldsymbol{h}'(\boldsymbol{x}_0) = 2 < 3$ which we would expect if we had three "independent" planes, and the tangent space of V is as it was before. Following the fundamental fact from linear algebra that the row rank of a matrix equals its rank (that is column rank), the dimension of the tangent space is the dimension of the ambient space, three in this case for $\mathbb{R}^3$, minus the rank of the Jacobian matrix at $\boldsymbol{x}_0$. Thus, in this example the dimension of the tangent space of $V = V_1 \cap V_2 = V_1 \cap V_2 \cap V_3$ at $\boldsymbol{x}_0$ is one, which aligns with our intuition since we have already seen that the intersection is a line.

If V were a smooth manifold, this dimension would be constant across V. Unfortunately, varieties do not enjoy such regularity; however, they do exhibit the next best thing – a local smooth manifold structure away from certain bad points. A nonsingular (regular, manifold) point of a variety V is a point $\boldsymbol{x}_0 \in V$ at which the behavior seen in the preceding example is the whole story – the dimension of the variety at $\boldsymbol{x}_0$, a more complicated quantity, is equal to the dimension of the tangent space at $\boldsymbol{x}_0$, written $\dim_{\boldsymbol{x}_0}(V) = \dim T_{\boldsymbol{x}_0}(V)$. Technically speaking the dimension of a variety at a point is the dimension of the largest irreducible component of that variety which contains the point, but this technicality is not needed here. What is important to realize is that the dimension (in the loose or the technical sense) of a variety may change across the variety; at some points it may look like a line while at others a surface and at still others, a high dimensional volume. An example of this is the variety $V = V((x^2 + y^2)z) \subset \mathbb{R}^3$ which is the union of the xy plane and the z axis

which is two dimensional at some places and one dimensional at others. Fortunately, this changing of dimensions neither happens at nonsingular points nor near them in the sense that for any nonsingular point $\boldsymbol{x}_0$ of a given dimension there exists an open neighborhood of $\boldsymbol{x}_0$ $U \subset \mathbb{R}^r$ in the ambient Euclidean space such that for all points $\boldsymbol{x} \in V \cap U$ in the variety V and the neighborhood of $\boldsymbol{x}_0$, the dimension of the points and their tangent spaces are equal and constant across the set. In other words, if a point $\boldsymbol{x}_0$ is a regular point of a variety V, then all the nearby points on the variety are also regular points. Even more, it is a well established fact in algebraic geometry that the same neighborhood $V \cap U$ (away from these singular points the variety) is a C^∞ submanifold (Kendig [1977] chapters 2.3 and 4.1-3 or Bochnak et al. [1998], pp. 66-68). Thus, at a local level of a nonsingular point varieties are very well behaved objects.

The basic intuition underpinning the asymptotic theory of $\widehat{\boldsymbol{\pi}}_{L2E}$ is that near nonsingular points algebraic statistical models look like their tangent spaces – hyperplanes. While the explicit form for projecting onto the model is not known, the local smooth manifold structure of the model near the true value $\boldsymbol{\pi}^\star$ guarantees that it is well approximated by the projection onto the tangent space at $\boldsymbol{\pi}^\star$ which can be exploited to demonstrate that the asymptotic behavior of $\widehat{\boldsymbol{\pi}}_{L2E}$ is the same as that of $\widehat{\boldsymbol{\pi}}_{T_{\boldsymbol{\pi}^\star}(\mathcal{M}_{r-1})}$, the projection of the empirical relative frequencies onto the tangent space of $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}^\star$. Since the explicit form of an L_2 projection onto a hyperplane is well known, the asymptotic distribution of the latter is readily characterized by the delta method, granting along with it that of $\widehat{\boldsymbol{\pi}}_{L2E}$.

PROPOSITION 4.8 (ASYMPTOTIC NORMALITY OF $\widehat{\boldsymbol{\pi}}_{L2E}$, KAHLE) *Suppose that* $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$, *where* $\mathcal{M}_{r-1}$ *is a nonempty algebraic statistical model with defining polynomi-*

als $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$. *Then if* $\boldsymbol{\pi}^\star > \mathbf{0}_r$ *is an s-dimensional nonsingular point of* $\mathcal{M}_{r-1}$,

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_{L2E} - \boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{\Sigma}_{L2E}\right) \tag{4.31}$$

with $\boldsymbol{\Sigma}_{L2E}$ *as discussed in the proof. In particular,* $\widehat{\boldsymbol{\pi}}_{L2E}$ *is asymptotically normal.*

PROOF 4.6 For any subset $\mathcal{A} \subset \mathbb{R}^r$, let $\boldsymbol{g}_{\mathcal{A}} : \mathbb{R}^r \to \mathbb{R}^r$ be the map

$$\boldsymbol{x} \overset{g_{\mathcal{A}}}{\mapsto} \arg\min_{\boldsymbol{\pi} \in \mathcal{A}} ||\boldsymbol{x} - \boldsymbol{\pi}||_2 \tag{4.32}$$

and for ease of notation let $\boldsymbol{g}_{\mathcal{M}} = \boldsymbol{g}_{\mathcal{M}_{r-1}}$ and $\boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}} = \boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}(\mathcal{M}_{r-1})}$. Letting $\boldsymbol{h}^\Delta = [\boldsymbol{h}\ \mathbf{1}_r'\boldsymbol{\pi} - 1]'$ be the defining polynomials of $\mathcal{M}_{r-1}$ including the simplex polynomial, differentiability implies that for any point $\boldsymbol{x} \in \mathbb{R}^r$

$$\boldsymbol{h}^\Delta(\boldsymbol{x}) = \boldsymbol{h}^\Delta(\boldsymbol{\pi}^\star) + (\boldsymbol{h}^\Delta)'(\boldsymbol{\pi}^\star)(\boldsymbol{x} - \boldsymbol{\pi}^\star) + o(||\boldsymbol{x} - \boldsymbol{\pi}^\star||) \tag{4.33}$$

as $\boldsymbol{x} \to \boldsymbol{\pi}^\star$. Since $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$, $\boldsymbol{h}^\Delta(\boldsymbol{\pi}^\star) = \mathbf{0}_r$ so that

$$\boldsymbol{h}^\Delta(\boldsymbol{x}) - (\boldsymbol{h}^\Delta)'(\boldsymbol{\pi}^\star)(\boldsymbol{x} - \boldsymbol{\pi}^\star) = o(||\boldsymbol{x} - \boldsymbol{\pi}^\star||), \tag{4.34}$$

and thus $\boldsymbol{g}_{\mathcal{M}}(\boldsymbol{x}) = \boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{x}) + o(||\boldsymbol{x} - \boldsymbol{\pi}^\star||)$. In particular, at $\widehat{\boldsymbol{\pi}}_{EMP}$ this is

$$\boldsymbol{g}_{\mathcal{M}}(\widehat{\boldsymbol{\pi}}_{EMP}) - \boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\widehat{\boldsymbol{\pi}}_{EMP}) = o_{P_{\boldsymbol{\pi}^\star}}\left(||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||\right), \tag{4.35}$$

where the change to the stochastic order notation is justified since $||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star|| \xrightarrow{P_{\boldsymbol{\pi}^\star}} \mathbf{0}_r$.[4] Of course, by definition these are

$$\widehat{\boldsymbol{\pi}}_{L2E} - \widehat{\boldsymbol{\pi}}_{T_{\boldsymbol{\pi}^\star}} = o_{P_{\boldsymbol{\pi}^\star}}\left(||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||\right), \tag{4.36}$$

where $\widehat{\boldsymbol{\pi}}_{T_{\boldsymbol{\pi}^\star}}$ is the projection of the empirical relative frequencies onto the tangent space of $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}^\star$. Multiplying left and right by $\sqrt{N}$, we have

$$\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_{L2E} - \widehat{\boldsymbol{\pi}}_{T_{\boldsymbol{\pi}^\star}}\right) = o_{P_{\boldsymbol{\pi}^\star}}\left(\sqrt{N}||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||\right). \tag{4.37}$$

[4] Lemma 2.12 of van der Vaart [2000].

Since $\sqrt{N}(\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star)$ converges in probability, Prohorov's theorem applies to guarantee that $\sqrt{N}||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||$ is uniformly bounded in probability by some finite constant M. Therefore by the standard rules of the o_P and O_P calculus we have

$$o_{P_{\boldsymbol{\pi}^\star}}\left(\sqrt{N}||\widehat{\boldsymbol{\pi}}_{EMP} - \boldsymbol{\pi}^\star||\right) = o_{P_{\boldsymbol{\pi}^\star}}(O_{P_{\boldsymbol{\pi}^\star}}(M)) = o_{P_{\boldsymbol{\pi}^\star}}(1), \tag{4.38}$$

and we conclude that $\widehat{\boldsymbol{\pi}}_{L2E}$ and $\widehat{\boldsymbol{\pi}}_{T_{\boldsymbol{\pi}^\star}}$ have the same asymptotic distribution.

To obtain that distribution, we apply the multivariate delta method to the result of Lemma 4.3 to yield

$$\sqrt{N}\Big(\boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\widehat{\boldsymbol{\pi}}_{EMP}) - \boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{\pi}^\star)\Big) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{g}'_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{\pi}^\star)\boldsymbol{\Sigma}\boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{\pi}^\star)\right). \tag{4.39}$$

Now, since $\boldsymbol{\pi}^\star$ is a regular point of $\mathcal{M}_{r-1}$, linear algebra provides that for any $\boldsymbol{x} \in \mathbb{R}^r$

$$\boldsymbol{g}_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{x} - \boldsymbol{\pi}^\star) + \boldsymbol{\pi}^\star, \tag{4.40}$$

where $\boldsymbol{A} = \boldsymbol{A}_{\boldsymbol{\pi}^\star}$ denotes the $r \times s$ matrix whose columns span the tangent space of $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}^\star$. Differentiating with respect to $\boldsymbol{x}$ we have that

$$\boldsymbol{g}'_{T_{\boldsymbol{\pi}^\star}}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}' \tag{4.41}$$

so that $\widehat{\boldsymbol{\pi}}_{L2E}$ is asymptotically normal with mean $\boldsymbol{\pi}^\star$ and variance-covarance matrix

$$\boldsymbol{\Sigma}_{L2E} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'. \tag{4.42}$$

□

The asymptotic theory for the other estimators can be similarly characterized. It has already been noted that the difficulty with dealing with minimum distance estimators in conditional independence models (and more generally algebraic statistics models) is that the models are implicit by definition while the asymptotic theory

has only been laid out for models with parametric descriptions. The theorem we state shortly guarantees that if the true probability vector is a nonsingular point of $\mathcal{M}_{r-1}$, i.e. it is part of a nice piece of $\mathcal{M}_{r-1}$, then a Jacobian condition allows for the asymptotic theory known for parametric models to be ported over into that for algebraic statistical models. Moreover, that asymptotic theory is characterized by the regularity of the model at $\boldsymbol{\pi}^\star$.

It is well known in the parametric case that the three estimators $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, and $\widehat{\boldsymbol{\pi}}_{X^2_N}$ are asymptotically equivalent in the sense that they achieve the same asymptotic normal distribution (Neyman [1949]). The form of that distribution is provided by what is known as Birch's theorem (Birch [1964]). The form presented here follows Bishop et al. [2007]; this is one of the few places where the more appropriate notation $\boldsymbol{f}$ is used to denote the map from the parameter space Θ to the simplex Δ_{r-1} (or model $\mathcal{M}_{r-1}$) as opposed to $\boldsymbol{\pi}$ as discussed previously.

THEOREM 4.1 (BIRCH'S THEOREM)

Suppose $\mathcal{M}_{r-1} \subset \Delta_{r-1}$ is a parametric model given by a function $\boldsymbol{f} : \Theta \to \Delta_{r-1}$ where $\Theta \subset \mathbb{R}^s$ and that the true value $\boldsymbol{\pi}^\star = \boldsymbol{f}(\boldsymbol{\theta}^\star)$ for some $\boldsymbol{\theta}^\star$ (i.e. the model is correctly specified). Moreover, let the following six conditions be granted.

1. *The point $\boldsymbol{\theta}^\star$ is an interior point of Θ.*
2. *$\boldsymbol{\pi}^\star = \boldsymbol{f}(\boldsymbol{\theta}^\star) > \mathbf{0}_r$.*
3. *The map $\boldsymbol{f}$ is totally differentiable at $\boldsymbol{\theta}^\star$.*
4. *The Jacobian matrix $\boldsymbol{f}'$ is of full rank s at $\boldsymbol{\theta}^\star$ so that $\boldsymbol{f}$ maps a sufficiently small s-dimensional neighborhood about $\boldsymbol{\theta}^\star$ into an s-dimensional neighborhood of $\boldsymbol{\pi}^\star$ in $\mathcal{M}_{r-1}$.*

5. *The inverse mapping* $\boldsymbol{f}^{-1} : \mathcal{M}_{r-1} \to \Theta$ *is continuous at* $\boldsymbol{\pi}^\star$. *In particular, for every* $\epsilon > 0$ *there exists a* $\delta > 0$ *such that if* $||\boldsymbol{\theta} - \boldsymbol{\theta}^\star|| \geq \epsilon$, $||\boldsymbol{f}(\boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{\theta}^\star)|| \geq \delta$.

6. $\boldsymbol{f}$ *is continuous at every point* $\boldsymbol{\theta} \in \Theta$.

Then

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_\delta - \boldsymbol{\theta}^\star) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, (\boldsymbol{J}'\boldsymbol{D}^{-1}\boldsymbol{J})^{-1}\right) \tag{4.43}$$

and by the multivariate delta method

$$\sqrt{N}(\boldsymbol{f}(\widehat{\boldsymbol{\theta}}_\delta) - \boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{J}(\boldsymbol{J}'\boldsymbol{D}^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}'\right), \tag{4.44}$$

where $\boldsymbol{J} = \boldsymbol{f}'(\boldsymbol{\theta}^\star)$ *is the Jacobian of* $\boldsymbol{f}$ *at* $\boldsymbol{\theta}^\star$, $\boldsymbol{J} = \boldsymbol{f}'(\boldsymbol{\theta}^\star)$, $\boldsymbol{D} = diag(\boldsymbol{\pi}^\star)$.

There are several things to notice about Birch's theorem. First, since minimum distance estimators are functionally invariant (Drossos and Philippou [1980]), $\boldsymbol{f}(\widehat{\boldsymbol{\theta}}_\delta) = \widehat{\boldsymbol{\pi}}_\delta$ for any $\delta = \delta_{MLE}$, δ_{X^2}, or $\delta_{X^2_N}$, so that $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, and $\widehat{\boldsymbol{\pi}}_{X^2_N}$ are asymptotically normal. Second, the parameterization requires that the parameter space contains an open set. This is as previously noted not possible with the trivial parameterization, so Birch's theorem does not apply directly to implicitly defined models. Third, the condition of total differentiability is a smoothness condition similar to that of the tangent space for $\widehat{\boldsymbol{\pi}}_{L2E}$ – the smoothness of the function is the smoothness of the surface in the simplex. Next, the condition on the Jacobian enforces a kind of local consistency of dimension, something also discussed in relation to tangent spaces at nonsingular points. The condition on the inverse of the parameterization of course has no analogue as algebraic statistics models do not have a parameterization, so this condition requires some kind of explanation in any result involving the asymptotic distribution of the same estimators in any implicit case.

In the course of demonstrating the asymptotic behavior of $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, and $\widehat{\boldsymbol{\pi}}_{X^2_N}$, we will find the following lemma useful.

LEMMA 4.4 (KAHLE)

Let $\mathcal{M}_{r-1}$ be a statistical model with $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$, $U \subset \mathbb{R}^r$ an open set containing $\boldsymbol{\pi}^\star$, and set $\mathcal{M}'_{r-1} = \mathcal{M}_{r-1} \cap U$. If $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}}$ is a minimum distance estimator for $\boldsymbol{\pi}^\star$ assuming the model $\mathcal{M}'_{r-1}$ with a central limit theorem $r_N(\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}} - \boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r(\mathbf{0}_r, \boldsymbol{\Sigma}_\delta)$ for some $r_N \to \infty$, then $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}}$, the minimum distance estimator for $\boldsymbol{\pi}^\star$ in the larger model $\mathcal{M}_{r-1}$, enjoys the same central limit theorem.

PROOF 4.7 By Slutsky's lemma it suffices to show $r_N\left(\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}} - \widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}}\right) \xrightarrow{P_{\boldsymbol{\pi}^\star}} \mathbf{0}_r$. By Proposition 5.1 $\widehat{\boldsymbol{\pi}}_{EMP}$ is strongly consistent for $\boldsymbol{\pi}^\star$ and thus for every $\epsilon > 0$ there exists an $N' \in \mathbb{N}$ such that for all $N \geq N'$, $P_{\boldsymbol{\pi}^\star}\left[\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}} \in U\right] \geq 1 - \epsilon$. However, since $\widehat{\boldsymbol{\pi}}_{EMP}$ and δ are identical for both estimators, if $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}} \in U$ then $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}}$ exists and $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}}$ and $\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}}$ coincide so that $P_{\boldsymbol{\pi}^\star}\left[r_N \left|\left|\widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}_{r-1}} - \widehat{\boldsymbol{\pi}}_\delta^{\mathcal{M}'_{r-1}}\right|\right| = 0\right] \geq 1 - \epsilon$, which confirms the lemma.

□

In addition to the above lemma, the following fact from real algebraic geometry which formalizes intuition gained from the intersecting surfaces example above will be needed. The theorem as well as its proof are provided as Proposition 3.3.10 in Bochnak et al. [1998], pp. 67-68.

PROPOSITION 4.9 *Let $V \subset \mathbb{R}^r$ be an affine variety, not necessarily irreducible, and $\boldsymbol{x}$ a point of V. The following properties are equivalent –*

1. *There exists an irreducible component V' of V, with $\dim(V') = s$, such that V' is the only irreducible component of V containing $\boldsymbol{x}$ and $\boldsymbol{x}$ is a nonsingular point of V'.*

2. *There exist* $r - s$ *polynomials* $\boldsymbol{f} = [f_1, \ldots, f_{r-s}]' \in \mathbb{R}[\boldsymbol{x}]^{r-s}$ *with* $f_i \in I(V)$ *for each* $i = 1, \ldots, r - s$ *and an open neighborhood* U *of* $\boldsymbol{x}$ *in* $\mathbb{R}^r$ *for the Euclidean topology such that* $V \cap U = V(f_1, \ldots, f_{r-s}) \cap U$ *and the rank of the Jacobian matrix* $\boldsymbol{f}'(\boldsymbol{x})$ *is equal to* $r - s$.

The typical problem faced in conditional independence models is that there are many redundant polynomials in the list $\boldsymbol{h}$. An example of this was seen in the introductory 2×2 independence model, where $\boldsymbol{h}^{\Delta}$ consisted of the four polynomials provided by definition, namely those in (2.48)–(2.51), along with the simplex polynomial $\mathbf{1}_4'\boldsymbol{\pi} - 1$; however, it is well known that the independence condition is equivalent to the single equation $\pi_{00}\pi_{11} - \pi_{01}\pi_{10} = 0$. Thus, while the definition of independence communicates the statistical model as a variety defined by five polynomials, only two polynomials are needed, namely, $\pi_{00}\pi_{11} - \pi_{01}\pi_{10}$ and $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} - 1$.

COROLLARY 4.4 (KAHLE) *Suppose* $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1}$ *is a nonempty algebraic statistical model with defining polynomials* $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$. *Then if* $\boldsymbol{\pi}^\star > \mathbf{0}_r$ *is an* s*-dimensional nonsingular point of* $\mathcal{M}_{r-1}$, *there are* $r - s$ *polynomials* $\boldsymbol{g} \in \mathbb{R}[\boldsymbol{x}]^{r-s}$ *such that there exists an open neighborhood* $U \subset \mathbb{R}^r$ *of* $\boldsymbol{\pi}^\star$ *where the model and the variety of* $\boldsymbol{g}$ *coincide, i.e.* $\mathcal{M}_{r-1} \cap U = V(\boldsymbol{g}) \cap U$, *and the rank of the Jacobian* $\boldsymbol{g}'(\boldsymbol{\pi}^\star)$ *is* $r - s$. *Moreover, we can calculate* $\boldsymbol{g}$ *with a simple elimination procedure applied to the Jacobian* $\boldsymbol{h}'(\boldsymbol{\pi}^\star)$. *We call these polynomials definitive polynomials of* $\mathcal{M}_{r-1}$ *at* $\boldsymbol{\pi}^\star$; *they are not unique.*

PROOF 4.8 The only part of the Corollary that is not a direct consequence of Proposition 4.9 is how to compute $\boldsymbol{g}$, which is more a point of theoretical interest than of practical interest. Nevertheless, the example above concerning the intersecting surfaces describes how it is done. Let $\boldsymbol{h}^{\Delta} = [\mathbf{1}_r'\boldsymbol{\pi} - 1 \ \boldsymbol{h}']'$ be the collection of polynomials

whose variety when intersected with the positive orthant is $\mathcal{M}_{r-1}$. Then perform the following algorithm.

1: **input** the polynomials $\boldsymbol{h}$

2: **output** a collection of definitive polynomials $\boldsymbol{g}$ of $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}^\star$

3: Set $g_1(\boldsymbol{\pi}) \leftarrow \mathbf{1}_r'\boldsymbol{\pi} - 1$

4: Set $i \leftarrow 2$

5: **for** $j = 2$ to $j = k$ **do**

6: **if** $\Big(\text{Row } j \text{ of } \boldsymbol{h}'(\boldsymbol{\pi}^\star)\Big) \notin \text{RowSpace}\big(\boldsymbol{g}'(\boldsymbol{\pi}^\star)\big)$ **then**

7: $g_i \leftarrow \text{Row } j \text{ of } \boldsymbol{h}'(\boldsymbol{\pi})$

8: $i \leftarrow i + 1$

9: **end if**

10: **end for**

The resulting polynomials $\boldsymbol{g}$ obviously have the property that the rows of $\boldsymbol{g}'$ are be linearly independent at $\boldsymbol{\pi}^\star$, which guarantees the rank condition and confirms the claim.

□

We are now able to state the general result which allows us to use pre-existing parametric machinery in implicit models. The proof follows from an application of the implicit function theorem and the previous results.

THEOREM 4.2 (ASYMPTOTIC NORMALITY OF $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, AND $\widehat{\boldsymbol{\pi}}_{X^2_N}$, KAHLE) *Suppose $\boldsymbol{\pi}^\star \in \mathcal{M}_{r-1} = V(\boldsymbol{h}) \cap \Delta_{r-1}$ is a nonempty algebraic statistical model with defining polynomials $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$ and that $\boldsymbol{\pi}^\star > \mathbf{0}_r$ is an s-dimensional nonsingular point of $\mathcal{M}_{r-1}$. Let $\boldsymbol{g}$ be $r - s$ definitive polynomials of $\mathcal{M}_{r-1}$ at $\boldsymbol{\pi}^\star$ and consider the*

matrix partition of the Jacobian

$$\boldsymbol{g}'(\boldsymbol{\pi}^\star) = \begin{bmatrix} -- & \nabla g_1(\boldsymbol{\pi}^\star) & -- \\ & \vdots & \\ -- & \nabla g_{r-s}(\boldsymbol{\pi}^\star) & -- \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \end{bmatrix} \tag{4.45}$$

where $\boldsymbol{X} \in \mathbb{R}^{(r-s)\times s}$ *and* $\boldsymbol{Y} \in \mathbb{R}^{(r-s)\times(r-s)}$. *If* $\boldsymbol{Y}$ *is invertible, then* $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, *and* $\widehat{\boldsymbol{\pi}}_{X^2_N}$ *are asymptotically normal with central limit theorem*

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_\delta - \boldsymbol{\pi}^\star) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{J}(\boldsymbol{J}'\boldsymbol{D}^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}'\right), \tag{4.46}$$

where $\boldsymbol{J} = \boldsymbol{f}'(\boldsymbol{\theta}^\star)$ *(described in the proof),* $\boldsymbol{D} = diag(\boldsymbol{\pi}^\star)$, *and* δ *is any one of* δ_{MLE}, δ_{X^2}, *or* $\delta_{X^2_N}$.

PROOF 4.9 By Corollary 4.4, there exists an open set $U \subset \mathbb{R}^r$ containing $\boldsymbol{\pi}^\star$ such that $\mathcal{M}_{r-1} \cap U = V(\boldsymbol{g}) \cap U$. Since $\boldsymbol{\pi}^\star > \mathbf{0}_r$, this U can be chosen in the positive orthant of $\mathbb{R}^r$ so that $U \cap \mathbb{R}^r_{>0} = U$. Writing $\mathcal{M}_{r-1} \cap U = V(\boldsymbol{g}) \cap U$ differently, we have

$$\{\boldsymbol{\pi} \in U : \boldsymbol{\pi} \in \mathcal{M}_{r-1}\} = \{\boldsymbol{\pi} \in U : \boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}_{r-s}\}. \tag{4.47}$$

Now, since $\boldsymbol{Y}$ is invertible and $g_1, \ldots, g_{r-1}$ are infinitely differentiable, by the implicit function theorem there exist open sets $V \subset \mathbb{R}^s, U' \subset U$ and a C^∞ function $\boldsymbol{p} : V \to V'$ such that

$$\{\boldsymbol{\pi} \in U' : \boldsymbol{\pi} \in \mathcal{M}_{r-1}\} = \{\boldsymbol{\pi} \in U' : \boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}_{r-s}\} \tag{4.48}$$

$$= \left\{\left[\boldsymbol{\pi}'_{1:s} \;\; \boldsymbol{p}(\boldsymbol{\pi}_{1:s})'\right]' \in U' : \boldsymbol{\pi}_{1:s} \in V\right\}, \tag{4.49}$$

where $\boldsymbol{\pi}_{1:s}$ denotes the first s elements of $\boldsymbol{\pi}$. Noticing that this is a parameterization of $\mathcal{M}_{r-1} \cap U$ of the form $\boldsymbol{\theta} = \boldsymbol{\pi}_{1:s}$, $\Theta = V$, and $\boldsymbol{f} = [\boldsymbol{\theta}' \;\; \boldsymbol{p}(\boldsymbol{\theta})']'$,[5] we have

$$\mathcal{M}_{r-1} \cap U' = \{\boldsymbol{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}. \tag{4.50}$$

[5]We adopt here the $\boldsymbol{f}$ notation instead of the $\boldsymbol{\pi}$ notation to avoid confusion.

Setting $\mathcal{M}'_{r-1} = \mathcal{M}_{r-1} \cap U'$, Birch's theorem (Theorem 4.1) applies so that

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_{\delta}^{\mathcal{M}'_{r-1}} - \boldsymbol{\pi}^{\star}) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{J}(\boldsymbol{J}'\boldsymbol{D}^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}'\right), \tag{4.51}$$

where $\widehat{\boldsymbol{\pi}}_{\delta}^{\mathcal{M}'_{r-1}}$ is the minimum distance estimator in the model $\mathcal{M}'_{r-1}$, $\boldsymbol{J} = \boldsymbol{f}'(\boldsymbol{\theta}^{\star})$, $\boldsymbol{D} = \text{diag}(\boldsymbol{\pi}^{\star})$ and δ is any one of δ_{MLE}, δ_{X^2}, or $\delta_{X^2_N}$. By application of Lemma 4.4 we conclude

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_{\delta} - \boldsymbol{\pi}^{\star}) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}_r, \boldsymbol{J}(\boldsymbol{J}'\boldsymbol{D}^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}'\right), \tag{4.52}$$

i.e. that the result holds for the estimators constructed with the full model, which confirms the asymptotic normality of the estimators $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{X^2}$, and $\widehat{\boldsymbol{\pi}}_{X^2_N}$.

□

An example is helpful to understand how the theorem is to be understood.

EXAMPLE 4.2 *Let $\mathcal{M}_{r-1} = V(\boldsymbol{h}) \cap \Delta_{r-1}$ be the "spherical" 2×2 independence model with probabilities $\boldsymbol{\pi} = [\pi_{00}\ \pi_{01}\ \pi_{10}\ \pi_{11}]'$ and defining equations*

$$h_1(\boldsymbol{\pi}) = \pi_{00} - \pi_{0+}\pi_{+0} \tag{4.53}$$

$$h_2(\boldsymbol{\pi}) = \pi_{01} - \pi_{0+}\pi_{+1} \tag{4.54}$$

$$h_3(\boldsymbol{\pi}) = \pi_{10} - \pi_{1+}\pi_{+0} \tag{4.55}$$

$$h_4(\boldsymbol{\pi}) = \pi_{11} - \pi_{1+}\pi_{+1} \tag{4.56}$$

$$h_5(\boldsymbol{\pi}) = (\boldsymbol{\pi} - \mathbf{1}_4/4)'(\boldsymbol{\pi} - \mathbf{1}_4/4) - (2/10)^2. \tag{4.57}$$

We call this a spherical independence model since each of the distributions in the model $\mathcal{M}_{r-1}$ exhibit independence and *are each a fixed distance away from the uniform distribution. Geometrically, the model corresponds to the intersection of a sphere and the independence model in $\mathbb{R}^4$, seen projected in $\mathbb{R}^3$ in Figure 4.4. Recall that these*

"projections" are caused by the deletion of elements and are thus not projections in the usual sense. It is for this reason that the spheres do not appear spherical.

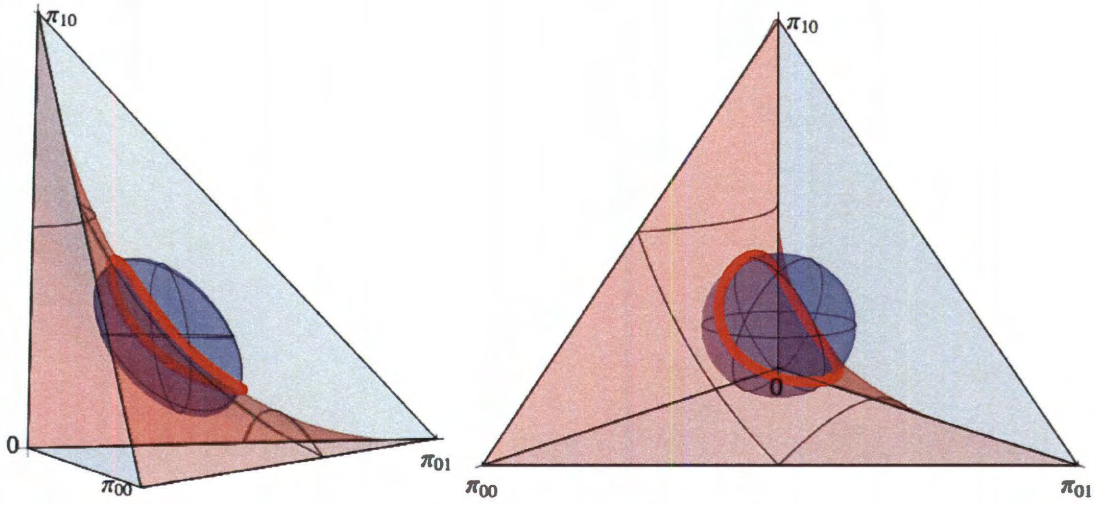

Figure 4.4 : The projected spherical 2×2 independence model $\mathcal{M}_{r-1}^{-}$ (in red) as the intersection of surfaces, seen from two directions

It can be checked that the point $\boldsymbol{\pi}^\star = [.35, .15, .35, .15]'$ *is in the model, i.e. on the red curve in Figure 4.4. Letting* $\boldsymbol{h}^\Delta(\boldsymbol{\pi}) = [\mathbf{1}_4'\boldsymbol{\pi} - 1 \;\; \boldsymbol{h}(\boldsymbol{\pi})']'$, *we have that the Jacobian of* $\boldsymbol{h}^\Delta$ *at* $\boldsymbol{\pi}^\star$ *is*

$$(\boldsymbol{h}^\Delta)'(\boldsymbol{\pi}^\star) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -\frac{1}{5} & -\frac{7}{10} & -\frac{1}{2} & 0 \\ -\frac{3}{10} & \frac{1}{5} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 & -\frac{1}{5} & -\frac{7}{10} \\ 0 & -\frac{1}{2} & -\frac{3}{10} & \frac{1}{5} \\ \frac{1}{5} & -\frac{1}{5} & \frac{1}{5} & -\frac{1}{5} \end{bmatrix}, \tag{4.58}$$

from which the procedure described in the proof of Corollary 4.4 indicates that only $g_1 = \mathbf{1}_4'\boldsymbol{\pi} - 1$, $g_2 = h_1$, *and* $g_3 = h_5$ *are needed to describe the model near* $\boldsymbol{\pi}^\star$. *Thus,*

$\boldsymbol{g} = [g_1\ g_2\ g_3]'$ *are the three definitive polynomials, and the Jacobian of* $\boldsymbol{g}$ *at* $\boldsymbol{\pi}^\star$ *is the matrix whose rows are the first, second, and last rows of the Jacobian of* $\boldsymbol{h}$ *above,*

$$\boldsymbol{g}'(\boldsymbol{\pi}^\star) = \left[\begin{array}{c|ccc} 1 & 1 & 1 & 1 \\ -\frac{1}{5} & -\frac{7}{10} & -\frac{1}{2} & 0 \\ \frac{1}{5} & -\frac{1}{5} & \frac{1}{5} & -\frac{1}{5} \end{array}\right]. \tag{4.59}$$

Since $\boldsymbol{g}$ *has three elements,* $\boldsymbol{X}$ *is the first column of* $\boldsymbol{g}'(\boldsymbol{\pi}^\star)$ *and* $\boldsymbol{Y}$ *is the last three columns. Since* $\boldsymbol{Y}$ *has a nonzero determinant, there exists open sets* $\Theta \subset \mathbb{R}$, $U \subset \mathbb{R}^4$ *and a function* $\boldsymbol{f} : \Theta \to U$ *such that* $\mathcal{M}_{r-1} \cap U = \{\boldsymbol{f}(\theta) : \theta \in \Theta\}$, *where* $\theta = \pi_{00}$ *but* $\Theta \neq [0, 1]$. *It can be checked that this function is*

$$\theta \overset{\boldsymbol{f}}{\mapsto} \left[\begin{array}{c} \theta \\ \frac{1}{40}\left(\sqrt{-400\theta^2+400\theta-42} - \sqrt{40\theta\left(\sqrt{-400\theta^2+400\theta-42}-20\right)+20\sqrt{-400\theta^2+400\theta-42}+58}-20\theta+10\right) \\ \frac{1}{40}\left(\sqrt{-400\theta^2+400\theta-42} + \sqrt{40\theta\left(\sqrt{-400\theta^2+400\theta-42}-20\right)+20\sqrt{-400\theta^2+400\theta-42}+58}-20\theta+10\right) \\ \frac{1}{20}\left(10-\sqrt{-400\theta^2+400\theta-42}\right) \end{array}\right] \tag{4.60}$$

and $\Theta \approx (.135, .405)$, *which contains* $\pi^\star_{00} = .35$ *(recall that the actual parameter here is* π_{00} *which we are simply rewriting as* θ*). By Theorem 4.8 and Theorem 4.2, any of the four minimum distance estimators are asymptotically normal for* $\boldsymbol{\pi}^\star$ *(and indeed any* $\boldsymbol{\pi} \in \mathcal{M}_{r-1}$*), and moreover their variance-covariance matrices are as specified in those theorems.*

||

Part II

Application

Chapter 5

Fitting

In Chapter 2 two separate examples were provided which involved actually calculating $\widehat{\boldsymbol{\pi}}_{MLE}$ and $\widehat{\boldsymbol{\pi}}_{L2E}$. While in both of those examples we were able to exploit a parameterization of the model, we noted that in general this method is not available and consequently considered different ways of computing the estimators. This chapter is devoted to introducing and better understanding this process for an arbitrary conditional independence model. Broadly speaking, two avenues are available to approach the problem of fitting. The first, more standard in statistics, is based on numerical routines. General nonlinear programming routines such as the limited memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton method with box constraints (L-BFGS-B) are available to compute all of the estimators considered in this work (Byrd et al. [1995]). However, these kinds of canned general-purpose routines fail to exploit the highly structured nature of the optimization problem at hand. Thus, for $\widehat{\boldsymbol{\pi}}_{L2E}$ we suggest a more tailored numerical method which exploits the polynomial structure to the full by using recently developed techniques from polynomial optimization. The second approach, distinct from the numerical routines in that it does not follow an iterative line search procedure, is based on an algebraic understanding of the underlying model and relies on an algebraic reformulation of the problem.

The current chapter proceeds as follows. In Section 5.1, the fitting problem is properly formulated as an optimization problem so that computing the estimator (fitting) is a clearly defined task. The primary difficulty with solving the problem

is also noted. Section 5.2 then discusses the numerical schemes, and Section 5.3 addresses the problem using Gröber basis methods. We conclude with a very brief discussion comparing the two methods.

5.1 The optimization problem

Understanding the formulation of the optimization problem is easiest in light of the example in Chapter 2. Thus, the reader is encouraged to revisit those examples as a springboard into this more general setting.

From Section 4.1, we know that each of the four estimators considered in this thesis has a representation in terms of a distance on the probability simplex. These distances are

$$\begin{aligned}
\delta_{L2E}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) &= (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)'(\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2) \\
\delta_{X^2}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) &= (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)'\mathrm{diag}(\boldsymbol{\pi}_2)^{-1}(\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2) \\
\delta_{X^2_N}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) &= (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)'\mathrm{diag}(\boldsymbol{\pi}_1)^{-1}(\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2) \\
\delta_{MLE}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) &= \sum_{k=1}^{r} \pi_k^{(1)} \log \frac{\pi_k^{(1)}}{\pi_k^{(2)}},
\end{aligned}$$

and we consider $\boldsymbol{\pi}_1$ to be the empirical distribution $\widehat{\boldsymbol{\pi}}_{EMP} = \boldsymbol{T}/N$ and $\boldsymbol{\pi}_2$ to be a candidate distribution in the model. As in the previous chapter, by definition the minimum δ-distance estimator of the conditional independence model with defining polynomials $\boldsymbol{h} \in \mathbb{R}[\boldsymbol{\pi}]^k$, $\mathcal{M}_{r-1} = V(\boldsymbol{h}) \cap \Delta_{r-1}$, is

$$\widehat{\boldsymbol{\pi}}_\delta = \arg \min_{\boldsymbol{\pi} \in \mathcal{M}_{r-1}} \delta(\widehat{\boldsymbol{\pi}}_{EMP}, \boldsymbol{\pi}). \tag{5.1}$$

We will of course assume such an estimator exists. Note that this equation generalizes (2.16) and (2.46) both in model and distance.

Also as before it will be convenient in various places to use the notation $\boldsymbol{h}^{\Delta} \in \mathbb{R}[\boldsymbol{x}]^{k+1}$ to denote the constraint polynomials along with the simplex polynomial $\mathbf{1}_r'\boldsymbol{\pi} - 1$, so that $\mathcal{M}_{r-1} = V(\boldsymbol{h}) \cap \Delta_{r-1} = V(\boldsymbol{h}^{\Delta}) \cap \mathbb{R}^r_{\geq 0}$, that is, the part of the variety of $\boldsymbol{h}^{\Delta}$ in the non-negative orthant.

Before we move on, locating the difficult part of the optimization problem helps in understanding its solution. The problem is not so much with the objective function δ (which is actually quite nice), but rather the feasibility region $\mathcal{M}_{r-1}$ – the model itself. It is the implicit nature of the problem which presents the difficulty.

5.2 Numerical schemes

In the Section 5.1 we identified the difficulty of the optimization problem as the feasibility region. To get around this, we use the method of Lagrange multipliers. Using Proposition 3.2, we can implement the Lagrange multiplier method in a number of ways. The most natural way to use Lagrange multipliers is simply to introduce one multiplier for each of the k conditional independence polynomials and one more for the simplex condition that the probabilities sum to unity, neglecting for the moment the positivity condition. Thus, if $\boldsymbol{h}^{\Delta}(\boldsymbol{\pi}) \in \mathbb{R}[\boldsymbol{\pi}]^{k+1}$ are the conditional independence polynomials including the simplex condition and $\boldsymbol{\lambda} \in \mathbb{R}^{k+1}$ the Lagrange multipliers, as in (2.17), (2.48)-(2.51), and (2.54), then the optimization problem of (5.1) is reduced to finding special critical points of

$$\Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \delta(\widehat{\boldsymbol{\pi}}_{EMP}, \boldsymbol{\pi}) + \boldsymbol{\lambda}'\boldsymbol{h}^{\Delta}(\boldsymbol{\pi}), \tag{5.2}$$

that is, solutions to

$$\nabla\Lambda(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \mathbf{0}. \tag{5.3}$$

Now, (5.3) involves $r+k+1$ variables in $r+k+1$ unknowns. Using Proposition 3.2, we could reduce this to using one h and one multiplier λ, so that (5.2) becomes

$$\begin{aligned} \Lambda(\boldsymbol{\pi}, \lambda) &= \delta(\widehat{\boldsymbol{\pi}}_{EMP}, \boldsymbol{\pi}) + \lambda \boldsymbol{h}^{\Delta}(\boldsymbol{\pi})' \boldsymbol{h}^{\Delta}(\boldsymbol{\pi}) && (5.4) \\ &= \delta(\widehat{\boldsymbol{\pi}}_{EMP}, \boldsymbol{\pi}) + \lambda h(\boldsymbol{\pi}), && (5.5) \end{aligned}$$

and (5.3),

$$\nabla \Lambda(\boldsymbol{\pi}, \lambda) = 0, \tag{5.6}$$

$r+1$ equations in $r+1$ unknowns. For example, in the binomial example this would change (2.17) to

$$\begin{aligned} \Lambda(\boldsymbol{\pi}, \lambda) &= ||\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}_n||_2^2 + \lambda h(\boldsymbol{\pi}) && (5.7) \\ &= (\pi_0 - \widehat{\pi}_0^{EMP})^2 + (\pi_1 - \widehat{\pi}_1^{EMP})^2 + (\pi_2 - \widehat{\pi}_2^{EMP})^2 && (5.8) \\ &\quad + \lambda\Big(\left(\pi_1^2 - 4\pi_2 + 4\pi_1\pi_2 + 4\pi_2^2\right)^2 \\ &\qquad + (\pi_0 + \pi_1 + \pi_2 - 1)^2 \Big). \end{aligned}$$

Alternatively, we could group together sets of the h_k's and allot to the sum of squares of each group a multiplier. The trade off between grouping all of the constraints into one and using a multiplier for each individually is that the constraint function h resulting from the grouping is more complicated than any of the h_k's. In fact if we use the sums of squares construction h is quartic for conditional independence models, because each of the h_k's is quadratic and the simplex is linear, which is not preferable since the closer to affine the varieties are the easier it is to solve them numerically. Moreover, h is nonnegative by design. That being said, there is information in the Lagrange multipliers concerning the optimization problem at hand. While we do not investigate this information, in this thesis we opt to use the full Lagrangian (i.e. the

one with one multiplier for each condition) and leave the other techniques for future investigation.

Calculating $\widehat{\pi}_{L2E}$, $\widehat{\pi}_{MLE}$, $\widehat{\pi}_{X^2}$, and $\widehat{\pi}_{X^2_n}$ numerically

Generally speaking, in principle a solution for the above optimization problem can be obtained via any nonlinear solver. Thus, when fitting any of the estimators any of such solver will do; we use the limited memory BFGS quasi-Newton method with box constraints (Byrd et al. [1995]). In practical situations, this method has performed satisfactorily for small scale problems (two and three dimensional tables) with computations on the order of seconds, with a typical 2x2x2 model taking about half a second for $\widehat{\pi}_{L2E}$.

For $\widehat{\pi}_{L2E}$ we suggest a more sophisticated method which try to take advantage of the highly structured nature of the optimization problem at hand. It is to this method which we now turn.

Calculating $\widehat{\pi}_{L2E}$ via Sums of Squares (SOS) relaxations

Computing $\widehat{\pi}_{L2E}$ is a particularly interesting problem because in this case the optimization problem (5.1) has as its objective function a polynomial and as its feasibility region a basic closed semialgebraic set.[1] A basic closed semialgebraic set is a set $\mathcal{K} = \{\boldsymbol{x} \in \mathbb{R}^n : h_1(\boldsymbol{x}) \geq 0, \ldots, h_k(\boldsymbol{x}) \geq 0\}$, where $h_i \in \mathbb{R}[\boldsymbol{x}]$ for $i = 1, \ldots, k$. Such problems are called polynomial optimization problems. In recent years it has been observed that while such polynomial optimization problems are NP-hard in general (i.e. at least as hard as any NP problem), approximations to the problem can be generated by solving a hierarchy of relaxations which are semidefinite programs which

[1] $\widehat{\pi}_{X^2}$ exhibits this same property but is not defined for all tables $\boldsymbol{T}$.

are convex and can be solved efficiently with interior point methods, for example. Excellent at-length surveys of polynomial optimization can be found in Sturmfels [2002] and Laurent [2009] from which the current exposition is heavily based in addition to Laurent [2003].

The relaxations to the problem involve sums of squares (SOS) of polynomials with the basic idea being as follows. In the unconstrained case, the polynomial programming problem is

$$p_{min} = \inf_{\boldsymbol{x}} p(\boldsymbol{x}), \tag{5.9}$$

where $p \in \mathbb{R}[\boldsymbol{x}]$ is a polynomial of even degree $2d$, $d \in \mathbb{N}$. This is equivalently expressed

$$p_{min} = \sup_{\mathcal{L}} \lambda \quad \text{with} \quad \mathcal{L} = \{\lambda : p(\boldsymbol{x}) - \lambda \geq 0, \boldsymbol{x} \in \mathbb{R}^n\}. \tag{5.10}$$

The problem of determining whether a polynomial is positive on all of $\mathbb{R}^n$ is known to be NP-hard (Laurent [2009]); however, the relaxation

$$p_{min} = \sup_{\mathcal{L}} \lambda \quad \text{with} \quad \mathcal{L} = \{\lambda : p(\boldsymbol{x}) - \lambda \text{ is SOS}\} \tag{5.11}$$

is based on determining whether a polynomial is a sum of squares of polynomials on $\mathbb{R}^n$, which can be solved via semidefinite programming using the following fact from polynomial optimization.

PROPOSITION 5.1 *The following are equivalent, throughout the notation* $|\boldsymbol{v}|$ *denotes* $\mathbf{1}'\boldsymbol{v}$.

1. $p(\boldsymbol{x})$ *is a sum of squares of polynomials (SOS).*

2. $p(\boldsymbol{x}) = \boldsymbol{z}'\boldsymbol{X}\boldsymbol{z}$ *for some positive semidefinite matrix* $\boldsymbol{X} = \left[X_{\boldsymbol{\beta},\boldsymbol{\gamma}}\right]_{|\boldsymbol{\beta}|,|\boldsymbol{\gamma}| \leq d}$, *where* $\boldsymbol{z} = \left[\boldsymbol{x}^{\boldsymbol{\beta}}\right]_{|\boldsymbol{\beta}| \leq d}$ *is the vector of monomials.*

3. *The semidefinite program*

$$\left\{ \begin{array}{c} \boldsymbol{X} \succeq 0 \\ \sum_{\boldsymbol{\beta},\boldsymbol{\gamma} \in S_n^d | \boldsymbol{\beta}+\boldsymbol{\gamma}=\boldsymbol{\alpha}} X_{\boldsymbol{\beta},\boldsymbol{\gamma}} = p_{\boldsymbol{\alpha}} \end{array} \right\} \tag{5.12}$$

is feasible, where $S_n^d = \left\{ \boldsymbol{\alpha} \in \mathbb{Z}_{\geq 0}^n : |\boldsymbol{\alpha}| \leq d \right\}$ *and* $p_{\boldsymbol{\alpha}}$ *is the coefficient of the monomial* $\boldsymbol{x}^{\boldsymbol{\alpha}}$ *in* $p(\boldsymbol{x})$.

For the constrained case, the optimization problem is

$$p_{min} = \inf_{\boldsymbol{x} \in \mathcal{K}} p(\boldsymbol{x}), \tag{5.13}$$

which is equivalent to

$$p_{min} = \sup_{\mathcal{L}} \lambda \quad \text{with} \quad \mathcal{L} = \{ \lambda : p(\boldsymbol{x}) - \lambda \geq 0, \boldsymbol{x} \in \mathcal{K} \}. \tag{5.14}$$

Again the problem is relaxed; lower bounds on p_{min} are obtained by

$$p_t = \sup_{\mathcal{L}} \lambda \quad \text{with} \quad \mathcal{L} = \left\{ \lambda : p(\boldsymbol{x}) - \lambda = \sum_{j=0}^{k} h_j(\boldsymbol{x}) p_j(\boldsymbol{x}) \right\}, \tag{5.15}$$

where $h_0(\boldsymbol{x}) := 1$, h_j is SOS for $j = 1, \ldots, k$, and the total degree of each of the polynomials $h_j(\boldsymbol{x})p_j(\boldsymbol{x})$ is less than or equal to $2t$.

5.3 Algebraic methods

The algebraic procedure recognizes (5.3) and (5.6) as systems of nonlinear polynomial equations in the π_k's and the λ_k's. Thus, any of the techniques from the algebraic community used to solve such problems are applicable. A survey of the modern techniques can be found in Sturmfels [2002] and Sommese et al. [2005], the former of which focuses on Gröbner type techniques, and the latter focusing on homotopies, i.e., tracing solutions from simpler systems back to the more complicated one presented.

This presentation considers only the basic notions of the first, leaving a more detailed development of the Gröbner technique and entire the homotopy technique for future inquiry.

As seen in Chapter 2, the basic notion of the Gröbner technique is as follows – compute a Gröbner basis for the ideal generated by $\boldsymbol{h}^{\Delta}$ with respect to a monomial order of interest, and then use the solve the ensuing system via univariate root solvers and back substitution. It may not always be the case that the problem is quite as clean as in the 2×2 example; however, the problem will be considerably easier. In many cases a similar structure will unfold due to the elimination theory of such bases (see, e.g., Cox et al. [2007]). Note that in this technique all of the solutions are determined as being the zero points of the gradient, and thus global solutions are obtained when available.

5.4 Comparing fitting methods

There are advantages and disadvantages to each of the methods described. For the standard nonlinear solver, while the routine will produce an output there is no guarantee that the optimal value will be found, if one exists. However, while the method is uncertain it is very simple to implement and since the surface is not believed to be very bumpy due to the geometric structures under consideration. Thus, such solvers provide a handy and useful basic routine which can be improved upon. The sums of squares relaxations method is vastly superior in terms of theoretical soundness, but its implementation is more complex. The Gröbner bases technique provides the most elegant solution to the problem; however, the devil is in computing the Gröbner basis itself, a process which is known to be in EXP-space via Buchberger's algorthm (Mittmann [2007]). In the smallest of circumstances, this method is actually very

feasible using efficient implementations of any Gröbner basis algorithm and well chosen monomial orders (and perhaps a Gröbner walk algorithm), and thus it might be preferable to the others.

Chapter 6

Implementation - the `catcim` and `mpoly` R packages

As noted in previous chapters, some of the fitting techniques are non-standard in statistics. In addition, since conditional independence models are defined implicitly and not parametrically with a particular functional form, it may be difficult for the non-expert to even determine a workable representation of the model in any computational environment. The `catcim` R package is here proposed to alleviate these difficulties by providing users with a sophisticated yet user friendly collection of functions to calculate estimators in an arbitrarily specified conditional independence model.

One problem which arises in dealing with conditional independence models in R is that of symbolic computing. Generally speaking, R is designed for numerical computing and has no real mechanism in either base R or any of its packages which satisfactorily performs symbolic computing with polynomials. The novel `mpoly` package presented below attempts to alleviate this problem by providing a simple yet efficient framework with which to perform symbolic computing with polynomials in R. Since `mpoly` forms the basis of `catcim`, we begin by describing `mpoly`.

6.1 `mpoly`

The `mpoly` package, written as a novel contribution of this thesis, is a general purpose collection of tools for symbolic computing with polynomials in R. Its utility is therefore obvious; however, it is not the first package proposed to deal with multivariate

polynomials. We begin with a discussion of the current package intended to fulfill this need, `multipol`, and highlight its limitations in order to motivate the need for `mpoly`.

6.1.1 The `multipol` package

R currently has three packages which deal with polynomials – `polynom`, `PolynomF`, and `multipol`. The first two, the second being a reboot of the first, deal exclusively with univariate polynomials and are thus not suited to our needs (Venables et al. [2009], Venables [2010]).

`multipol` is an implementation of multivariate polynomials which uses `arrays` as its underlying data structure, defining an S3 class object called `multipol` (Hankin [2009], Hankin [2008]). Thus, under the hood an object of class `multipol` is simply an *unnamed* multidimensional array.

```
> a <- as.multipol(array(1:12,c(2,3,2)))
> a
, , z^0

    y^0 y^1 y^2
x^0   1   3   5
x^1   2   4   6

, , z^1

    y^0 y^1 y^2
x^0   7   9  11
```

```
x^1   8  10  12
```

Note the manner in which the multivariate polynomial is specified. After a moment with pencil and paper, it is evident that the polynomial being described is

$$1 + 2x + 3y + 4xy + 5y^2 + 6xy^2 + 7z + 8xz + 9yz + 10xyz + 11y^2z + 12xy^2z \quad (6.1)$$

in the ring $\mathbb{R}[x, y, z]$. (A similar pretty printing can be enabled by setting `options(showchars = TRUE)`.) There are at least three different problems or serious inconveniences with this method of implementation which together call for a fresh reenvisioning of multivariate polynomials in R.

1. There are some similarities between representing a polynomial as an `array` and representing a contingency table as an `array` as discussed in the `catcim` package section. First, at the outset they both seem like a very reasonable thing to do. While it takes a bit to translate the `array` format of a polynomial to the common inline format, it is nevertheless obvious how one goes about doing it. However, `array`s are awkwardly handled in R, sufficiently so in fact to require an entire package just to perform various common tasks with them (the `aperm` package). Consequently, the data structure itself can make the simple act of defining a polynomial tedious and even challenging.

   Unfortunately, even if we accept the `array` scheme, the current implementation lacks basic functionality. The general constructor for an object of class `multipol` is the function `multipol` (or `as.multipol`); however, this is just a wrapper to force an `array` object's class to be `multipol`, without any checking of the underlying data structure. Thus, the coefficients can be any R data type – logicals, characters, numerics, etc., even raw bytes or missing altogether. This

is a common criticism of the use of S3 classes which when done correctly requires data checks at the function level which the `multipol` constructors do not perform. To complicate matters further, the general R function which provides the user with the data structure of the object, `str`, throws an exception.

```
> str(a)
Error in `[.multipol`(object, seq_len(ile)) :
  incorrect number of dimensions
```

While the motivation behind this problem is reasonable (subseting with indices which are zero to refer to terms with zero degree indeterminates), the consequence is a very user unfriendly framework.

2. Notice that the definition of a multivariate polynomial in `multipol` makes no reference to the variable names `x` and `y`. Surprisingly, they are defaulted *when the* `print` *method is called.* Thus a `multipol` object has no notion of variables whatsoever, the variables are only generated when asked for via `print.multipol`. An obvious problem in and of itself, this issue creates a much more serious restriction of the package's practical value – the inability to do polynomial arithmetic with two polynomials with a different number of variables (due to a lack of variable specification at the time of definition). As far as R is concerned, the polynomials can have many variables, but neither the names of the variables nor their order can be recorded, so arithmetic is only *possible* if two `multipol` objects have the same number of dimensions (variables), and only *meaningful* if those dimensions represent the same variables in the same order, the latter of which is never checked because there are no names specified which can be checked. Thus, when performing arithmetic, the onus is

on the user to ensure that the arithmetic is possible and meaningful. Moreover, `multipol` does not issue a warning message to alert the user to potential inappropriate use. These issues are demonstrated in the following examples, the first of which shows the inability to do arithmetic with two polynomials with a differing number of variables, and the second which shows how the arithmetic can be wrong (or at least not what the user intended).

Suppose one wants to add or multiply the polynomials

$$a = 1 + 2x + 3y + 4xy \tag{6.2}$$

and

$$b = 1 + 2x + 3y + 4xy + 5z + 6xz + 7yz + 8xyz, \tag{6.3}$$

both in $\mathbb{R}[x, y, z]$. Mathematically speaking, it is obvious that the sums and products are well-defined. However, they cannot be calculated with `multipol`

—

```
> a <- multipol( array(1:4,c(2,2)) )
> b <- multipol( array(1:8,c(2,2,2)) )
> a + b
Error: length(dima) == length(dimb) is not TRUE
> a * b
Error: length(dim(a)) == length(dim(b)) is not TRUE
```

Again, these errors are caused by an oversimplistic implementation. While this is not an inherent problem with the `array` data structure, it is a problem with the current implementation of multivariate polynomials in R.

More basically, suppose that one wants to add the polynomial $a = x$ to $b = 1 + y$.

```
> options(showchars = TRUE)
> ( a <- multipol( as.array(c('x^0' = 0, 'x^1'=1) ) ) )
[1] 1*x^1
> ( b <- multipol( as.array(c('y^0' = 1, 'y^1'=1) ) ) )
[1] 1 + 1*x^1
> a + b
[1] 1 + 2*x^1
```

The location of the problem here is clearly seen since we have chosen to print the polynomials at each step – regardless of our attempt to force (in a simple way) the labeling of the variables, we achieve the same incorrect result.

Apart from the difficulty of variable specification (which alone makes the package so inconvenient that it is not practically usable for our purposes), to ensure that the arithmetic could even be done would require preprocessing each polynomial in anticipation of arithmetic with the others. In other words, it would require padding `arrays` with zeros to ensure that they are similarly shaped and then transposing them to make sure they are aligned correctly. This suggests a more fundamental drawback to the `array` representation which makes the package totally unusable for the current work.

3. The `array` representation does not allow for sparse polynomials. A sparse (multivariate) polynomial is a polynomial of multidegree $\boldsymbol{d}$ which has "few" terms $c_{\boldsymbol{d}'}\boldsymbol{x}^{\boldsymbol{d}'}$ with multidegree $\boldsymbol{d}' \leq \boldsymbol{d}$ (element-wise) with nonzero coefficients.

As an example, consider the polynomial

$$ab^2 + bc^2 + cd^2 + \cdots + yz^2 + za^2 \tag{6.4}$$

and its representation in `multipol`. Of course, since `multipol` defines multivariate polynomials using `arrays`, the user must first determine the polynomial's array representation in order to input it into the computer. However, a moment's consideration proves this is not feasible since the representation requires a 26 dimensional array (a dimension for each variable) each with three levels (e.g., `a^0`, `a^1`, `a^2`). This requires an array with $3^{26} = 2,541,865,828,329$ cells, all but 26 of which are nonzero. Storing each of the coefficients in a double-precision floating point format (64 bits each), this amounts to 20.33 terabytes of space, slightly more than the text content of the Library of Congress ($\approx$ 20TB). This is of course ridiculous, as the actual character string itself

```
a b^2 + b c^2 + c d^2 + ... + y z^2 + z a^2
```

stored in Unicode (16 bits per character) requires only .41 kilobytes. This storage problem is a fatal flaw for the `multipol` package in the current context. Since this thesis deals with "large" polynomial rings (ones with many variables), `multipol` is not a realistic option.

6.1.2 The `mpoly` package

`mpoly` is a complete reenvisioning of how multivariate polynomials and symbolic computing with multivariate polynomials should be implemented in R. Unlike `multipol`, `mpoly` uses as its most basic data structure the `list`. This fundamental change allows us to dispense of every issue with `multipol` discussed in the previous subsection.

However, while it uses `lists`, the motivation can be most easily seen from the angle of `data.frames`, so we begin there.

`data.frames` are very similar to matrices – they are two dimensional arrays. In a multivariate polynomial with p variables, one improvement over `multipol` would be to make `data.frames` the basis of an S3 class called `mpoly`. The `data.frame` would have $p + 1$ columns, one for each variable degree and one for the coefficient. Each row of the `data.frame` would represent a term of the polynomial, and only terms with nonzero coefficients would be stored. Considering an `mpoly` object as a classed `data.frame` of dimension $n \times (p+1)$ in this way, it would represent a polynomial in p variables with n terms. This is demonstrated in the following example of a polynomial in the ring $\mathbb{R}[x, y, z]$.

```
> df <- data.frame(
+   x = c(0,10,2,0,1),
+   y = c(0,0,0,5,1),
+   z = c(0,0,0,0,0),
+   coef = c(1,2,3,4,5)
+ )
> df
   x y z coef
1  0 0 0    1
2 10 0 0    2
3  2 0 0    3
4  0 5 0    4
5  1 1 0    5
> class(df) <- 'mpoly'
```

```
> str(df)
List of 4
 $ x   : num [1:5] 0 10 2 0 1
 $ y   : num [1:5] 0 0 0 5 1
 $ z   : num [1:5] 0 0 0 0 0
 $ coef: num [1:5] 1 2 3 4 5
 - attr(*, "row.names")= int [1:5] 1 2 3 4 5
 - attr(*, "class")= chr "mpoly"
```

The polynomial represented is $1 + 2x^{10} + 3x^2 + 4y^5 + 5xy$. The representation is quite a bit better than the `multipol` representation, which would require an array of dimension 11×6 ($\times$ 1); however, after some reflection it is clear inefficiencies are still present – we still store a large number of zeros. While the `data.frame` representation is able to deal with sparse polynomials (few terms) well, there is still room for improvement. In particular, the `data.frame` representation is poor for what we call super sparse (multivariate) polynomials – sparse polynomials whose terms do not involve many variables. The polynomial in (6.4) is in fact super sparse, since each of its terms only involves two of the 26 possible variables. Were an `mpoly` object a classed `data.frame`, representing the polynomial in (6.4) would require 26 rows (one for each term) and 27 columns (one for each variable degree and one for the coefficient); however, in each row only three of the entries would be nonzero – the two variables present and the coefficient. Thus, the `data.frame` would have $26 \times 27 = 702$ entries, but only $26 \times 3 = 78$ of those entries would be nonzero, making the object only 11.1% nonzero. In other words, eight out of every nine stored numbers would be wasted as zeros stored in double precision. Thus, the problem of sparsity can easily be significant in terms of storage even for a `data.frame` representation, and the issue

becomes even more prohibitive as the number of variables increases.

A solution is available using the `list` data structure. `lists` are `vectors` whose elements can be any other R object – `data.frames`, `lists`, even `functions`, `formulas`, and `languages`. In the `mpoly` package, an `mpoly` object is an S3 class object which is a `list`. The elements of the `list` are each named `numeric vectors`, with unique names including `coef`. Naturally, each element of an `mpoly` object corresponds to a term in the polynomial, and the element of each term named `coef` is the coefficient of that term. Constructing an `mpoly` object this way is very straightforward using the constructor `mpoly`.

```
> library(mpoly)
Loading required package: stringr
Loading required package: rSymPy
Loading required package: rJython
Loading required package: rJava
Loading required package: rjson
> polyList <- list(
+   c(x = 0,  y = 0, z = 0, coef = 1),
+   c(x = 10, y = 0, z = 0, coef = 2),
+   c(x = 2,  y = 0, z = 0, coef = 3),
+   c(x = 0,  y = 5, z = 0, coef = 4),
+   c(x = 1,  y = 1, z = 0, coef = 5)
+ )
> polyList
[[1]]
   x    y    z coef
```

```
   0    0    0    1
[[2]]
   x    y    z coef
  10    0    0    2
[[3]]
   x    y    z coef
   2    0    0    3
[[4]]
   x    y    z coef
   0    5    0    4
[[5]]
   x    y    z coef
   1    1    0    5
> poly <- mpoly(polyList)
> class(polyList)
[1] "mpoly"
```

Looking at the list before using the `mpoly` constructor, it seems like this is essentially the same formulation as the `data.frame` representation. However, the constructor does all the work to drop extraneous information as we can see when we unclass the object. In the following notice that the zeros from the `data.frame` representation are discarded; this is the key to properly incorporating sparsity when storing multivariate polynomials in R.

```
> unclass(polyList)
[[1]]
coef
```

```
   1
[[2]]
   x coef
  10    2
[[3]]
   x coef
   2    3
[[4]]
   y coef
   5    4
[[5]]
   x    y coef
   1    1    5
```

Unlike multivariate polynomials in `multipol`, those in `mpoly` not only have variable names but also an intrinsic variable order which is taken to be the `unique` names of elements of the `mpoly` minus `coef`.[1]

```
> vars(poly)
[1] "x" "y"
```

Viewing a multivariate polynomial as a `list` is a cumbersome task. To make things easier, a `print` method for `mpoly` objects exists and is dispatched when the object is queried by itself at the command prompt.

```
> poly
```

[1] To be clear, this is in the `unique(names(unlist(mpoly)))` sense. Thus, the order of the terms matters when determining the intrinsic order.

```
1  +  2 x^10  +  3 x^2  +  4 y^5  +  5 x y
```

Notice the order of the terms presented in the printed version of the `mpoly` object; it is the order in which the terms are coded in the `mpoly` object itself. This can be changed in either of two ways. First, it can be changed via the `print` method, which accepts arguments `order` for the total order used (lexicographic, graded lexicographic, and graded reverse lexicographic) and `varorder` for a variable order different than the intrinsic order. When an order is requested but a variable order is not specified, the method messages the user to alert them to the intrinsic variable order being used.[2]

```
> print(poly, order = 'lex')
using variable ordering - x, y
2 x^10  +  3 x^2  +  5 x y  +  4 y^5  +  1
> print(poly, order = 'grlex')
using variable ordering - x, y
2 x^10  +  4 y^5  +  3 x^2  +  5 x y  +  1
> print(poly, order = 'lex', varorder = c('y','x'))
4 y^5  +  5 y x  +  2 x^10  +  3 x^2  +  1
> print(poly, order = 'glex', varorder = c('y','x'))
2 x^10  +  4 y^5  +  5 y x  +  3 x^2  +  1
```

Second, the elements of the `mpoly` object can be reordered to create a new `mpoly` object using the `reorder` method.

```
> poly
```

[2]This is a subtle point. It is very possible that a polynomial in the ring $\mathbb{R}[x, y]$ is coded with the intrinsic order `y, x` and that, by consequence, the lexicographic order will not be the one intended. The messaging is used to make clear what order is being used.

```
1  +  2 x^10  +  3 x^2  +  4 y^5  +  5 x y
> ( poly2 <- reorder(poly, order = 'lex') )
using variable ordering - x, y
2 x^10  +  3 x^2  +  5 x y  +  4 y^5  +  1
> unclass(poly2)
[[1]]
   x coef
  10    2
[[2]]
   x coef
   2    3
[[3]]
   x    y coef
   1    1    5
[[4]]
   y coef
   5    4
[[5]]
coef
   1
```

The major workhorse of the package is the constructor itself. In particular, polynomial reduction (combining of like terms) and regularization (combining of coefficients and like variables within terms) are both performed when the multivariate polynomials are constructed with `mpoly`.

```
> list4mpoly <- list(
```

```
+   c(x = 1, coef = 1),
+   c(x = 1, coef = 2)
+ )
> mpoly(list4mpoly)
3 x
>
> list4mpoly <- list(  c(x = 5, x = 2, coef = 5, coef = 6, y = 0) )
> mpoly(list4mpoly)
30 x^7
```

While the `mpoly` constructor is nice, it is inconvenient to have to keep specifying `lists` in order to define polynomials. The `mp` function was constructed for this purpose and is intended to make defining multivariate polynomials quick and easy by taking them in as character strings and parsing them into `mpoly` objects.

```
> ( p <- mp('10 x + 2 y 3 + x^2 5 y') )
10 x  +  6 y  +  5 x^2 y
> is.mpoly(p)
[1] TRUE
> unclass(p)
[[1]]
   x coef
   1   10
[[2]]
   y coef
   1    6
[[3]]
```

```
x    y coef
2    1    5
```

This parsing is a nontrivial process and depends heavily on the specification of the polynomial in the string. The `mp` function must first determine the variables that the user is specifying (which must be separated by spaces for disambiguation) and then construct the `list` to send to `mpoly` to construct the object. Because it is passed through `mpoly`, the `mpoly` object returned by `mp` is reduced and regularized.

```
> mp('x^2 + 10 x 6 x + 10 x 6 x y y 2')
61 x^2  +  120 x^2 y^2
```

The `mpoly` package has much more to offer than simply defining polynomials. Methods are available for addition, subtraction, multiplication, exponentiation and equality as well. Moreover, since `mpoly` objects know their variable names intrinsically, we can perform arithmetic with whichever polynomials we like. For example, the arithmetic with a and b from (6.2) and (6.3) is easy –

```
> a <- mp('1 + 2 x + 3 y + 4 x y')
> b <- mp('1 + 2 x + 3 y + 4 x y + 5 z + 6 x z + 7 y z + 8 x y z')
> a + b
2  +  4 x  +  6 y  +  8 x y  +  5 z  +  6 x z  +  7 y z  +  8 x y z
> b - a
5 z  +  6 x z  +  7 y z  +  8 x y z
> a * b
1  +  4 x  +  6 y  +  20 x y  +  5 z  +  16 x z  +  22 y z  +
60 x y z  +  4 x^2  +  16 x^2 y  +  12 x^2 z  +  40 x^2 y z  +
9 y^2  +  24 x y^2  +  21 y^2 z  +  52 x y^2 z  +  16 x^2 y^2  +
```

```
32 x^2 y^2 z
```

Exponentiation and equality are also available.

```
> a^2
1 + 4 x + 6 y + 20 x y + 4 x^2 + 16 x^2 y + 9 y^2 +
24 x y^2 + 16 x^2 y^2
> a == b
[1] FALSE
> ( c <- mpoly(a[c(2,1,4,3)]) )
4 y x + 3 y + 2 x + 1
> a == c
[1] TRUE
```

Here also each of the results are reduced and regularized. While the computations are done entirely in R, they are quite efficient; each of the above calculations is virtually instantaneous.

But `mpoly` does not stop there. The basic operations of the differential calculus, partial differentiation and gradients, are also available to the user and are efficient. A `deriv` method exists for `mpoly` objects which can be dispatched,[3] and the `gradient` function is built on `deriv` to compute gradients.

```
> deriv(b, 'x')
8 y z + 4 y + 6 z + 2
> gradient(b)
8 y z + 4 y + 6 z + 2
```

[3] `deriv` does not call the default method from the `stats` package

```
8 x z  +  4 x  +  7 z  +  3
8 x y  +  6 x  +  7 y  +  5
```

The gradient is a good example of another class object in the `mpoly` package, the `mpolyList`. `mpolyLists` are simply `lists` of `mpoly` objects and are used to hold vectors of multivariate polynomials. They can be easily specified using the `mp` function on a `vector` of character strings.

```
> mp(c('x + y + z', 'x + z^2'))
x  +  y  +  z
x  +  z^2
>
> str(mp(c('x + y + z', 'x + z^2')), 1)
List of 2
 $ :List of 3
  ..- attr(*, "class")= chr "mpoly"
 $ :List of 2
  ..- attr(*, "class")= chr "mpoly"
 - attr(*, "class")= chr "mpolyList"
```

The viewing of `mpolyList` objects is made possible by a `print` method for `mpolyList` objects just like for `mpoly` objects. Moreover addition, subtraction, and multiplication are defined for `mpolyList` objects as well; they each operate element-wise.

In addition to differentiation, `mpoly` can also compute Gröbner bases of collections of multivariate polynomials (`mpolyList` objects) by passing the proper objects in the proper syntax to the `rSymPy` package which has an implementation of Buchberger's

algorithm.[4] The computations are performed by a Java based Python implementation and are quite fast once the Java Virtual Machine (JVM) has been initialized. The Gröbner basis is then returned as an `mpolyList` object.

```
> polys <- mp(c('t^4 - x', 't^3 - y', 't^2 - z'))
> gb <- grobner(polys)
using variable ordering - t, x, y, z
> gb
-1 z  +  t^2
t y  -  z^2
-1 y  +  z t
x  -  z^2
y^2  -  z^3
> class(gb)
[1] "mpolyList"
```

Moreover, `grobner` can calculate Gröbner bases with respect to various monomial orders and any variable ordering.

```
> polys <- mp(c('x^2 - 2 y^2', 'x y - 3'))
> grobner(polys, varorder = c('x', 'y'))
3 x  -  2 y^3
-9  +  2 y^4
>
```

[4]Buchberger's algorithm is the standard method of calculating Gröbner bases, however, faster methods are known. One such method is Faugère's F4 and F5 algorithms. Unfortunately, there are exceedingly few implementations of these faster algorithms which exist, and none available in R.

```
> grobner(polys, varorder = c('x', 'y'), order = 'grlex')
-3 x  +  2 y^3
x^2  -  2 y^2
-3  +  x y
```

Unfortunately, there is currently no Gröbner walk algorithm available to convert a Gröbner basis in one monomial order to a Gröbner basis in another, a technique often used to quickly compute Gröbner bases in more difficult orders (e.g. lexicographic) from "easier" ones (e.g. graded reverse lexicographic), so there are still a number of improvements which can be made.

Apart from being interesting algebraic objects, polynomials can of course be thought of as functions. To access this functional perspective, we can consider a multivariate polynomial as a function by converting an `mpoly` or `mpolyList` object to a `function` object using an `mpoly` method for the generic `as.function` method. This is particularly suited to R's strengths since R is geared towards numerically maximizing functions.

```
> library(ggplot2); theme_set(theme_bw())
> ( p <- mp('x') * mp('x - .5') * mp('x -  1') ) # 0's at 0, .5, 1
x^3  -  1.5 x^2  +  0.5 x
> f <- as.function(p)
f(x)
> s <- seq(-.1, 1.1, length.out = 201)
> df <- data.frame(x = s, y = f(s))
> qplot(x, y, data = df, geom = 'path') +
+   geom_hline(yintercept = 0, colour = I('red'))
```

The plot generated is included in Figure 6.1, where one can see that the function has the correct zeros. For multivariate polynomials the syntax is the same, and

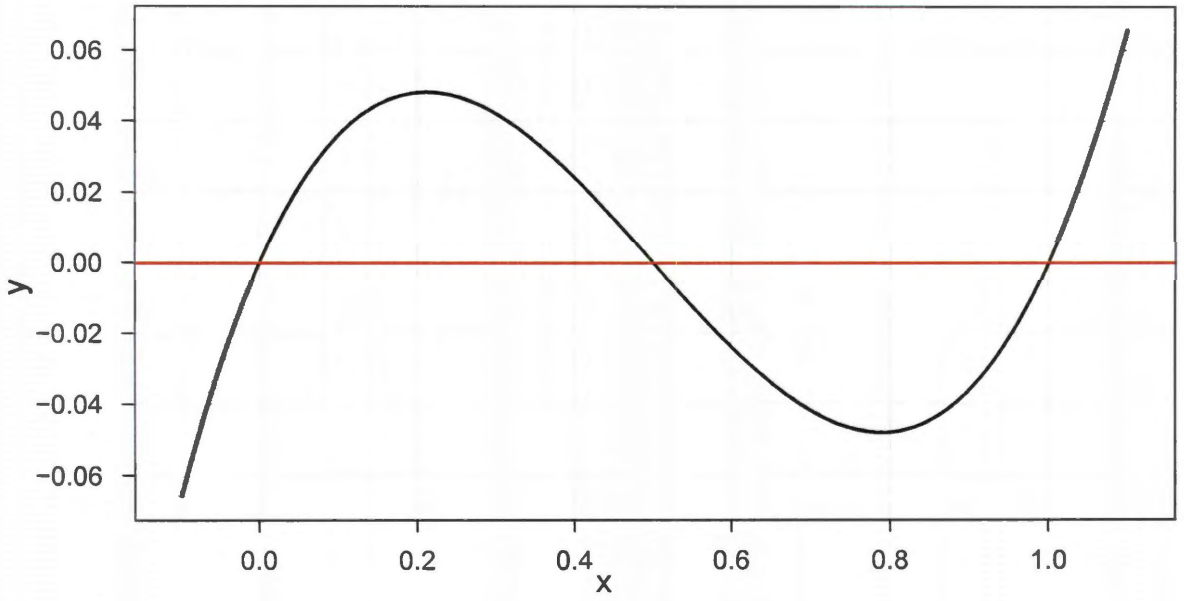

Figure 6.1 : The `as.function` method for `mpoly` objects

`as.function` can provide the function one of a `vector` argument (ideal for passing to optimization routines such as `optim`) or a sequence of arguments.

```
> mpoly <- mp('x + 3 x y + z^2 x')
>
> f <- as.function(mpoly)
f(.) with . = (x, y, z)
> f(1:3)
[1] 16
>
```

```
> f <- as.function(mpoly, vector = FALSE)
f(x, y, z)
> f(1, 2, 3)
[1] 16
```

`mpolyList` objects can be converted into functions in the same way. Note that in both cases the user is messaged with the appropriate syntax for the resulting function.

```
> polys <- mp(c('x + 1', 'y^2 + z'))
> f <- as.function(polys)
f(.) with . = (x, y, z)
> f(1:3)
[1] 2 7
```

While other functionality is provided (such as a `terms` method and function to compute least common multiples), these are the main contributions of the `mpoly` package. Put together, they provide a package which is not only user-friendly, efficient, and useful for polynomial algebra, but also a solid foundation upon which further developments can be made to make polynomials more accessible in R. In particular, it provides a nice starting point for any future package dealing with algebraic statistics. For now, it provides much needed functionality for estimating categorical conditional independence models with the novel `catcim` package.

6.2 `catcim`

The basic problem with a practical implementation for fitting categorical conditional independence models in R is that of structuring the problem. In particular, questions concerning how the data and the conditional independence model can be easily and

efficiently represented in R become of paramount importance. To illustrate how these problems are resolved we base our discussion on the following example.

EXAMPLE 6.1 *Edwards and Kreiner [1983] consider the unpublished data set in Table 6.1 from a study conducted at the Institute for Social Research, Copenhagen, collected in 1978-1979. The data represents 1,591 employed men aged 18-67 and asks whether in the past year they had done any work on their home which they previously would have paid a skilled craftsmen to do. Five categorical variables are represented – age, residence, employment, mode, and response – whose levels can be seen in the table.*

||

Question 1 *How can we represent the data in a statistical computing environment?*

Anyone who has dealt with multi-way contingency table data in the `data.frame`-driven statistical computing environment R can easily identify with the problem. R prefers to treat data sets in the classical multivariate context as matrices whose rows are single observations and columns are variables of a particular type (numeric, string, factor, logical, etc.), the data structure it calls a `data.frame`. If we represent this data as a `data.frame` in this way, we will be forced to have a 1,592 $\times$ 5 data structure where many of the rows are redundant. Such a representation has the form

```
     Age Residence Employment Mode Response
1   < 30     House     Office  Own      Yes
2 46-67 Apartment  Unskilled Rent      Yes
3   < 30     House    Skilled  Own      Yes
4   < 30     House    Skilled  Own       No
```

				Age		
Residence	**Employment**	**Mode**	**Response**	**< 30**	**31-45**	**46-67**
Apartment	Skilled	Rent	Yes	18	15	6
			No	15	13	9
		Own	Yes	5	3	1
			No	1	1	1
	Unskilled	Rent	Yes	17	10	15
			No	34	17	19
		Own	Yes	2	0	3
			No	3	2	0
	Office	Rent	Yes	30	23	21
			No	25	19	40
		Own	Yes	8	5	1
			No	4	2	2
House	Skilled	Rent	Yes	34	10	2
			No	28	4	6
		Own	Yes	56	56	35
			No	12	21	8
	Unskilled	Rent	Yes	29	3	7
			No	44	13	16
		Own	Yes	23	52	49
			No	9	31	51
	Office	Rent	Yes	22	13	11
			No	25	16	12
		Own	Yes	54	191	102
			No	19	76	61

Table 6.1 : Data on home repairs from Edwards and Kreiner [1983]

```
5 31-45     House     Office  Own      Yes
6 31-45 Apartment  Unskilled  Own       No
.     .         .          .    .        .
.     .         .          .    .        .
```

```
   .     .          .          .    .        .
1590 31-45     House  Unskilled  Own      Yes
1591  < 30     House    Skilled  Own      Yes
```

Alternatively, and perhaps most intuitively, we can represent the data as a 5-way array of numbers in the data structure designed for such tasks, the `array` object in R. While this representation is more efficient than the long `data.frame` format, it is woefully unwieldy – as noted in the `mpoly` section multidimensional arrays are notoriously difficult to work with in R. Moreover, standard modeling syntax in R asks for `data.frames`, not `arrays`. Thus, while the `array` is very intuitive, it is not commensurate with a practical implementation. A view of the `array` representation of the data is presented below.

```
, , Employment = Skilled, Mode = Rent, Response = Yes

       Residence
Age     Apartment House
  < 30         18     1
  31-45        15    17
  46-67         5    34

, , Employment = Unskilled, Mode = Rent, Response = Yes

       Residence
Age     Apartment House
  < 30          2    25
```

```
  31-45          3      8
  46-67         30      4
```

.

.

.

```
, , Employment = Office, Mode = Own, Response = No

        Residence
Age      Apartment House
  < 30          49    12
  31-45         51   102
  46-67         11    61
```

We advocate the use of a "short" `data.frame` format which we will call the `df` or `count` format. Like the usual data matrix, the `df` format is a `data.frame` object; however, instead of having only p columns as in the classical multivariate context it has $p + 1$ columns. The additional column called `count` contains the number of repetitions of the preceding p columns. Using the language of Wickham [2007], it is a kind of "melted" `array` object which preserves all of the variable names, levels, and count information of the natural `array` format in a wieldy `data.frame` data structure. This shortened data representation has the form

```
Age Residence Employment Mode Response count
```

1	< 30	Apartment	Skilled	Rent	Yes	18
2	< 30	Apartment	Skilled	Rent	No	15
3	< 30	Apartment	Skilled	Own	Yes	5
4	< 30	Apartment	Skilled	Own	No	1
5	< 30	Apartment	Unskilled	Rent	Yes	17
6	< 30	Apartment	Unskilled	Rent	No	34
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
71	46-67	House	Office	Own	Yes	102
72	46-67	House	Office	Own	No	61

While these are the three data representations considered for the implementation presented in this work, they are by no means the only way we could represent the data. For example, the widely successful technique of multiple correspondence analysis is based on still another representation of categorical data using a special kind of indicator matrix. That being said, while these other representations are worthwhile avenues for investigation they are not considered any further in this work as the current representation easily suffices for the task at hand.

Once the data can be easily handled in R, we can do basic operations with it such as computing the empirical distribution (relative frequencies) as well as marginal tables. More importantly, the `data.frame` structure used for our representation is the canonical data structure used for model fitting in R. However before we can consider model fitting we are presented with a more subtle problem of representing

the conditional independence model.

Question 2 *How can we represent a conditional independence model in a statistical computing environment?*

Conditional independence models are, by definition, collections of statements such as "A and B are conditionally independent of E given C" and "D is conditionally independent of B and C given E and F," and so forth. If we are to write an R package which is flexible enough to fit any conditional independence model, it must first have a syntax with which one can easily specify conditional independence models and then be able determine the proper restrictions on the cell probabilities themselves from those specifications.

The issue of model specification is resolved by using a combination of the existing R `formula` abstraction in concert with string operations. For example, using the data from Example 6.1 Edwards and Kreiner [1983] consider the model

$$\text{(Response, Employment)} \perp\!\!\!\perp \text{Residence} \mid \text{(Age, Mode)}, \tag{6.5}$$

which we interpret as "if we know the age of an individual and whether he rents or owns his accommodation, knowing whether he repaired those accommodations in the past month and how he is employed is irrelevant in predicting whether he lives in an apartment or a house." In other words, the experimental model dictates that all we need to know is how old he is and if he rents or owns, then we can predict whether he lives an apartment or a house. The novel syntax allows the easy specification of this model as

```
c(Response, Employment) || Residence ~ c(Age, Mode)        (6.6)
```

within an environment where those variables are well-defined, for example, the general purpose fitting function of the `catcim` package `cim`. The syntax is clean

enough to be legible regardless of level of expertise and is generalized to any number of such statements using `lists` of such formulas.

Once the models are specified, they need to be translated into the system of polynomial constraints. This is where `mpoly` comes in. Using `mpoly` along with carefully designed string operations in R we can translate the specified conditional independence model into its polynomial representation using the method described in the paragraph following (3.34), a method of expanding pluses and rearranging. This representation can then be used to calculate the steps common to all of the fitting algorithms in Chapter 5, namely, to change the nonlinear optimization problem with a complicated feasibility region into a Lagrangian, which is itself a polynomial and therefore representable as an `mpoly` object in the `mpoly` package.

We now come to the point where we can actually use `cim`. Recall that the model to be fit is

$$\text{(Response, Employment)} \perp\!\!\!\perp \text{Residence} \mid \text{(Age, Mode)}. \tag{6.7}$$

From a user perspective, the primary function in the `catcim` package is `cim`, which stands for conditional independence model. The labeling is meant to parallel pre-existing syntaxes such as `gam` for generalized additive models or `glm` for generalized linear models. Thus, we simply enter

```
cim(c(Response, Employment) || Residence ~ c(Age, Mode), data = Repairs)
```

at the command prompt and `cim` returns a data frame with $p+2$ columns. The first p are used to enumerate the configurations, and the last two report $\widehat{\pi}_{EMP}$ and $\widehat{\pi}_{L2E}$.

In addition to the L2E, `catcim` can also compute the other estimators discussed in Chapter 4. It does so with the `dist` argument which accepts `'l2e'`, `'mle'`, `'mcs'`,

and `'ncs'`. (It also accepts `'emp'` if only the empirical is desired.) Whichever is specified, the result has the same structure of a `data.frame` with the estimate of the joint distribution appended as a column.

Of course, various fitting methods are available as well and are accessed with the `method` argument. The default is `'auto'`, which will implement the limited memory boxed BFGS omnibus nonlinear solver discussed in Chapter 5. In addition to `'auto'`, `method` accepts arguments `'grobner'`, `'sdp'`, and `'BB'`; the first two of which are only available for the L2E. The last, `'BB'`, implements yet another nonlinear scheme to solve the optimization problem discussed in the previous chapter; it uses primarily spectral methods (Gilbert and Varadhan [2009]).

Chapter 7

Dr. Pangloss' Method – Simulation

In Chapter 4 we presented various theoretical results concerning the L2E, MLE, minimum chi-squared, and minimum Neyman modified chi-squared estimators. The primary focus in that chapter was the asymptotic behavior of the estimators. In this chapter we consider the other extreme by looking at the finite (small) sample properties of each of the estimators by considering not their distance from normality in some kind of functional sense but their bias and mean squared error. Since we cannot check all possible models, we conduct the assessment using a small collection of models and sample sizes which can be expected to arise in practical situations. Instead of a rigorous mathematical proof, large scale simulations are run. All of the simulations make use of reduced models which serve as a proxy and are representative of the true conditional independence models. Using these surrogates, two kinds of simulation are used. For small tables and sample sizes, we use high-performance computing to literally enumerate each of the possible outcomes along with their probabilities and calculate the theoretical quantities of interest for each of the estimators across the reduced model. This procedure, detailed in Section 7.3, is entirely deterministic – it simply uses the computer to do large-scale arithmetic.[1] For moderate to large problems, the method fails due to combinatorial explosion and we are forced to use stochastic simulation techniques in order to get a grasp on the quantities of interest.

[1] In the smallest of cases this can even be done symbolically (i.e., without using a reduced model proxy), and the results have been seen to agree with those using the surrogate model method.

Both of the methods are highly computationally intensive and attempt to illuminate the deeper theory underpinning the estimators.

7.1 Bias and Mean Squared Error

For any statistical model $\mathcal{M}_{r-1}$ and collection of data $\mathcal{D} = \boldsymbol{T}$ (a contingency table), the bias of a statistic $\widehat{\pi} = \boldsymbol{S}(\boldsymbol{T})$ at a point $\pi \in \mathcal{M}_{r-1}$ in the model is defined

$$\boldsymbol{Bias}(\pi, \boldsymbol{S}) \;=\; \boldsymbol{Bias}_{\pi}[\boldsymbol{S}] \;=\; \mathbb{E}_{\pi}\left[\boldsymbol{S}(\boldsymbol{T})\right] - \pi \;\in\; \mathbb{R}^{r}. \tag{7.1}$$

Thus, $\boldsymbol{Bias}_{\pi}(\bullet)$ is a collection of functionals indexed by $\mathcal{M}_{r-1}$ which take statistics $\boldsymbol{S}$ to vectors in $\mathbb{R}^r$. Equivalently, for a fixed statistic $\boldsymbol{S}$ $\boldsymbol{Bias}_{\bullet}(\boldsymbol{S})$ is a function over the model $\mathcal{M}_{r-1}$. If $\boldsymbol{Bias}_{\pi}[\boldsymbol{S}] = \boldsymbol{0}$ for all $\pi \in \mathcal{M}_{r-1}$ so that $\boldsymbol{Bias}_{\bullet}(\boldsymbol{S})$ is the zero function, $\boldsymbol{S}$ is said to be unbiased. A related quantity, the mean squared error (MSE) is defined

$$\boldsymbol{MSE}(\pi, \boldsymbol{S}) \;=\; \boldsymbol{MSE}_{\pi}[\boldsymbol{S}] \;=\; \mathbb{E}_{\pi}\left[(\boldsymbol{S}(\boldsymbol{T}) - \pi)(\boldsymbol{S}(\boldsymbol{T}) - \pi)'\right] \;\in\; \mathbb{R}^{r\times r}, \tag{7.2}$$

and exhibits the same functional/function duality as $\boldsymbol{Bias}$ since formally $\boldsymbol{MSE}$: $\mathcal{M}_{r-1} \times \mathcal{S} \to \mathbb{R}^{r\times r}$ (where $\mathcal{S}$ is the collection of all statistics). While bias and mean squared error are typically considered in the context of the estimation of a parameter $\theta \in \Theta$ (e.g. "$\overline{X}$ is unbiased for the mean μ in a normal model"), here they refer to the estimation of the distributions themselves as a consequence of the fact noted in the introduction to Chapter 4 that categorical models are trivially self-parameterizable.

As defined above, $\boldsymbol{Bias}_{\pi}[\boldsymbol{S}]$ and $\boldsymbol{MSE}_{\pi}[\boldsymbol{S}]$ are difficult to interpret because they are 1. multivariate, the former being vector-valued and the latter matrix-valued, and 2. defined on a multivariate domain ($\mathcal{M}_{r-1}$). While bias and mean squared error are classical means of comparing estimators, they are typically considered in

the univariate context; the above definitions are simply natural generalizations to the multivariate context. In order to compare them, therefore, we consider their L_2 and Frobenius norms, respectively. The reason for the L_2 norm being obvious, the motivation for the Frobenius norm is that 1. it is the L_2 entrywise norm, but more relevantly 2. it is equivalent to the sum of the squares of the singular values of a matrix, which are related to the lengths of the axes of an ellipsoid. To deal with the multivariate domain, we simply average the values over the model. Thus, for every model/sample size/estimator triple, a single quantity is available which describes the bias of the estimator and a single quantity is available which describes the dispersion of the estimator. These single quantities can then be compared across estimators for a fixed model and sample size to draw conclusions concerning their relative merit.

As with the other aspects of this thesis, here also one of the major difficulties in our analysis is that in general none of the estimators has a closed form. It therefore does not seem possible to analytically calculate either the bias or the mean squared error of the estimators as functions of the unknown distribution (parameter) $\boldsymbol{\pi}$. Worse, $\widehat{\boldsymbol{\pi}}_{X^2_N}$ is not even well defined for every possible outcome (table) $\boldsymbol{T}$. For this reason, we omit Neyman's modified chi-squared statistic from our simulations.

As noted, holding the estimator fixed the bias and MSE are functions of the target parameter $\boldsymbol{\pi}$. Since for each simulation we discretize the model, we only know the values of these functions at those points. This is not so great a concern, however, as both are continuous functions of that parameter.

PROPOSITION 7.1 (KAHLE) *The $\boldsymbol{Bias}_{\boldsymbol{\pi}}$ and $\boldsymbol{MSE}_{\boldsymbol{\pi}}$ of $\widehat{\boldsymbol{\pi}}_{L2E}$ are continuous functions of $\boldsymbol{\pi}$ for any conditional independence model; the same is true for the estimators $\widehat{\boldsymbol{\pi}}_{MLE}$ and $\widehat{\boldsymbol{\pi}}_{X^2}$ for conditional independence model with the additional condition that $\boldsymbol{\pi}^\star \geq \epsilon \mathbf{1}_r$, for some $\epsilon \in (0, 1)$.*

Since $\boldsymbol{Bias}_\pi$ and $\boldsymbol{MSE}_\pi$ are continuous, knowing them on a mesh in particular is quite indicative of what is going on over the entire model. Since these norms are continuous, the averaging effectively approximates an integral; moreover, the approximation is reasonable since the mesh is uniform. This is exactly the case in Dr. Pangloss' method described next.

7.2 Dr. Pangloss' all possible worlds

The method here termed "Dr. Pangloss' method" is a nonstochastic simulation which allows us to understand and make pseudo-theoretical conclusions concerning the various estimators in various settings (models) without direct knowledge of the theoretical objects themselves, i.e. the closed forms of the estimators and their finite sample distributions. The namesake, Dr. Pangloss, is a character from Voltaire's 1759 satirical chef d'oevre *Candide* who parodies the Leibniz' philosophy and theology with the mantra "tout est pour le mieux dans le meilleur des mondes possibles" ("everything is for the best in the best of all possible worlds"). Since the method attempts to generate "all possible worlds" – a finite collection of distributions which serve as a model proxy – the method bears his name. An illustration for the concept is provided in Figure 7.2 for the independence model in the 2×2 contingency table discussed in Section 2.2.

The method is more than the typical simulation. It is a unique and extremely useful convenience of categorical statistics that one is able to consider virtually every possible distribution without fear of missing important ones. This is certainly not the case with continuous distributions where simulations are only performed on a barrage of distributions with different shapes and functional forms. For example, the widely cited simulation study Antoniadis et al. [2001] considers wavelet denoising

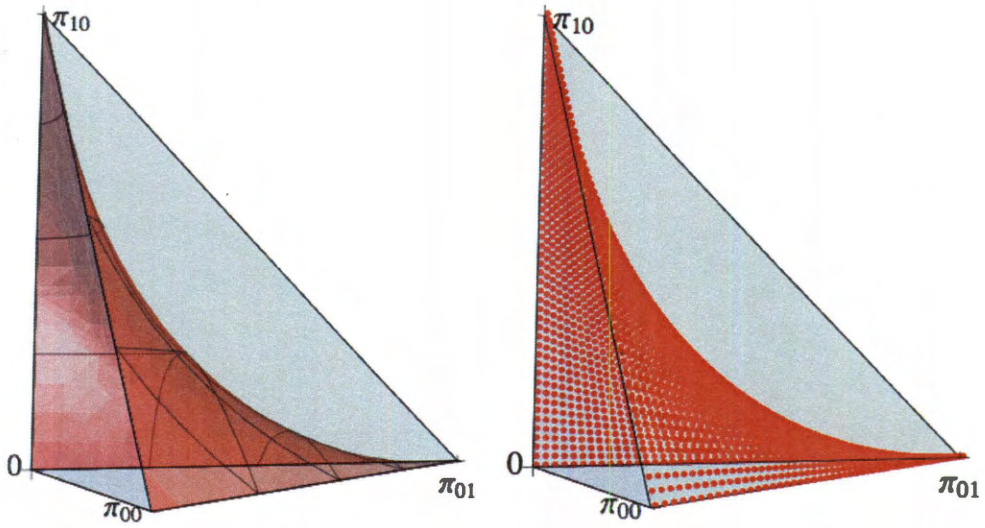

Figure 7.1 : A Pangloss points model reduction for the 2×2 independence model

of univariate signals with one of the most extensive simulations ever performed for method comparison, involving 12 different signal forms (step, wave, blip, etc.) and 22 different methods (minimax, scad, sureshrink, etc.) along with various method configurations, sample sizes and signal to noise ratios. And yet, although highly indicative of the overall picture, the results are still confined to those signals in the sense that if you pick a signal with a longer tail or a higher peak here or there the results may change. Dr. Pangloss' method effectively bypasses this problem by laying a mesh over the entire model in the simplex so that no "more extreme distributions" exist.[2] It is this pseudo-exhaustive enumeration of distributions we consider to be "all

[2]As described this can only be achieved with some of the conditional independence models. Models with singularities which are not parameterizable do not quite meet this requirement as presented here and are thus left for further study. Note that none of the models presented in this chapter exhibit this behavior.

possible worlds" or "Pangloss points". Conclusions concerning the performance of the various estimators when the model is these enumerated points we then extrapolate to the performance of the estimators for the entire model $\mathcal{M}_{r-1}$ (again, for various models). Since the selection of points is very representative of the model itself, we label the results "pseudo-theoretical".

7.3 Algorithmic calculation of exact theoretical quantities

Fortunately, computational methods present another way that we can get a handle on the bias and MSE of the estimators. Since the estimators themselves are discrete random vectors, their bias and mean squared error can be computed via

$$\boldsymbol{Bias}_{\pi}[\widehat{\pi}] \;=\; \mathbb{E}_{\pi}\left[\widehat{\pi}\right] - \pi \;=\; \left(\sum_{\boldsymbol{T}} P_{\pi}\left[\boldsymbol{T}\right] \boldsymbol{S}(\boldsymbol{T})\right) - \pi \tag{7.3}$$

and

$$\boldsymbol{MSE}_{\pi}[\widehat{\pi}] \;=\; \mathbb{E}_{\pi}\left[(\widehat{\pi}-\pi)(\widehat{\pi}-\pi)'\right] \;=\; \sum_{\boldsymbol{T}} P_{\pi}\left[\boldsymbol{T}\right](\boldsymbol{S}(\boldsymbol{T})-\pi)(\boldsymbol{S}(\boldsymbol{T})-\pi)', \tag{7.4}$$

where $\boldsymbol{T}$ ranges over all possible tables for the fixed sample size N and $P_{\pi}\left[\boldsymbol{T}\right]$ is the probability of $\boldsymbol{T}$ assuming π is the true probability of each cell, which follows a $\text{Multinom}_r(N, \pi)$ distribution. If the value of $\boldsymbol{S}$ were known for every $\boldsymbol{T}$, then in principle we could calculate both of the bias and MSE for each of the Pangloss points (i.e. mesh points) and make conclusions based on those quantities. It is this program which we carry out in the basic algorithm, Algorithm 7.1.

EXAMPLE 7.1 *For this example we consider the L2E estimator in the* 2×2 *independence model with a sample size of ten (the others are done similarly). As listed in Table 7.1, with a sample size of ten there are 286 possible tables which once scaled by*

Algorithm 7.1 Dr. Pangloss' method

1: **input** sample size N, reduced model $\mathcal{M}_{r-1}^P = \left\{\boldsymbol{\pi}_j^P\right\}_{j=1}^J$

2: **output** average bias and mean squared error norm for $\mathcal{M}_{r-1}^P$

3: enumerate tables $\{\boldsymbol{T}_k\}_{k=1}^{C_N}$

4: **for all** tables $\boldsymbol{T}_k$ **do**

5: calculate $\widehat{\boldsymbol{\pi}}_{MLE}$, $\widehat{\boldsymbol{\pi}}_{L2E}$, $\widehat{\boldsymbol{\pi}}_{X^2}$

6: **end for**

7: **for all** Pangloss points $\boldsymbol{\pi}_j^P$ **do**

8: calculate probabilities $\left\{P_{\boldsymbol{\pi}_j^P}\left[\boldsymbol{T}_k\right]\right\}_{k=1}^{C_N}$

9: calculate bias $\left|\left|\boldsymbol{Bias}_{\boldsymbol{\pi}_j^P}[\widehat{\boldsymbol{\pi}}]\right|\right|_2$ and $\left|\left|\boldsymbol{MSE}_{\boldsymbol{\pi}_j^P}[\widehat{\boldsymbol{\pi}}]\right|\right|_F$ for each estimator

10: **end for**

11: average $\left\{\left|\left|\boldsymbol{Bias}_{\boldsymbol{\pi}_j^P}[\widehat{\boldsymbol{\pi}}]\right|\right|_2\right\}_{j=1}^J$, $\left\{\left|\left|\boldsymbol{MSE}_{\boldsymbol{\pi}_j^P}[\widehat{\boldsymbol{\pi}}]\right|\right|_F\right\}_{j=1}^J$ and return

the sample size are the 286 possible values of $\widehat{\boldsymbol{\pi}}_{EMP}$. For each of these empirical relative frequencies we fit the model using the techniques of Chapter 5. This is illustrated in Figure 7.2.

Now, since for any fixed distribution $\boldsymbol{\pi}$ and sample size N the distribution $\widehat{\boldsymbol{\pi}}_{EMP}$ is $Multinom_r(N, \boldsymbol{\pi})$, we can use this distribution to attribute probabilities to all possible values of $\widehat{\boldsymbol{\pi}}_{L2E}$. We can therefore visualize the distributions in the projected simplex using point size to represent the relative probability of tables and estimators. This is done for the distribution $\boldsymbol{\pi} = [.25\ .25\ .25\ .25]' \in \mathcal{M}$ in Figure 7.3.

||

EXAMPLE 7.2 *For this example we consider the bias of $\widehat{\boldsymbol{\pi}}_{X^2}$ in the 2×2 independence model with various the sample sizes $N = 5$, 10, and 30. Since $\widehat{\boldsymbol{\pi}}_{X^2}$ is consistent, we*

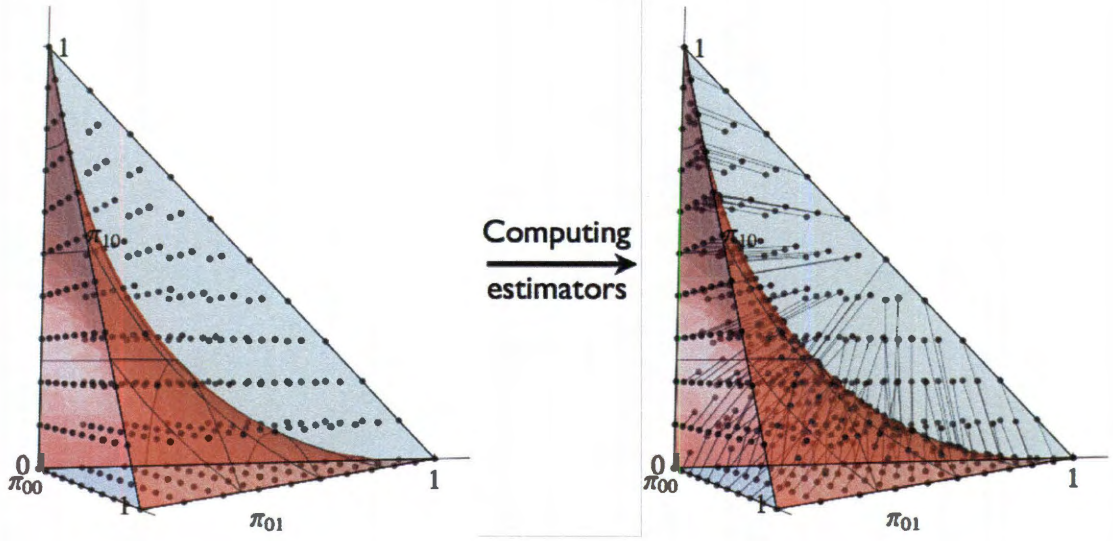

Figure 7.2 : All possible empirical relative frequencies $\widehat{\boldsymbol{\pi}}_{EMP}$ with the sample size $N = 10$ (left) along with the $\widehat{\boldsymbol{\pi}}_{L2E}$s (right)

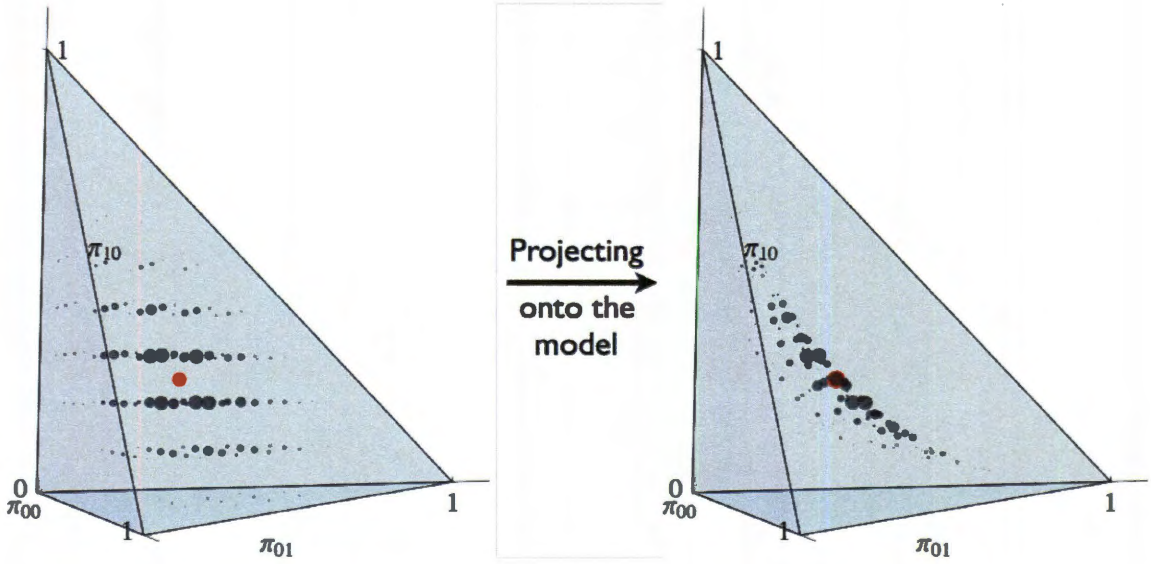

Figure 7.3 : Values of $\widehat{\boldsymbol{\pi}}_{EMP}$ (left) and $\widehat{\boldsymbol{\pi}}_{L2E}$ (right) sized according to their likelihood assuming the Pangloss point $\boldsymbol{\pi}^P = [.25\ .25\ .25\ .25]' \in \mathcal{M}$ and the sample size $N = 10$

know that the bias norm eventually converges to 0. This is resounded by the images in Figure 7.4. In that figure the model surface is colored according to the level of the bias norm. While the bias actually seems to get worse going from $N = 5$ to $N = 10$ samples, the effect is diminished by the time $N = 30$. The surfaces also show where the "hard" places to estimate are – the values intermediate between the simplex boundary and the center of the model. The same effect is observed for each of the estimators.

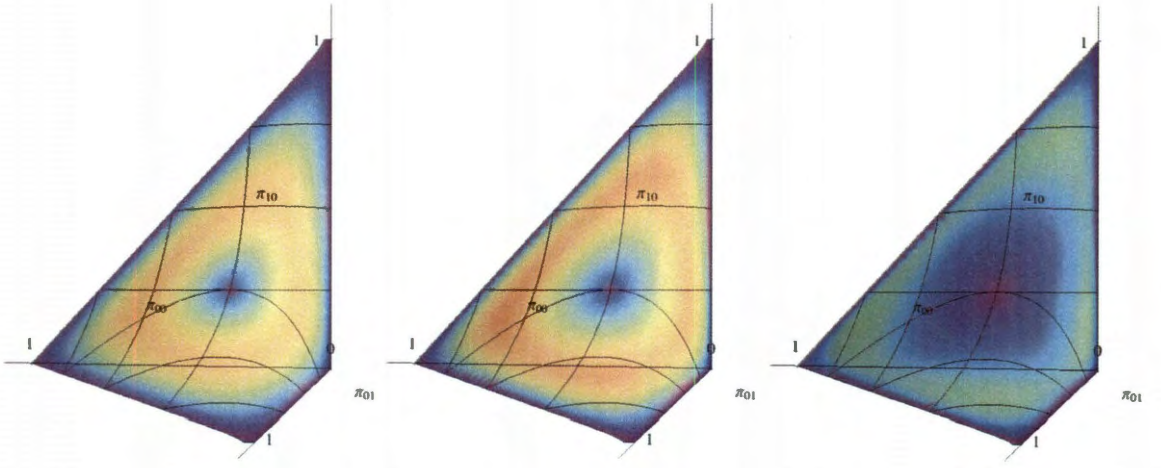

Figure 7.4 : $||\boldsymbol{Bias}_{\boldsymbol{\pi}}[\widehat{\boldsymbol{\pi}}_{X^2}]||_2$ in the 2×2 independence model with sample sizes $N = 5, 10, 30$. The largest bias (red) corresponds to a norm of roughly .05, the smallest (purple), 0.

||

Assuming that $\boldsymbol{S}$ is known, one important question with this for Algorithm 7.1 is – how hard is it to compute (7.3) and (7.4)? The question depends on how many tables there are since for each table we must compute each of the estimators. For a table with sample size N, we denote the number of tables (the number of terms in

the sums) C_N. Since N is known (see Remark 3.1), C_N is fixed and equal to

$$\begin{aligned} C_N &= \#\text{ terms in the sums} \\ &= \#\text{ of tables with sample size } N \\ &= \binom{N+r-1}{r-1} \\ &= \#\text{ of outcomes in a Multinom}_r(N, \boldsymbol{\pi}) \text{ experiment,} \end{aligned} \tag{7.5}$$

an enormous number for even relatively small sample sizes and tables. Table 7.1 provides a feel for the size of C_N for a number of such table sizes as well as how the number of tables explodes combinatorially for even modest tables. Even if we

N	$r = 4$	$r = 6$	$r = 8$	$r = 16$
5	56	252	792	15,504
10	286	3,003	19,448	3,268,760
15	816	15,504	170,544	155,117,520
20	1,771	53,130	888,030	3,247,943,160
25	3,276	142,506	3,365,856	40,225,345,056
30	5,456	324,632	10,295,472	344,867,425,584
50	23,426	3,478,761	264,385,836	207,374,699,821,536

Table 7.1 : C_N for 2×2, 2×3, $2 \times 2 \times 2$ and $2 \times 2 \times 2 \times 2$ tables with sample size N

could compute the estimators instantly, we would not be able to store the list of tables for large sample sizes or large tables. Moreover, even if we used a running total algorithm which did not require storing the estimators, the sheer number of additions and multiplications grows so fast that the quantities would not be computable. The fact that the sums are only computable with small sample sizes is actually not the

worst case scenario – the small sample size cases are precisely those where the resulting estimators are furthest away from their asymptotic distributions.

7.4 Monte Carlo simulation of theoretical quantities

Stochastic simulation provides a way around the combinatorial explosion of number of tables. The basic idea is that instead of enumerating all the possible tables, for every Pangloss point we only sample a fixed number K of tables, determine the bias and MSE norms for each table, average the K norms to find an estimate of the bias and MSE norm for that distribution, and then average the averages across the reduced model. This idea is formalized in Algorithm 7.2. The convergence of the simulations to the true values holds as a consequence of the law of large numbers.

Algorithm 7.2 Dr. Pangloss' method (Monte Carlo)

1: **input** sample size N, reduced model $\mathcal{M}_{r-1}^P = \left\{\boldsymbol{\pi}_j^P\right\}_{j=1}^J$, iterations K (500 here)
2: **output** average bias and mean squared error norm for $\mathcal{M}_{r-1}^P$
3: **for all** Pangloss points $\boldsymbol{\pi}_j^P$ **do**
4: sample $\{\boldsymbol{T}_k\}_{k=1}^K \sim \text{Multinom}_r(N, \boldsymbol{\pi}^{P_j})$
5: calculate $\{\widehat{\boldsymbol{\pi}}_{MLE}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\pi}}_{L2E}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\pi}}_{X^2}\}_{k=1}^K$
6: average $\left\{\widehat{\boldsymbol{\pi}}_\delta - \boldsymbol{\pi}^{P_j}\right\}_{k=1}^K$ and take norm to estimate the bias
7: average $\left\{(\widehat{\boldsymbol{\pi}}_\delta - \boldsymbol{\pi}^{P_j})(\widehat{\boldsymbol{\pi}}_\delta - \boldsymbol{\pi}^{P_j})'\right\}_{k=1}^K$ and take norm for an estimate of the MSE
8: **end for**
9: average bias and MSE norm estimates

7.5 Simulation results and discussion

The results from the simulations discussed in this chapter are contained in Tables 7.2 and 7.3. In those tables, the numbers listed in blue represent quantities calculated by the exact, deterministic Pangloss method described in Section 7.3. The numbers listed in red represent quantities determined by the stochastic Pangloss method (monte carlo simulation) described in Section 7.4. The models considered are the 2×2 independence model, the independence model in a 2×3 table, and the conditional independence model $X_1 \perp\!\!\!\perp X_2 | X_3$ with each variable binary. The results of this third model would be suitable for the data discussed in Example 3.7, for example. The sample sizes range from $N = 5$ to $N = 30$.

As we know from Chapter 4, each of the estimators is asymptotically unbiased. The simulations provide a perspective of what happens in finite – and in fact very small – sample sizes. Perhaps most striking in this area is that the minimum chi-squared statistic ($\widehat{\boldsymbol{\pi}}_{X^2}$) exhibits a strong bias when compared to the L2E and the MLE, which is exactly unbiased.[3] Thus, for the models considered the MLE is recommended over the minimum chi-squared statistic due to its relatively substantial bias, with the L2E following close behind the MLE.

A similar observation is made in terms of the mean squared error of the estimators. As a benchmark for comparison, we include the empirical relative frequencies ($\widehat{\boldsymbol{\pi}}_{EMP}$) in this part of the simulation and calculate its mean squared error norm in the same manner. Not surprisingly, the statistics which incorporate the additional model assumptions perform better in terms of their mean squared error. That is to say, their distribution is more concentrated that that of the empirical relative frequencies;

[3]The simulations are being checked to ensure the veracity of this result.

however, not drastically so. Again the MLE performs well in terms of mean squared error; however, it appears to be slightly outperformed by the minimum chi-squared statistic implying that while the minimum chi-squared statistic is more biased than the MLE, it's dispersion is smaller. Thus, the distinction between the MLE and the minimum chi-squared statistic follows the typical bias-variance tradeoff with the MLE being less biased but also less efficient while the minimum chi-squared statistic is both more biased and more efficient. The L2E, while being significantly less biased than the minimum chi-squared statistic, is more disperse than both the MLE and the minimum chi-squared statistic. The user is thus recommended to select the statistic which best suits the application at hand with these properties in mind.

Apart from the strict comparisons of the averages over the model, it is clear that the functions $||\boldsymbol{Bias}_\pi||_2$ and $||\boldsymbol{MSE}_\pi||_F$ vary substantially over the surface of the model, a fact illustrated by Figure 7.4. The fact that the functions vary is by itself not a surprising observation – we should expect that in cases near degeneracy that the estimators are quite stable (i.e. highly focused) while those away from it are more disperse. What is surprising is that there are locales away from degeneracy which exhibit lower bias. This effect is clearly seen in Figure 7.4 where the bias is lessened in the area surrounding the uniform distribution; a similar effect is seen in the MSE and the effect is observed in each of the statistics considered. A possible avenue for future investigation is to try to characterize these zones of relative low variability.

REMARK 7.1 Another consideration left for future work is that of robustness. While the L2E is outperformed in bias and mean squared error by the MLE, it is likely to be the case that it is more robust to outliers and contaminated data sets than the other estimators since this kind of behavior has been observed in other applications (Parr and Schucany [1980], Donoho and Liu [1988], Millar [1981]). Robustness is a

property of great interest for conditional independence models. While the conditional independence models are very expressive, they affect a substantial simplification of statistical model. This is observed to a small degree in the 2×2 independence model case, where the model is reduced from three parameters to two; however, the effect is even more pronounced in larger models, a fact exploited in graphical models where a table in several dimensions can be reduced to a model parameterized by a very small number of parameters (i.e., a low dimensional manifold in a high dimensional ambient space). If the L2E exhibits robustness properties in the setting of categorical conditional independence models (which is expected with some minor assumptions similar to those seen in Chapter 4), then it is indeed a very reasonable competitor to the MLE. The simulations demonstrate that the loss in bias and efficiency which come with the L2E are not great, and they would certainly be an acceptable price to pay for deviations from the strict experimental assumptions if the L2E does in fact enjoy robustness properties.

		$N = 5$	$N = 10$	$N = 20$	$N = 30$
$\widehat{\boldsymbol{\pi}}_{L2E}$	2×2	.0047	.0035	.0022	.0016
	2×3	.0066	.0056	.0080	.0062
	$2 \times 2 \times 2$	.0173	.0123	.0083	.0070
$\widehat{\boldsymbol{\pi}}_{MLE}$	2×2	.0000	.0000	.0000	.0000
	2×3	.0000	.0000	.0068	.0053
	$2 \times 2 \times 2$	.0155	.0109	.0074	.0062
$\widehat{\boldsymbol{\pi}}_{X^2}$	2×2	.0289	.0307	.0220	.0159
	2×3	.1252	.1120	.0212	.0168
	$2 \times 2 \times 2$	.0192	.0204	.0172	.0140

Table 7.2 : Mean bias L_2 norm

		$N = 5$	$N = 10$	$N = 20$	$N = 30$
$\widehat{\boldsymbol{\pi}}_{EMP}$	2×2	.0791	.0395	.0196	.0131
	2×3	.0790	.0394	.0197	.0132
	$2 \times 2 \times 2$	.0740	.0371	.0185	.0124
$\widehat{\boldsymbol{\pi}}_{L2E}$	2×2	.0763	.0380	.0188	.0126
	2×3	.0757	.0373	.0184	.0122
	$2 \times 2 \times 2$	.0713	.0357	.0177	.0118
$\widehat{\boldsymbol{\pi}}_{MLE}$	2×2	.0711	.0352	.0175	.0117
	2×3	.0692	.0335	.0165	.0110
	$2 \times 2 \times 2$	.0672	.0331	.0164	.0109
$\widehat{\boldsymbol{\pi}}_{X^2}$	2×2	.0692	.0343	.0172	.0115
	2×3	.1002	.0599	.0153	.0103
	$2 \times 2 \times 2$	.0637	.0304	.0152	.0102

Table 7.3 : Mean MSE Frobenius norm

Bibliography

A. Agresti. *Categorical data analysis.* Wiley-Interscience, 2002.

A. Agresti and C. A. Franklin. *Statistics : the art and science of learning from data.* Prentice Hall, 2007.

A. Antoniadis, J. Bigot, and T. Sapatinas. Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, 6(6): 1–83, 2001.

J. Berkson. Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, pages 457–487, 1980.

P. Billingsley. *Probability and measure.* Wiley Series in Probability and Mathematical Statistics, third edition, 1995.

M. W. Birch. A new proof of the Pearson-Fisher theorem. *The Annals of Mathematical Statistics*, 35(2):817–824, 1964. ISSN 0003-4851.

Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis.* Springer Verlag, 2007.

J. Bochnak, M. Coste, and M. F. Roy. *Real algebraic geometry.* Springer Verlag, 1998.

D. Böhning and H. Holling. On minimizing chi-square distances under the hypothesis

of homogeneity or independence for a two-way contingency table. *Statistics & Probability Letters*, 4(5):253–258, 1986.

R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

G. Casella and R. L. Berger. *Statistical inference.* Duxbury Advanced Series, second edition, 2002.

R. Christensen. *Log-linear models and logistic regression.* Springer Verlag, 1997.

D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms.* Springer, third edition, 2007.

J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, pages 522–539, 1980.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

A. P. Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, pages 598–617, 1980.

P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998.

D. L. Donoho and R. C. Liu. The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

J. L. Doob. Probability and statistics. *Transactions of the American Mathematical Society*, 36(4):759–775, 1934. ISSN 0002-9947.

C.A. Drossos and A.N. Philippou. A note on minimum distance estimates. *The Annals of Statistics*, 32(1):121–123, 1980.

M. Drton and S. Sullivant. Algebraic statistical models. *Statistica Sinica*, 17:1273–1297, 2007.

M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Birkhauser Basel, 2009.

D. Edwards. *Introduction to graphical modelling*. Springer, second edition, 2000.

D. Edwards and S. Kreiner. The analysis of contingency tables by graphical models. *Biometrika*, 70(3):553–565, 1983.

S. E. Fienberg and A. B. Slavkovic. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. *Data Mining and Knowledge Discovery*, 11(2):155–180, 2005.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.

D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *The Annals of Statistics*, 34(3):1463, 2006.

P. Gilbert and R. Varadhan. Bb: An r package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(i04), 2009.

R. K. S. Hankin. Programmers Niche : Multivariate polynomials in R. *R News*, 8/1, 2008.

Robin K. S. Hankin. *multipol: multivariate polynomials*, 2009. URL `http://CRAN.R-project.org/package=multipol`. R package version 1.0-4.

S. Hosten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Foundations of Computational Mathematics*, 5(4):389–407, 2005.

H. Hotelling. The consistency and ultimate distribution of optimum statistics. *Transactions of the American Mathematical Society*, pages 847–859, 1930. ISSN 0002-9947.

K. Kendig. *Elementary algebraic geometry.* Springer-Verlag, 1977.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951. ISSN 0003-4851.

M. Laurent. Polynomial optimization and sums of squares of polynomials. *Unpublished. Online [available] at http://homepages.cwi.nl/ monique/. Accessed - July 20, 2011*, 2003.

M. Laurent. Sums of squares, moment matrices and optimization over polynomials. *Emerging applications of algebraic geometry*, 149:157–270, 2009.

S. L. Lauritzen. *Graphical models.* Oxford University Press, 1996.

S. L. Lauritzen. *Lectures on contingency tables.* Aalborg University, electronic edition, 2002.

E. L. Lehmann and G. Casella. *Theory of point estimation.* Springer-Verlag, second edition, 2003.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses.* Springer-Verlag, third edition, 2005.

P. W. Millar. Robust estimation via minimum distance methods. *Probability Theory and Related Fields*, 55(1):73–89, 1981.

J. Mittmann. Gröbner bases: Computational algebraic geometry and its complexity. 2007.

D. S. Moore and G. P. McCabe. *Introduction to the practice of statistics.* W. H. Freeman and Company, second edition, 1998.

J. Neyman. Contribution to the theory of the X^2 test. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 239–273, 1949.

L. Pachter and B. Sturmfels. *Algebraic statistics for computational biology.* Cambridge Univ Pr, 2005.

W. C. Parr. Minimum distance estimation: a bibliography. *Communications in Statistics-Theory and Methods*, 10(12):1205–1224, 1981.

W. C. Parr and W. R. Schucany. Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624, 1980.

J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference.* Morgan Kaufmann San Mateo, Ca, 1988.

J. Pearl. *Causality : models, reasoning, and inference.* Cambridge University Press, 2000.

K. 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. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

R. Peck, C. Olsen, and J. L. DeVore. *Introduction to statistics and data analysis.* Duxbury, 2008.

C. R. Rao. *Linear statistical Inference and Its Applications.* John Wiley & Sons, Inc., 1965.

T. R. C. Read and N. A. C. Cressie. *Goodness-of-fit statistics for discrete multivariate data.* Springer, 1988.

S. I. Resnick. *A probability path.* Birkhäuser, 1999.

D. W. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285, 2001.

J. Shao. *Mathematical statistics.* Springer, second edition, 2003.

A. Slavkovic. *Statistical disclosure limitation beyond the margins: characterization of joint distributions for contingency tables.* PhD thesis, Ph. D. Thesis, Department of Statistics, Carnegie Mellon University, 2004.

R. D. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28(1):9–12, 1974.

A. J. Sommese, C. W. Wampler, and C. W. Wampler. *The numerical solution of systems of polynomials arising in engineering and science.* World Scientific Publishing Company Inc, 2005.

G. Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1973.

B. Sturmfels. *Solving systems of polynomial equations*. American Mathematical Society, 2002.

S. M. Sullivant. *Toric ideals in algebraic statistics*. PhD thesis, Ph.D. Thesis, Department of Mathematics, University of California, Berkeley, 2006.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Bill Venables. *PolynomF: Polynomials in R*, 2010. URL `http://CRAN.R-project.org/package=PolynomF`. R package version 0.94.

Bill Venables, Kurt Hornik, and Martin Maechler. *polynom: A Collection of Functions to Implement a Class for Univariate Polynomial Manipulations*, 2009. URL `http://CRAN.R-project.org/package=polynom`. R package version 1.3-6. S original by Bill Venables, packages for R by Kurt Hornik and Martin Maechler.

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007. URL `http://www.jstatsoft.org/v21/i12/paper`.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.