Markov Bases for Noncommutative Harmonic Analysis of Partially Ranked Data

Ann Johnston

Michael Orrison, Advisor

Matt Davis, Reader

May, 2011

Harvey Mudd College
Department of Mathematics

Abstract

Given the result v_0 of a survey and a nested collection of summary statistics that could be used to describe that result, it is natural to ask which of these summary statistics best describe v_0. In 1998 Diaconis and Sturmfels presented an approach for determining the conditional significance of a higher order statistic, after sampling a space conditioned on the value of a lower order statistic. Their approach involves the computation of a Markov basis, followed by the use of a Markov process with stationary hypergeometric distribution to generate a sample.

This technique for data analysis has become an accepted tool of algebraic statistics, particularly for the study of fully ranked data. In this thesis, we explore the extension of this technique for data analysis to the study of partially ranked data, focusing on data from surveys in which participants are asked to identify their top k choices of n items. Before we move on to our own data analysis, though, we present a thorough discussion of the Diaconis–Sturmfels algorithm and its use in data analysis. In this discussion, we attempt to collect together all of the background on Markov bases, Markov processes, Gröbner bases, implicitization theory, and elimination theory, that is necessary for a full understanding of this approach to data analysis.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Professor Michael Orrison for introducing me to this thesis topic and exploring it with me as my thesis advisor. Working with him and learning from him over the past four years has been an inspiration. I would also like to extend my thanks to Professor Matt Davis for his suggestions as my second reader, and to Aaron Pribadi and Jaron Kent-Dobias for their help in working with Mathematica and `4ti2`. Lastly, I would like to thank my high school calculus teacher, Jim Voytas, for introducing me to mathematics and for his continued friendship and support.

Chapter 1

Introduction

Participants in a survey can be asked to provide a variety of types of information. For example, participants can be asked to select from a list of items, those items of which they would approve, or to order the items according to their preferences. We call data that is obtained from these two types of surveys, respectively, approval data and fully ranked data. Data that is obtained from surveys in which participants do not give a full ranking of a list of items is grouped together under the term *partially ranked data*.

Given any such survey data, it is natural to try to summarize that data so as to interpret the result. This gives rise to many different summary statistics. For example, in most political elections (viewing these elections as surveys) the most natural summary statistic simply tells us how many votes each candidate receives. For fully ranked or partially ranked data, a summary statistic might list how many times each item in the list was ranked first, or how many times each pair of items was ranked in a certain order. In analyzing approval data, it might be natural to use a summary statistic that lists the number of times that each pair of items shows up in the approved subset. This statistic might be particularly appealing, for example, if the survey is intended to select a pair of individuals to chair a committee.

The possible results of each of these types of surveys can be viewed as $\mathbb{Z}$-valued functions on the set of all possible individual survey responses. For example, the result of a political election between n candidates is generally represented as an n-tuple of integers, with the ith entry corresponding to the number of votes that the ith candidate received. This n-tuple is a function from the set of all possible individual voter responses to $\mathbb{Z}$. Similarly, in surveys that yield approval data and ask for participants to choose from among a list of n items, individual respondents have a choice of $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ subsets of the items, and so each possible survey result takes the form of a 2^n-tuple of integers, which can be viewed as a $\mathbb{Z}$-valued function on the set of all possible individual survey responses.

Recalling that the collection of all $\mathbb{R}$-valued functions defined on a given set forms a vector space over $\mathbb{R}$, the result of a survey can be viewed as a $\mathbb{Z}$-valued element in the vector space of $\mathbb{R}$-valued functions defined on the set of all possible individual survey responses. In this way, there is a natural association between any survey and its *underlying vector space*. The standard basis for this vector space consists of the set of indicator functions on the possible individual survey responses.

With this framework, the procedures that compute summary statistics on a given survey result can be viewed as linear transformations that act on the survey's associated underlying vector space.

A natural question to ask, given the result of a survey and a collection of summary statistics that could be used to describe that result, is which of the statistics are actually essential in capturing the significant information in the data. That is, which of the summary statistics are necessary for describing trends in the responses to the survey?

1.1 The Significance of a Summary Statistic, and a Motivating Example

Analyzing Approval Voting Election Results

In the case of an approval voting election between n candidates, in which voters are asked to select a subset of the n candidates of which they would approve, the natural action of the symmetric group S_n on the subsets of an n-element set allows the associated underlying vector space to be viewed as an $\mathbb{R}S_n$-module. After decomposing the underlying vector space into orthogonal subspaces according to the number of approved candidates k, representation theory enables us, for each $0 \leq k \leq n$, to decompose the associated vector subspace into $k+1$ orthogonal isotypic subspaces, with respect to the action of S_n. For $0 \leq i \leq k$, these isotypic subspaces are called, respectively, the *pure ith order effects space*.

We focus on the subspace associated with a specific fixed k. This limits our attention to the responses of voters who approved of exactly k candidates. Here, the projection of an election result into the zeroth of these isotypic subspaces corresponds to the effect on the election result of the number of people who chose k candidates.

The projection of an election result into the first of these isotypic subspaces corresponds to the effect of the individual candidates on the election result after removing the effect of the number of people who chose k candidates—that is, after removing the pure zeroth order effects. If, in an approval election, voters tend to have a strong preference for individual candidates, then we might expect this pure first order effects statistic to be large. If, instead of removing these zeroth order effects, we consider the projection of an election result into both the zeroth and first isotypic subspaces of the underlying vector space, then we have a summary statistic that describes the *first order effects* in the election. This first order effects statistic simply counts the number of voters approving of each individual candidate.

Similarly, the summary statistic that counts the number of voters approving of each different pair of candidates is called the *second order effects* statistic, and consists of the projections of the election result into the pure zeroth, pure first, and pure second order effects spaces. If, in an approval election, the projection into the pure second order effects space is large, then we might expect the voters in the election to tend to have a strong preference for certain pairs of candidates. If this is the case, then neither the zeroth nor the first order effects statistic is sufficient to capture all of the information present in the election result.

This generalizes to the *ith order effects statistic*, for $0 \leq i \leq k$, which counts the number of voters approving of each different i-subset of the n candidates, and consists of the projections of the election result into the pure zeroth through pure ith order effects spaces. If the voters in an election have a strong preference for certain i-subsets of the candidates, then this will show up as the election result having a large projection into the pure ith order effects space, in which case, we know that the ith order effects statistic captures information about the election that is not captured by any of the lower order effects statistics.

Thus, for our fixed k, the collection of ith order effects statistics, with $0 \leq i \leq k$, forms a nested collection of summary statistics that indicate underlying trends in the will of the voters. As such, it is natural to use these summary statistics in analyzing the result of an approval voting election, and to attempt to determine the largest order effects statistic that carries significant information.

For a more thorough discussion of the analysis of approval voting data, see Hansen and Orrison (2008). The above discussion generalizes naturally to surveys that yield approval data. Here, we have focused on elections, because our discussion has been guided by that in Hansen and Orrison (2008), which focuses on election data. We more thoroughly present the summary statistics used in this analysis of approval data, in our discussion of partially ranked data found in Chapter 5.

The Question of Conditional Significance

In the discussion above, we have referred to the significance of the ith order effects statistic for a survey result in terms of the magnitude of the projection of the result into its pure ith order effects space. If the magnitude of this projection is large, then we might consider the ith order effects statistic to be significant. However, we must ask whether the magnitude of this projection, alone, is enough to indicate the significance of the ith order effects statistic, or whether it is possible that the magnitude of the projection into the pure ith order effects space could just be a residual effect of one of the lower order statistics.

To clarify the question being asked, we return to our example of an approval voting election. Consider an approval election between n candidates with result v_0, and fix $0 \leq k \leq n$. Let $t_0, \ldots, t_k$, be the ith order effects statistics associated with this election, respectively, for $0 \leq i \leq k$. Further, let t_j and t_ℓ, with $1 \leq j < \ell \leq k$, be much larger in magnitude than all other t_i. Does this indicate that t_ℓ contains significant information about v_0, or is it possible that v_0 is best described by t_j, and that t_ℓ is large only as a result of t_j? That is, does the statistic t_ℓ have significance, after we have conditioned on the lower order statistic t_j?

It may be, if we consider all possible results of the same election that produced v_0 having the summary statistic t_j in common, that we find most of these election results also have large projections into the pure ℓth order effects space. If this is the case, then t_ℓ, despite its large magnitude, does not seem to capture information about v_0 that was not already captured by t_j, and so we might say that t_ℓ is an artifact of t_j rather than having its own significance.

Parallel questions arise in the analysis of other types of survey data, with their own natural collections of nested summary statistics. Given the result of a survey and a nested collection of summary statistics that could be used to describe that result, a means is desired for determining whether or not a given summary statistic captures essential information about the result. That is, given a summary statistic t, we would like to be able to determine t has significance even after conditioning on any of the lower order statistics.

In practice, given the result of a survey and an associated collection of summary statistics that could be used to describe that result, the set of all possible survey results that share a specific summary statistic is very large. Thus, in general, it is not reasonable to consider the magnitudes of the projections associated with each of the different summary statistics on this entire set. In their 1998 paper, Diaconis and Sturmfels provide a method for avoiding this difficulty, by presenting a Markov chain that can be used to generate a representative sample of the set of all possible survey results that share a given summary statistic. It is reasonable to compute, on the elements of this representative sample, the magnitudes of each of the different summary statistics for the survey. The result of these computations can then be used in analyzing the conditional significance of the different summary statistics.

1.2 Goals and Overview of this Thesis

The algorithm that Diaconis and Sturmfels present for generating a Markov basis (Theorem 3.2 of Diaconis and Sturmfels (1998)) has become an accepted tool in algebraic statistics.

In Diaconis and Eriksson (2006), the authors explore its usefulness in the analysis of fully ranked data, letting their analysis be driven by the isotypic decomposition of the associated underlying vector space. In Hansen and Orrison (2008), the use of this algorithm in the analysis of approval data is explored, focusing again on isotypic subspaces. This algorithm's usage is also discussed extensively in Riccomagno (2009), where Riccomagno refers to it as "the Diaconis–Sturmfels algorithm". In this thesis, I further explore the technique for data analysis that was first introduced in Diaconis and Sturmfels (1998), with particular focus on partially ranked data. Again, I primarily focus on the summary statistics associated with projections into the isotypic subspaces of the underlying vector space.

The usefulness of the Diaconis–Sturmfels algorithm in data analysis draws upon a large body of theory, relying the theory of Markov processes and on Gröbner basis techniques and implicitization and elimination theory from algebraic geometry. The papers that I have found discussing the Diaconis–Sturmfels algorithm, with the exception of Riccomagno (2009), have made little or no attempt at collecting together and clearly presenting this underlying theory. The first goal for my thesis has thus been to collect together and clearly present the supporting theory for the Diaconis–Sturmfels algorithm, and then to identify and clarify this algorithm, as well as its usage. I have attempted, in doing so, to improve upon the straightforwardness and clarity of the discussion of this algorithm that can be found in Riccomagno (2009). The first half of my thesis has been devoted to this effort.

In Chapter 2, we present the theory of Markov bases for a space conditioned on a summary statistic, as well as some terminology and theory for Markov processes. We then discuss the use of a Markov basis in generating a representative sample for its associated conditioned space, presenting the Markov chain that was introduced by Diaconis and Sturmfels (1998). In Chapter 3, we present sufficient algebraic geometry for an understanding of implicitization and elimination theory. After discussing the implicitization and elimination theorems, we discuss their use in generating a Markov basis for a space conditioned on a summary statistic, as it was introduced in Diaconis and Sturmfels (1998).

Then, in Chapter 4, we pull together the theory from the previous two chapters to present a coherent picture of Diaconis' and Sturmfels' approach to data analysis. In a paper assuming prior knowledge of Markov processes and of the Diaconis–Sturmfels algorithm and its supporting theory, the content of Chapter 4 could be found in an introduction. In my thesis, this chapter serves as a transition into the discussion of our own work in analyzing partially ranked data, which can be found in Chapter 5.

Chapter 2

Markov Bases and Markov Processes for Generating a Representative Sample

Consider a survey result v_0, with summary statistic t that is computed from v_0 via the linear transformation T. In Diaconis and Sturmfels (1998), the authors present a Markov process for generating a representative sample of the space conditioned on t. It is natural to ask what information from v_0 and t is necessary in carrying out this Markov process. The definition of a Markov basis captures this necessary information.

In this chapter, we first present the definitions of a Markov basis and a Markov process. We then discuss their use in generating a representative sample of the space conditioned on a summary statistic, as it was presented in Diaconis and Sturmfels (1998), focusing on the Metropolis algorithm that they present. The notation that we use has primarily been taken from that paper.

2.1 Markov Bases

Let v_0 be the result of a survey, and let t be a summary statistic for v_0. Additionally, let $\mathcal{X}$ be the set of all possible responses to the survey that produced the result v_0. Further, let $V = L_{\mathbb{R}}(\mathcal{X})$ be the underlying vector space associated with this survey, and let T be the linear transformation on V that computes t from v_0, so that $T(v_0) = t$. Here, we are letting $L_{\mathbb{R}}(\mathcal{X})$ denote the collection of $\mathbb{R}$-valued functions defined on $\mathcal{X}$. In general, we take $\mathcal{X}$ to be finite.

Example 2.1.1. Consider a survey in which participants are asked to identify their top two choices of four items. For this survey, the set $\mathcal{X}$ consists of the $\binom{4}{2} = 6$ possible subsets of the four items, and so the underlying vector space V is 6-dimensional. We can record the result of this survey in a vector

$$v_0 = (a_1, a_2, a_3, a_4, a_5, a_6),$$

where each a_i is the number of survey responses that chose the ith pair of items. A summary statistic t for v_0 might, for example, count the number of survey responses, or the number of times that each item shows up in the survey responses. Both of these statistics can be computed via linear transformations on the underlying vector space V.

Let $\mathcal{F}$ be the set of all possible results of this survey (so $v_0 \in \mathcal{F}$), and let the set of all $v \in \mathcal{F}$ with $T(v) = t$ be denoted $\mathcal{F}_t$. We would like to generate a representative sample of $\mathcal{F}_t$ to use in an analysis of the conditional significance of other summary statistics for v_0, after conditioning on t.

Definition 2.1.1. Let V be the underlying vector space associated with a result v_0 of a survey, and let T be a linear transformation on V, with $T(v_0) = t$. A *Markov basis* for $\mathcal{F}_t$ is a set of vectors $f_1, \ldots, f_L$ in $L_{\mathbb{Z}}(\mathcal{X}) \subset V$ such that

(a) $T(f_i) = 0, \qquad \text{for } 1 \leq i \leq L, \quad \text{and}$

(b) for any $u, v \in \mathcal{F}_t$ there exist ordered pairs $(\epsilon_1, f_{i_1}), \ldots, (\epsilon_A, f_{i_A})$

with $\epsilon_j = \pm 1$ for $1 \leq j \leq A$, and with $\quad v = u + \sum_{j=1}^{A} \epsilon_j f_{i_j},$

where for $1 \leq \alpha \leq A$, we have $u_\alpha = u + \sum_{j=1}^{\alpha} \epsilon_j f_{i_j} \in \mathcal{F}.$

Remark 2.1.1. Notice that for every $u, u' \in \mathcal{F}_t$ we can write $u' = u + f$ for some $f \in \mathcal{F}$. Then, because T is a linear transformation, we have

$$T(f) = T(u' - u) = T(u') - T(u) = t - t = 0. \tag{2.1}$$

Thus, viewing f as a "move" that keeps us in $\mathcal{F}_t$, a Markov basis can be viewed as a set of "moves" such that any two elements of $\mathcal{F}_t$ can be connected via a finite number of "moves". This is the idea that Diaconis and

Eriksson focus on in their definition of a Markov basis (Diaconis and Eriksson, 2006: see Definition 5), in contrast to the definition that we give, which comes more directly from Diaconis and Sturmfels (1998: see Equation 1.5).

Our requirement in (b), that $u_\alpha \in \mathcal{F}$, differs in form from the requirement given in (1.5)(b) of Diaconis and Sturmfels (1998), which requires that the entries of each u_α be nonnegative with respect to standard basis of the indicator functions on $\mathcal{X}$. Both the condition that we have given and the condition given in (1.5)(b) of Diaconis and Sturmfels (1998) are equivalent to the requirement that at each "step" from u to v, the result u_α is a valid survey result. Our definition is simply independent of the choice of basis for the underlying vector space V.

Notice that the vectors in a Markov basis need not be linearly independent.

In general, we can think of a Markov basis for $\mathcal{F}_t$ as conditioning on the summary statistic matrix T, noticing that Definition 2.1.1 depends in no way on the specific value of t. In this way, the Markov basis associated with a specific T and V only needs to be computed once. This fact is particularly valuable, because the computation of a Markov basis using the Diaconis–Sturmfels algorithm (which we discuss in Chapter 3) turns out to be very computationally intensive.

Notice that if the number of participants in a survey is fixed, then the finiteness of $\mathcal{X}$ gives the finiteness of $\mathcal{F}$, and so also of $\mathcal{F}_t \subset \mathcal{F}$. Under this restriction, the set of differences of elements in $\mathcal{F}_t$ is finite, and each of these differences must be zero-valued under T. It is straightforward to show that this finite collection of differences forms a Markov basis for $\mathcal{F}_t$. In this way, if the number of participants in a survey is fixed, the existence of a finite Markov basis for $\mathcal{F}_t$ is guaranteed.

Also, notice that a Markov basis for $\mathcal{F}_t$ is trivial if and only if $\mathcal{F}_t$ consists of a single element. This corresponds to the case where the linear transformation T is injective. Particularly, if T is the identity operator on V, then the Markov basis for $\mathcal{F}_t$ is trivial.

Example 2.1.2. Consider the survey from Example 2.1.1. Let the four items in this survey be labeled $1, 2, 3, 4$, and let the 2-subsets of these elements be ordered according to the list

$$(\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}).$$

The summary statistics described in Example 2.1.1 are, respectively, computed by the matrices

$$T_0 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad T_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The Markov bases for conditioning on these summary statistics are, respectively,

$$\begin{aligned} \mathcal{B}_0 = \{&(1,-1,0,0,0,0),(1,0,-1,0,0,0),\\ &(1,0,0,-1,0,0),(1,0,0,0,-1,0),\\ &(1,0,0,0,0,-1)\}, \end{aligned}$$

and

$$\mathcal{B}_1 = \{(1,-1,0,0,-1,1),(1,0,-1,-1,0,1)\}.$$

We will see in Section 2.3 that a Markov basis for $\mathcal{F}_t$ can be used to define a Markov process, which can then be used to generate a representative sample of $\mathcal{F}_t$. This leaves the question of an algorithm for generating a Markov basis for F_t. The Diaconis–Sturmfels algorithm satisfies this purpose, as we will see in Chapter 3.

2.2 Markov Processes

Let $\{X_n\}$ be a collection of random variables that each take on values in a given space. We call this collection of X_n a *system*, and we call the space in which these X_n take values the *state space* of the system. We call n the *index parameter* of the system, and we interpret X_n as the "state of the system at time n". If the index parameter n takes on values in the natural numbers, then we say that the system is a *discrete time process*, whereas, if n takes on nonnegative real values, then we say that the system is a *continuous time process*. In this discussion, we will limit our attention to Markov processes that are time-independent.

Example 2.2.1. Consider the random variable X associated with the possible results of a given survey. Taking a countable number of copies of X and indexing them by $\mathbb{N}$, we get a collection of random variables $\{X_n\}$. This collection of random variables forms a system indexed by $\mathbb{N}$, and so, it is a discrete time process. The state space for X_n is the space $\mathcal{F}$ of all possible survey results.

A system $\{X_n\}$ is said to be a *Markov process* if it is a "memoryless" system. By this we mean that, given the value of X_n (the "present state of the system"), the values of X_s with $s > n$ (the "future states of the system"), are independent of the values of X_r with $r < n$ (the "previous states of the system"). If, further, the system $\{X_n\}$ has a countable state space, then this Markov process is called a *Markov chain*. This notion is formalized below, in Definition 2.2.1, which we have taken from Example 1.3c of Karlin and Taylor (1975) and Section 2.1 of Karlin and Taylor (1975).

Notice that in Example 2.2.1, where we take each X_n to be associated to a survey result, the state space is countable. We thus choose to focus primarily on discrete time Markov chains, because we will use such a Markov chain in Section 2.3 to generate a representative sample of $\mathcal{F}_t$.

Definition 2.2.1. A collection of random variables $\{X_n\}$ is a *discrete time Markov chain* if it is a Markov process whose state space is a countable set and whose index parameter n takes on values in $\mathbb{N}$.

A collection of random variables $\{X_n\}$ is a *Markov process* if, for all n, we have

$$\begin{aligned}\Pr\{a < X_n \leq b | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m\} \\ = \Pr\{a < X_n \leq b | X_{t_m} = x_m\}\end{aligned}$$

whenever $t_1 < t_2 < \cdots < t_m < n$.

Here, we are assuming that the random variables X_n take on values in an ordered set, so that we can write $a < X_n \leq b$. If the state space of X_n is countable, then we can rewrite the condition in Definition 2.2.1 as

$$\begin{aligned}\Pr\{X_n = b | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_m} = x_m\} \\ = \Pr\{X_n = b | X_{t_m} = x_m\},\end{aligned}$$

for all n, whenever $t_1 < t_2 < \cdots < t_m < n$.

Observe that this definition of a Markov chain makes the the idea of a *transition probability function* well-defined. That is, given a Markov chain with state space S, and given states i and j in S, and $n \in \mathbb{N}$, the function

$$P(i, j, n) := \Pr(X_{m+n} = j | X_m = i)$$

is a well-defined function in i, j, n.

The definition of a Markov chain given in Section 4.1 of Ross (1996) has a much stronger focus on these transition probabilities. It states that a collection of random variables $\{X_n\}$ is a *Markov chain* if, for all $n \geq 0$, and for all $i_0, \ldots, i_{n-1}, i, j$ elements of the state space of X_n, we have

$$\Pr\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P_{ij},$$

where P_{ij} is the fixed probability that the next state of the system will be j when the current state of the system is i.

Generally, the study of a given Markov chain reduces to the study of the associated transition probability function defined on its state space.

When the state space of a Markov chain is very large, and we would like to use a Markov chain to generate a representative sample of its state space, the existence of a Markov basis for its state space greatly simplifies the sampling process. Given a Markov basis for the state space S, rather than calculating the transition probabilities for each pair in S and each $n \in \mathbb{N}$, we need only calculate the transition probabilities $P(s, s \pm f_i, 1)$, with $s \in S$ and f_i an element of the given Markov basis. This collection of transition probabilities is sufficient for generating a representative sample for S (see Section 2.3).

Communicating Classes, Irreducibility, and Aperiodicity

We now present some terminology for Markov chains that will be useful in Section 2.3. We are taking these definitions essentially directly from Sections 4.2–4.7 of Ross (1996).

Definition 2.2.2. Let S be the state space for a Markov chain, and let P_{ij}^n denote the probability, given that the chain starts in state i, that the chain will be in state j after n transitions, where $n \geq 0$, or more formally,

$$P_{ij}^n = \Pr\{X_{n+m} = j | X_m = i\}, \quad \text{where} \quad n \geq 0, \quad i, j \geq 0.$$

Then, for $i, j \in S$, we say that state j is *accessible* from state i if there exists some $n \in \mathbb{N}$ for which $P_{ij}^n > 0$, and we say that states i and j *communicate* if i and j are mutually accessible.

Notice that under this definition, communicating is an equivalence relation on the state space of a Markov chain, and we call the associated equivalence classes *communicating classes* (see Proposition 4.2.1 of Ross (1996)).

Definition 2.2.3. A Markov chain is said to be *irreducible* or *connected* if its state space S consists of a single communicating class. That is, if i and j communicate with each other for every $i, j \in S$, or more formally, if for every $i, j \in S$ there exists an $n \in \mathbb{N}$ such that $P_{ij}^n > 0$.

Definition 2.2.4. Given a Markov chain with state space S, we say that a state $i \in S$ is *periodic* with period d, if d is the greatest integer such that $P_{ii}^n = 0$ whenever d does not divide n. We say that a state i is *aperiodic* if it is periodic with period 1. If for all $i \in S$, state i is aperiodic, then we say that the Markov chain with state space S is *aperiodic*.

Given a Markov chain with state space S and state $i \in S$, it is helpful in understanding Definition 2.2.4, to notice that state i is periodic with period d, if upon starting at state i, the Markov chain has a nonzero probability of returning to state i only after a number of transitions that is a multiple of d. In this way, we can view state i as being aperiodic if and only if, upon starting at state i, the Markov chain returns to i at "irregular time intervals", or more formally, if and only if

$$\gcd\{n : P_{ij}^n > 0\} = 1.$$

By Proposition 4.2.2 of Ross (1996), periodicity is a communicating class property. In this way, if any element in the state space of an irreducible Markov chain is aperiodic, then the entire Markov chain is aperiodic.

Stationary Distributions and Reversible Markov Chains

The next few definitions and remarks build up terminology for Definition 2.2.9, the definition of a *reversible* Markov chain. We will use this definition in Section 2.3.

Definition 2.2.5. Let S be the state space for a Markov chain, and let $i, j \in S$. We define f_{ij}^n to be the probability, given that the Markov chain starts in state i, that the Markov chain transitions into state j for the first time after exactly n transitions. Additionally, we define

$$f_{ij} := \sum_{n=1}^{\infty} f_{ij}^n,$$

so that f_{ij} denotes the probability of the Markov chain ever making a transition into state j, given that the the Markov chain began in state i. If $f_{jj} = 1$, for $j \in S$, then we say that state j is *recurrent*. Otherwise, we say that state j is *transient*.

Note that by Corollary 4.2.4 of Ross (1996), being recurrent is a property of communicating classes. That is, if i and j communicate in a Markov chain, then i and j are either both recurrent or both transient.

Definition 2.2.6. Consider a Markov process with state space S. Define, for $j \in S$

$$\mu_{jj} := \begin{cases} \infty & \text{if } j \text{ is transient,} \\ \sum_{n=1}^{\infty} n f_{jj}^{n} & \text{if } j \text{ is recurrent,} \end{cases}$$

so that μ_{jj} denotes the expected number of transitions needed, upon starting at state j, for the Markov process to return to state j. Given a recurrent state $j \in S$, we say that j is *positive recurrent*, if $\mu_{jj} < \infty$. If $\mu_{jj} = \infty$ then we say that state j is *null recurrent*.

Notice that under this definition, a state j in the state space of a Markov chain is positive recurrent if and only if, upon starting at state j, the Markov chain is expected to return to state j within a finite number of transitions.

By Proposition 4.3.2 of Ross (1996), positive recurrence and null recurrence are both communicating class properties.

Definition 2.2.7. Consider a Markov chain $\{X_n\}$ with state space S, where the elements in S are indexed by $\mathbb{N}$. Let P_{ij} denote the transition probabilities of this Markov chain on this state space S, and let $\{\pi_j : j \in \mathbb{N}\}$ be a probability distribution defined on the random variable X_n. π_j is said to be *stationary* if, for all $j \in \mathbb{N}$,

$$\pi_j = \sum_{i \in \mathbb{N}} \pi_i P_{ij}.$$

It can be shown (Ross, 1996: pg. 174) that if the probability distribution on X_0 in a Markov chain is a *stationary distribution* for that Markov chain, then for all X_n in the Markov chain, $n \geq 0$, the distribution on X_n is this same stationary distribution. Thus, the use of the term "stationary" in Definition 2.2.7 is natural.

Given an irreducible, aperiodic Markov chain with state space S indexed by $\mathbb{N}$, we define the function π_j on the random variable X_n, for $j \in \mathbb{N}$, by

$$\pi_j := \lim_{n \to \infty} P_{ij}^{n}, \quad \text{for } i \in \mathbb{N}.$$

It can be shown (see Theorem 4.3.3 of Ross (1996)) that if the states in S are positive recurrent, then π_j is positively valued, and further, that this π_j is the *unique* stationary distribution for this irreducible, aperiodic Markov process.

Definition 2.2.8. Consider an irreducible and positive recurrent Markov chain. We say that this Markov chain is *stationary*, if its initial state is chosen according to the stationary distribution for this Markov chain. That is, a Markov chain is stationary if the distribution on X_0 is the stationary distribution of the Markov chain.

Remark 2.2.1. Let v_0 be the result of a survey with summary statistic t, and let the set of all of possible results of this survey be $\mathcal{F}$. If we assume that the distribution on $\mathcal{F}$ is hypergeometric, and if we define an irreducible and positive recurrent Markov chain with initial state v_0 and state space $\mathcal{F}$, and with hypergeometric stationary distribution, then notice that we have defined a stationary Markov chain. Thus, at each step X_n in the Markov chain, the distribution on X_n is hypergeometric. In this way, if we start at v_0 and take sufficiently many steps along this Markov chain, and then repeat this process many times, the collection of the final states from these runs should form a representative sample of $\mathcal{F}$. If we restrict our moves in this Markov chain, to preserve the value of the summary statistic t, then the result should instead be a representative sample of $\mathcal{F}_t$. We will see in Section 2.3, that we can use a Markov basis for $\mathcal{F}_t$ to define such a Markov chain.

Consider a stationary Markov chain $\{X_n\}$ with state space S, having transition probabilities P_{ij} and stationary distribution π_i. Ross (1996: pg. 203) shows that the reverse process, starting at some $n \geq 0$ and taking the sequence of states $X_n, X_{n-1}, \ldots$, is in fact a Markov chain, with transition probabilities

$$P^*_{ij} = \frac{\pi_j P_{ji}}{\pi_i}.$$

This fact leads naturally to Definition 2.2.9, as well as to the idea that follows it. Both of these are taken essentially directly from page 203 of Ross (1996).

Definition 2.2.9. A stationary Markov chain with state space S, having transition probabilities P_{ij} and stationary distribution π_j, for $i, j \in S$, is said to be *time reversible* or *reversible*, if $P^*_{ij} = P_{ij}$ for all $i, j \in S$, where we define

$$P^*_{ij} := \frac{\pi_j P_{ji}}{\pi_i}.$$

This definition gives the requirement that, in order for a Markov chain with state space S to be reversible, it must satisfy

$$\pi_i P_{ij} = \pi_j P_{ji},$$

for all $i, j \in S$. We can view $\pi_j P_{ji}$ as the "rate" at which the Markov chain goes from state j to state i, recalling that

$$\Pr\{X_m = j\} = \pi_j \qquad \text{and} \qquad P_{ji} = \Pr\{X_{m+1} = i | X_m = j\}.$$

for $m \geq 0$ and $i, j \in S$. Similarly, we can view $\pi_i P_{ij}$ as the "rate" at which the Markov chain goes from state i to state j. Thus, the given requirement for the reversibility of a Markov chain can be viewed as the requirement that the Markov chain must proceed from state i to state j at the same rate that it proceeds from state j to state i.

2.3 Generating a Representative Sample Space

With the idea of a Markov basis for $\mathcal{F}_t$ from Section 2.1, and with the definitions and terminology of Markov processes from Section 2.2, we can now address the use of a Markov chain to generate a representative sample of $\mathcal{F}_t$.

Section 2 of Diaconis and Sturmfels (1998) presents two methods for setting up a Markov chain that can be used to generate a representative sample of $\mathcal{F}_t$, given a Markov basis for $\mathcal{F}_t$. The first of these, presented in their Lemma 2.1, is a Metropolis algorithm, and the second of these, presented in their Lemma 2.2, is a Gibbs sampler. We limit our focus to the Metropolis algorithm, as this is the method that we have used in our own analysis, which we present in Chapter 5.

The Metropolis Algorithm

We first present Diaconis and Sturmfels's Lemma 2.1, and then observe that, as it claims to do, this lemma defines a Markov chain with stationary distribution proportional to σ, according to the standard Metropolis algorithm. Thus, with appropriate choice of σ, it is possible to use this Markov chain to generate a representative sample of $\mathcal{F}_t$.

Lemma 2.3.1. *(Lemma 2.1 from Diaconis and Sturmfels (1998)) Let σ be a positive function on $\mathcal{F}_t$. Given a Markov basis $f_1, \ldots, f_L$ for $\mathcal{F}_t$, generate a Markov chain on $\mathcal{F}_t$ by choosing I uniformly in $\{1, \ldots, L\}$ and $\epsilon = \pm 1$ with probability $\frac{1}{2}$ independent of I. If the chain is currently at $g \in \mathcal{F}_t$, it moves to $g + \epsilon f_I$ with probability*

$$\min\left\{\frac{\sigma(g + \epsilon f_I)}{\sigma(g)}, 1\right\},$$

provided $g + \epsilon f_I$ has only nonnegative entries (i.e., is a legal survey result). In all other cases the chain stays at g. This is a connected, reversible, aperiodic Markov chain on $\mathcal{F}_t$ with stationary distribution proportional to σ.

In general, the Metropolis algorithm takes the following form (see Section 10.2 of Ross (1997)):

> **Metropolis Algorithm:** Let S be a state space, and let $i \in S$ be the current state, and $j \in S$ be a proposed new state. Let $P(x)$ be a probability distribution on S, and let Q be an irreducible Markov transition probability matrix on S, with $Q(i,j)$ representing the entry in Q at row i and column j. Choose α uniformly on $(0,1)$. Then, if
>
> $$\alpha < \min\left\{\frac{P(j)Q(i,j)}{P(i)Q(j,i)}, 1\right\} \tag{2.2}$$
>
> we accept the new state j. Otherwise, we remain at state i.

Notice that in this more general Metropolis algorithm, if $Q(i,j) = Q(j,i)$, then Equation 2.2 becomes

$$\alpha < \min\left\{\frac{P(j)}{P(i)}, 1\right\}. \tag{2.3}$$

Further, if we take the distribution from which we are sampling to be $P = \gamma\sigma$, where γ is a scalar that causes the range of P to be $[0,1]$ (with σ defined as in Lemma 2.3.1), then Equation 2.3 becomes

$$\alpha < \min\left\{\frac{\gamma\sigma(j)}{\gamma\sigma(i)}, 1\right\} = \min\left\{\frac{\sigma(j)}{\sigma(i)}, 1\right\}.$$

Also notice that accepting j when uniformly chosen α on $(0,1)$ is less than $\min\left\{\frac{\sigma(j)}{\sigma(i)}, 1\right\}$ is equivalent to accepting j with probability $\min\left\{\frac{\sigma(j)}{\sigma(i)}, 1\right\}$. Thus, if we take $j = g + \epsilon f_I$ and $i = g$, we see that Lemma 2.3.1 does indeed present a specific case of the general Metropolis algorithm. Here, the choice of a proposed new state is determined by the choice of a Markov basis element such that the resulting state is a legal survey result.

According to Section 10.2 of Ross (1997), it is known that this general Metropolis algorithm produces a connected, reversible Markov chain with state space S and stationary distribution $P(x)$. In this way, we see that the Markov chain defined in Lemma 2.3.1 has stationary distribution $\gamma\sigma$, which is indeed proportional to σ, as the lemma claims.

The Choice of a Hypergeometric Stationary Distribution

In the remark following Lemma 2.1 in Diaconis and Sturmfels (1998), the authors present a class of functions that they have found useful in defining positive functions σ (from Lemma 2.3.1) for the examples presented in the paper. Letting $g \in \mathcal{F}_t$, the authors point out that if

$$\sigma(g) = \prod_{x \in \mathcal{X}} \frac{1}{g(x)!},$$

then σ is distributed hypergeometrically. Thus, with this choice of σ, the Markov chain defined in Lemma 2.3.1 has hypergeometric stationary distribution.

There is a strong emphasis in Diaconis and Sturmfels (1998) on the choice of a Markov chain with a hypergeometric stationary distribution. This results from the fact that we would like to generate a representative sample of $\mathcal{F}_t$, under the assumption that the distribution on the individual participant responses to the survey is uniform. That is, we would like to compare our result v_0 to $\mathcal{F}_t$, assuming that when each participant completes the survey, their response is uniformly random from among all possible survey responses.

Just as when two dice are rolled the outcome of their combined roll being a seven or a two is not equally likely, this assumed uniformity in the responses of the individual survey participants does not carry over to a uniformity in the distribution on the possible survey results. Rather, according to Section 1 of Diaconis and Sturmfels (1998), this assumed uniformity in the responses of the individual survey participants translates to survey results that are distributed hypergeometrically. Thus, in our analysis of v_0, it is natural to assume that the distribution on the set of all possible survey results $\mathcal{F}$ is hypergeometric.

Recall from Remark 2.2.1, that under the assumption that $\mathcal{F}$ is hypergeometrically distributed, if we have a Markov chain with hypergeometric stationary distribution and Markov basis conditioning on $\mathcal{F}_t$, then we can use this Markov chain to generate a representative sample of $\mathcal{F}_t$.

The method that we have used in Chapter 5 to generate a representative sample of $\mathcal{F}_t$ follows precisely the procedure that we have outlined in Remark 2.2.1.

Chapter 3

Implicitization and Elimination Theory for Generating a Markov Basis

We have seen in Chapter 2 that, given a survey result v_0 with summary statistic t, and given a Markov basis for $\mathcal{F}_t$, we can carry out a Markov process to generate a representative sample of $\mathcal{F}_t$. Let us now consider the method for generating such a Markov basis, given the result of a survey and an associated summary statistic.

The method that we present was first introduced by in Diaconis and Sturmfels (1998), and is referred to in Riccomagno (2009) as the Diaconis–Sturmfels algorithm. This algorithm relies deeply on implicitization and elimination theory from algebraic geometry, which themselves rely deeply on Gröbner bases. In this chapter, we thus first present the necessary background on these subjects, before introducing the Diaconis–Sturmfels algorithm. The theory that we present in Sections 3.1 and 3.2 can be found in Chapters 1–3 of Cox et al. (1997).

3.1 Gröbner Bases

Let k be a field and let $x_1, \ldots, x_n$ be indeterminates. Consider the polynomial ring $k[x_1, \ldots, x_n]$. If $I \subseteq k[x_1, \ldots, x_n]$ is an ideal, and $f \in k[x_1, \ldots, x_n]$ is a polynomial, a natural question to ask is whether f is an element of I.

In a polynomial ring $k[x]$, letting $f_1, \ldots, f_m$ and g be polynomials in $k[x]$, the division algorithm (Proposition 1.5.2 in Cox et al. (1997)) guarantees the existence of $a_1, a_2, \ldots, a_m, r \in k[x]$ such that

$$g = a_1 f_1 + a_2 f_2 + \cdots + a_m f_m + r, \tag{3.1}$$

where either $r = 0$ or the polynomial degree of r is less than the polynomial degree of g. Further, it guarantees that these r and a_i, for $1 \leq i \leq m$, are unique. In this way, we see that if $f_1, \ldots, f_m$ are the generators of an ideal I, then $g \in I$ if and only if we have $r = 0$ in Equation 3.1.

We can extend this division algorithm to rings of polynomials in more than one variable by defining a *monomial ordering* (Definition 2.2.1 in Cox et al. (1997)).

Definition 3.1.1. Given a monomial $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \in k[x_1, \ldots, x_n]$, we define the notation

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} := x^{\alpha},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{N}^n$ (where we take $0 \in \mathbb{N}$). A *monomial ordering* on $k[x_1, \ldots, x_n]$ is a relation $>$ on the set of monomials $x^{\alpha} \in k[x_1, \ldots, x_n]$, such that, for $\alpha, \beta \in \mathbb{N}^n$,

$$x^{\alpha} > x^{\beta}$$

if and only if $\alpha \succ \beta$, where $\succ$ is a total (or linear) ordering on $\mathbb{N}^n$ that satisfies the following two properties:

(i) If $\alpha \succ \beta$ and $\gamma \in \mathbb{N}^n$, then $\alpha + \gamma \succ \beta + \gamma$.

(ii) $\succ$ is a well-ordering on $\mathbb{N}^n$, and so every nonempty subset of $\mathbb{N}^n$ has a smallest element under $\succ$.

Example 3.1.1. The *Lexicographic order* is a monomial ordering, where for $\alpha, \beta \in \mathbb{N}^n$ we say that $x^{\alpha} > x^{\beta}$ if, in the vector difference $\alpha - \beta$, the leftmost nonzero entry is positive (see Proposition 2.2.4 of Cox et al. (1997)). For example, under the Lexicographic order

$$x^{(1,2,0)} > x^{(0,3,4)}, \quad \text{because } \alpha - \beta = (1, -1, -4).$$

Similarly,

$$x^{(3,2,4)} > x^{(3,2,1)}, \quad \text{because } \alpha - \beta = (0, 0, 3).$$

This is the default monomial ordering in Mathematica and many other computer algebra systems.

Example 3.1.2. Let $\alpha, \beta \in \mathbb{N}^n$. Define $x^\alpha > x^\beta$ if either

(1) $|\alpha| = \sum_{i=1}^n \alpha_i > |\beta| = \sum_{i=1}^n \beta_i$, or

(2) $|\alpha| = |\beta|$ and the rightmost nonzero entry of $\alpha - \beta$ is negative.

We call this relation the *Graded Reverse Lexicographic order*, or *Degree Reverse Lexicographic order*, and this relation is a monomial ordering (see Definition 2.2.6 of Cox et al. (1997)). For example, under the Degree Reverse Lexicographic order

$$x^{(4,7,1)} > x^{(4,2,3)}, \quad \text{because } |(4,7,1)| = 12 > 9 = |(4,2,3)|.$$

Similarly,

$$\begin{aligned} x^{(1,5,2)} > x^{(4,1,3)}, \quad &\text{because } |(1,5,2)| = |(4,1,3)|, \\ &\text{and } (1,5,2) - (4,1,3) = (-3,4,-1). \end{aligned}$$

According to Cox, Little, and O'Shea (1997: pg. 58), it has recently been shown that the Degree Reverse Lexicographic monomial ordering is the most computationally efficient monomial ordering for many computations. As a result, we have chosen to use this monomial ordering in our own implementation of the Diaconis–Sturmfels algorithm, in Mathematica 7.

The computational effectiveness of the Degree Reverse Lexicographic monomial ordering is also supported by Theorem 6.1 of Diaconis and Sturmfels (1998), which gives a degree bound on the elements of a reduced Gröbner basis. Their Remark (i) following this theorem states that this degree bound only necessarily holds with the use of the Degree Reverse Lexicographic monomial ordering.

Choosing a monomial ordering, and letting g and $f_1, \ldots, f_m$ be polynomials in $k[x_1, \ldots, x_n]$, we are able to determine which term in each of these polynomials is the *leading term*. For a polynomial $f = \sum_{i=1}^k a_i x^{\alpha_i}$, the leading term of f, $\mathrm{LT}(f)$, is the term $a_j x^{\alpha_j}$, such that $x^{\alpha_j} > x^{\alpha_i}$, for all $1 \leq i \leq k$, with respect to our chosen monomial ordering.

In Section 2.3 of their paper, Cox, Little, and O'Shea (1997) show that the choice of a monomial ordering, together with its associated leading terms, enables us to extend the division algorithm for polynomials in one variable to polynomials in $k[x_1, \ldots, x_n]$. However, unlike in the division algorithm for polynomials in one variable, in this extension of the division algorithm to n variables the resulting remainder r need not be unique.

Example 3.1.3. Consider the polynomial $g = x^2y + xy^2 + y^2$, and let us divide it by the polynomials $f_1 = xy - 1$ and $f_2 = y^2 - 1$. If in our division algorithm we first use f_1, then we obtain

$$x^2y + xy^2 + y^2 = (x+y)(xy-1) + (1)(y^2-1) + (x+y+1). \tag{3.2}$$

Whereas, if we first use f_2 in our division algorithm, then we obtain

$$x^2y + xy^2 + y^2 = (x+1)(y^2-1) + (x)(xy-1) + (2x+1). \tag{3.3}$$

Notice that in Equation 3.2, the remainder is $r = x + y + 1$, and in Equation 3.3, the remainder is $r = 2x + 1$. Thus, in both cases f_1 and f_2 do not divide r, and so the division algorithm has been completed. However, in these two cases, the value of r differs, and so we see that this remainder is not unique.

This nonuniqueness of the remainder r greatly reduces the strength of this division algorithm. Given an ideal $I = \langle f_1, \ldots, f_m \rangle$ and a polynomial g, it is no longer necessarily the case that the remainder, upon dividing g by $f_1, \ldots, f_m$, will be zero if and only if $g \in I$. Thus, it is no longer possible to determine the membership of g in I simply by looking at its remainder upon division by the generators of I.

However, if we choose the right collection of polynomials f_i to generate I, then the uniqueness of the remainder r, upon the division of a polynomial g by the polynomials f_i, is preserved. We call this "right collection of generating polynomials" a Gröbner basis for I.

This idea is formalized in the following definition (Definition 2.5.5 in Cox et al. (1997)).

Definition 3.1.2. Fix a monomial ordering, and let $I \subset k[x_1, \ldots, x_n]$ be an ideal. Additionally, $\mathrm{LT}(I)$ denote the set of leading terms of elements of I. A finite subset $G = \{f_1, \ldots, f_m\}$ of I is said to be a *Gröbner basis* for I if

$$\langle \mathrm{LT}(f_1), \ldots, \mathrm{LT}(f_m) \rangle = \langle \mathrm{LT}(I) \rangle.$$

Note that in Proposition 3.1.1 we will be able to make our definition of a Gröbner basis less abstract by framing it in terms of the divisibility of the leading terms of I by the leading terms of the elements in the Gröbner basis for I. The following definition (Definition 2.4.1 of Cox et al. (1997)) will enable us to conclude the equivalence of Proposition 3.1.1 to Definition 3.1.2.

Definition 3.1.3. An ideal $I \subseteq k[x_1, \ldots, x_n]$ is a *monomial ideal* if there is a (possibly infinite) subset $A \subseteq \mathbb{N}^n$ such that I consists of all polynomials which are finite sums of the form $\sum_{\alpha \in A} h_\alpha x^\alpha$, where $h_\alpha \in k[x_1, \ldots, x_n]$. In this case, we write $I = \langle x^\alpha : \alpha \in A \rangle$.

By Lemma 2.4.2 of Cox et al. (1997), a monomial ideal $I = \langle x^{\alpha} : \alpha \in A \rangle$ has the property that $x^{\beta} \in I$ if and only if x^{β} is divisible by x^{α} for some $\alpha \in A$. By Proposition 2.5.3 of Cox et al. (1997), if $I \subseteq k[x_1, \ldots, x_n]$ is an ideal, then $\langle \mathrm{LT}(I) \rangle$ is a monomial ideal. This gives us the following equivalence, which is presented on page 77 of Cox et al. (1997).

Proposition 3.1.1. *A set $G = \{f_1, \ldots, f_m\}$ is a Gröbner basis for I if and only if the leading term of any element of I is divisible by one of the $LT(f_i)$, where $f_i \in G$.*

Note that the existence of a Gröbner basis for any ideal $I \subseteq k[x_1, \ldots, x_n]$ is guaranteed by the Hilbert Basis theorem. Further, the proof of the Hilbert Basis theorem given in Cox et al. (1997) shows that a Gröbner basis for I must also be a generating set for I.

With the more concrete definition of a Gröbner basis given in Proposition 3.1.1, we obtain the following proposition (Proposition 2.6.1 of Cox et al. (1997)) that strengthens the division algorithm.

Proposition 3.1.2. *Let $G = \{f_1, \ldots, f_m\}$ be a Gröbner basis for an ideal $I \subseteq k[x_1, \ldots, x_n]$, and let $g \in k[x_1, \ldots, x_n]$. Then there is a unique $r \in k[x_1, \ldots, x_n]$ with the following two properties:*

(i) No term of r is divisible by any of $LT(f_1), LT(f_2), \ldots, LT(f_m)$.

(ii) There exists some $f \in I$ such that $g = f + r$.

In particular, r is the remainder on division of g by G no matter how the elements of G are listed when using the division algorithm.

Notice that by this proposition, if G is a Gröbner basis for a polynomial ideal I, then g is generated by G if and only if the unique remainder upon dividing g by G, is $r = 0$. In this case we have $g \in I$. Thus, with a Gröbner basis for I and the division algorithm, we have a test for membership in I.

It is important to realize that a Gröbner basis need not be unique. This is seen in the following example, which we take from Section 2.7 of Cox et al. (1997).

Example 3.1.4. Consider the ring $k[x, y]$ with the Degree Reverse Lexicographic order, and let

$$I = \langle x^3 - 2xy, x^2y - 2y^2 + x \rangle.$$

Then for any constant $a \in k$, the polynomials

$$f_1 = x^2 + axy, \quad f_2 = xy, \quad f_3 = y^2 - \frac{1}{2}x \tag{3.4}$$

form a Gröbner basis for I.

The following definition (Definition 2.7.5 in Cox et al. (1997)) enables us to associate with each polynomial ideal a unique Gröbner basis.

Definition 3.1.4. A *reduced Gröbner basis* for a polynomial ideal I is a Gröbner basis G for I such that:

(i) $\mathrm{LT}(p)$ has coefficient 1 for all $p \in G$.

(ii) For all $p \in G$, no monomial of p lies in $\langle \mathrm{LT}(G - \{p\}) \rangle$.

By Proposition 2.7.6 of Cox et al. (1997), given a nonzero polynomial ideal I, together with a monomial ordering, I has a unique reduced Gröbner basis.

Example 3.1.5. In Example 3.1.4, if $a = 0$, then f_1, f_2, f_3 of Equation 3.4 form the unique reduced Gröbner basis for I.

This leaves the question of how to compute the reduced Gröbner basis for a given polynomial ideal. If we have an ideal I with generating set $g_1, \ldots, g_k$, there are several algorithms for turning this generating set for I into the reduced Gröbner basis for I. One such algorithm is Buchberger's algorithm, which is presented in Section 2.7 of Cox et al. (1997). Algorithms for turning a generating set for an ideal into a reduced Gröbner basis for that ideal are implemented in most computer algebra systems. In Mathematica 7, one such algorithm is implemented under the command `GroebnerBasis`. It is this command that we have used in our own implementation of the Diaconis–Sturmfels algorithm.

3.2 Elimination Theory and Implicitization Theory

In this section, we introduce the ideas from elimination theory and implicitization theory sufficient for an understanding of the Diaconis–Sturmfels algorithm, which computes a Markov basis for $\mathcal{F}_t$. We will then present this algorithm in Section 3.3. Most of the material in this section comes directly from Chapter 3 of Cox et al. (1997).

First, we give the definition of an affine variety (Definition 1.2.1 of Cox et al. (1997)).

Definition 3.2.1. Let k be a field and let $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$. The *affine variety defined by* $f_1, f_2, \ldots, f_s$ is the set

$$\mathbf{V}(f_1, \ldots, f_s) = \{(a_1, \ldots, a_n) \in k^n : f_i(a_1, \ldots, a_n) = 0 \text{ for all } 1 \leq i \leq s\}.$$

The *affine variety of an ideal* $I \subseteq k[x_1, \ldots, x_n]$, denoted $\mathbf{V}(I)$, is the affine variety defined by a generating set for I.

We have seen in Section 3.1 that a Gröbner basis is a valuable tool for determining ideal membership in a polynomial ideal. Elimination theory and implicitization theory also depend heavily on the theory of Gröbner bases. These two theories arise in the solutions to the following two problems, which we take almost verbatim from the beginning of Section 2.1 in Cox et al. (1997):

(a) Let k be a field. Find all common solutions in k^n of the system of polynomial equations

$$f_1(x_1, \ldots, x_n) = \cdots = f_s(x_1, \ldots, x_n) = 0.$$

This is equivalent to asking for the points in $\mathbf{V}(f_1, \ldots, f_s)$.

(b) Let k be a field. Let V be a subset of k^n given parametrically as

$$\begin{aligned} x_1 &= g_1(t_1, \ldots, t_m), \\ x_2 &= g_2(t_1, \ldots, t_m), \\ &\vdots \\ x_n &= g_n(t_1, \ldots, t_m), \end{aligned}$$

where, for $1 \leq i \leq n$, we have $x_i \in k$, and where this collection of x_i together forms an n-tuple $(x_1, \ldots, x_n) \in k^n$. If the g_i are polynomials in the variables t_j, then V will be an affine variety (or part of one). Find a system of polynomial equations in the x_i that defines the variety.

Here, problems a and b are "inverse problems". By this we mean that problem a asks for the set of solutions to a given system of polynomial equations, whereas problem b provides us with a set of solutions and asks us to find a system of polynomial equations with these solutions. We refer to the problem given in b as the *implicitization problem*, because its basic idea is to convert an explicit parameterization for an affine variety V into a set of equations that implicitly define V. It is for the purpose of answering the implicitization problem that elimination theory has been developed.

Definition 3.2.2. Given an ideal $I = \langle f_1, \ldots, f_s \rangle \subseteq k[x_1, \ldots, x_n]$, and given $0 \leq l \leq n$, the *lth elimination ideal*, denoted I_l is the ideal in the polynomial ring $k[x_{l+1}, x_{l+2}, \ldots, x_n]$ defined by

$$I_l = I \cap k[x_{l+1}, x_{l+2}, \ldots, x_n].$$

Our Definition 3.2.2 is Definition 3.1.1 from Cox et al. (1997). Notice that, under this definition, when we take $l = 0$, the zeroth elimination ideal is $I_0 = I$. Also notice that different orderings of the indeterminates $x_1, \ldots, x_n$ in the polynomial ring $k[x_1, \ldots, x_n]$, result in different elimination ideals.

The following definition is presented in Exercise 5 in Section 3.1 of Cox et al. (1997). Its usefulness is made apparent by the role that it plays in Theorem 3.2.1.

Definition 3.2.3. Consider the polynomial ring $k[x_1, \ldots, x_n]$, and fix an integer $1 \leq l \leq n$. We say that a monomial ordering on $k[x_1, \ldots, x_n]$ is of l-*elimination type* if any monomial involving at least one of $x_1, \ldots, x_l$ is greater than all monomials in $k[x_{l+1}, x_{l+2}, \ldots, x_n]$.

The following result is referred to as the *Elimination Theorem*, and can be found in Exercise 5 in Section 3.1 of Cox et al. (1997).

Theorem 3.2.1. *If $I \subseteq k[x_1, \ldots, x_n]$ is an ideal and G is a Gröbner basis for I with respect to a monomial ordering of l-elimination type, then*

$$G_l = G \cap k[x_{l+1}, x_{l+2}, \ldots, x_n]$$

is a Gröbner basis of the lth elimination ideal $I \cap k[x_{l+1}, x_{l+2}, \ldots, x_n]$.

Restated, this theorem tells us that given G, a Gröbner basis for I, we can find the Gröbner basis for I_l by taking exactly those elements of G that do not involve any of the first l indeterminates of $k[x_1, \ldots, x_n]$. This is a very powerful and useful statement.

This brings us to a result that is referred to as the *polynomial implicitization theorem*, which is Theorem 3.3.1 of Cox et al. (1997).

Theorem 3.2.2. *Let k be an infinite field, and let $F : k^m \to k^n$ be the function determined by the polynomial parameterization*

$$F(t_1, \ldots, t_m) = (f_1(t_1, \ldots, t_m), f_2(t_1, \ldots, t_m), \ldots, f_n(t_1 \ldots, t_m)).$$

Let I be the ideal

$$I = \langle x_1 - f_1, x_2 - f_2, \ldots, x_n - f_n \rangle \subseteq k[t_1, t_2, \ldots, t_m, x_1, x_2, \ldots, x_n]$$

and let $I_m = I \cap k[x_1, \ldots, x_n]$ be the mth elimination ideal. Then $\mathbf{V}(I_m)$ is the smallest variety in k^n containing $F(k^m)$.

Theorems 3.2.1 and 3.2.2 together give us the following algorithm (see Section 3.3 of Cox et al. (1997)):

> **Implicitization algorithm:** Let k be an infinite field, and let $f_1, \ldots, f_n$ be polynomials in $k[t_1, \ldots, t_m]$. This allows us to view each f_i as a coordinate function in the polynomial parameterization $F : k^m \to k^n$, as described in Theorem 3.2.2. Construct the ideal
>
> $$I = \langle x_1 - f_1, \ldots, x_n - f_n \rangle,$$
>
> and let $>$ be a monomial ordering of m-elimination type (such as the Degree Reverse Lexicographic order, with $t_1 > t_2 > \cdots > t_m > x_1 > x_2 > \cdots > x_n$). Then, we can compute a Gröbner basis, G, for I with respect to $>$, and from the elimination theorem we have as a Gröbner basis for the mth elimination ideal, with respect to $>$, the collection of polynomials
>
> $$G_m = G \cap k[x_1, \ldots, x_n].$$
>
> Further, $\mathbf{V}(G_m) \subseteq k^n$ is the smallest variety containing $F(k^m)$.

We will see in Section 3.3, that it is this implicitization algorithm that is the principle actor in generating Markov bases for conditioning on a summary statistic using the Diaconis–Sturmfels algorithm.

3.3 Generating a Markov Basis for $\mathcal{F}_t$

We now have sufficient background to address the problem of computing a Markov basis for $\mathcal{F}_t$ given a survey result v_0 with summary statistic t. The theory presented in this section comes primarily from Section 3 of Diaconis and Sturmfels (1998). Our discussion of toric ideals comes from Chapter 4 of Sturmfels (1996).

Let V be the underlying vector space associated with the survey that produced v_0, and let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the set of all possible participant responses to this survey. We assume that $\mathcal{X}$ is finite.

Recall that the standard basis for V consists of the indicator functions on $\mathcal{X}$. For each $x \in \mathcal{X}$, we denote its indicator function by $\mathbf{x}$, so that we have $\mathbf{x} \in L_{\mathbb{N}}(\mathcal{X})$ defined, for all $x_i \in \mathcal{X}$, by

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

Let $\mathcal{F}$ be the set of all possible results of the same survey that produced v_0. This gives us $\mathcal{F} \subseteq L_{\mathbb{N}}(\mathcal{X}) \subseteq L_{\mathbb{Z}}(\mathcal{X}) \subseteq V$, where we take $L_{\mathbb{N}}(\mathcal{X})$ and

$L_{\mathbb{Z}}(\mathcal{X})$ to be the sets of all $\mathbb{N}$-valued and $\mathbb{Z}$-valued functions defined on $\mathcal{X}$, respectively. In this way, we can represent any result $f \in \mathcal{F}$ as an element of $L_{\mathbb{N}}(\mathcal{X})$.

Let k be an infinite field. We can represent any $f \in L_{\mathbb{N}}(\mathcal{X})$ as a monomial $\mathcal{X}^f \in k[x_1, \ldots, x_n]$ by writing

$$\mathcal{X}^f := \prod_{i=1}^{n} x_i^{f(x_i)}.$$

Here, we are letting the indeterminates of the polynomial ring $k[x_1, \ldots, x_n]$ correspond to the possible participant responses to the survey. That is, these indeterminates correspond to the elements of $\mathcal{X}$. By the notation $f(x_i)$, we mean the entry of f that is associated with x_i, when we are encoding $f \in L_{\mathbb{N}}(x_1, \ldots, x_n)$ with respect to the standard basis. Thus, given a survey result f, we see that $f(x_i)$ counts the number of survey participants who chose x_i.

We can extend this association between survey results and elements in $k[x_1, \ldots, x_n]$, to all $f \in L_{\mathbb{Z}}(\mathcal{X})$, by writing $f = f^+ - f^-$, where $f^+, f^- \in L_{\mathbb{N}}(\mathcal{X})$ are defined by

$$f^+(x_i) = \begin{cases} f(x_i) & \text{if } f(x_i) \geq 0, \\ 0 & \text{if } f(x_i) < 0, \end{cases}$$

$$f^-(x_i) = \begin{cases} -f(x_i) & \text{if } f(x_i) \leq 0, \\ 0 & \text{if } f(x_i) > 0. \end{cases}$$

For $f \in L_{\mathbb{Z}}(\mathcal{X})$ we write

$$\mathcal{X}^f := X^{f^+} - X^{f^-}.$$

Remark 3.3.1. Notice that this mapping from $L_{\mathbb{Z}}(\mathcal{X})$ to $k[x_1, \ldots, x_n]$ is not structure preserving in the sense that we might expect.

Consider two functions $f, g \in L_{\mathbb{N}}(\mathcal{X})$. Then, $(f+g)(x) = f(x) + g(x)$, and so we have

$$\mathcal{X}^{f+g} = \prod_{i=1}^{n} x_i^{f(x_i)+g(x_i)}.$$

This is a monomial, whereas $\mathcal{X}^f + \mathcal{X}^g$ is a binomial, unless g and f are scalar multiples of each other. Thus, we see that in general $\mathcal{X}^{f+g} \neq \mathcal{X}^f + \mathcal{X}^g$.

Similarly, with multiplication we have $(fg)(x) = f(x)g(x)$ for $f, g \in L_{\mathbb{N}}(\mathcal{X})$, which gives us $\mathcal{X}^f \mathcal{X}^g \neq \mathcal{X}^{fg}$. In fact, instead we have

$$\mathcal{X}^f \mathcal{X}^g = \prod_{i=1}^{n} x_i^{f(x_i)} x_i^{g(x_i)} = \prod_{i=1}^{n} x_i^{f(x_i)+g(x_i)} = \mathcal{X}^{f+g}.$$

This makes it clear that the collection of monomials associated with elements of $L_{\mathbb{Z}}(\mathcal{X})$ under the map that we have defined does not inherit the ring structure of $L_{\mathbb{Z}}(\mathcal{X})$.

Let T be the linear transformation on the underlying vector space V that computes t from v_0. Letting d be the number of entries in the statistic t, or equivalently the number of rows in T, we have $T : L_{\mathbb{Z}}(\mathcal{X}) \to \mathbb{N}^d$. For $f \in L_{\mathbb{Z}}(\mathcal{X})$, let the notation $T(f)_i$ denote the ith coordinate of $T(f) \in \mathbb{N}^d$.

For each $x \in \mathcal{X}$, viewing x as a monomial in $k[\mathcal{X}]$, define the map

$$\varphi_T : k[\mathcal{X}] \to k[t_1, \dots, t_d],$$

such that

$$\varphi_T(x) = t_1^{T(\mathbf{x})_1} t_2^{T(\mathbf{x})_2} \cdots t_d^{T(\mathbf{x})_d},$$

and extend this map linearly and multiplicatively from the elements of $\mathcal{X}$ to the rest of $k[\mathcal{X}]$. It is clear from its linear and multiplicative definition that this map φ_T is a ring homomorphism.

Moreover, letting each of the columns of T be denoted $\mathbf{a_i}$, for $1 \le i \le n$, we can view $\mathbf{a_i}$ as an element in $L_{\mathbb{N}}(\mathbb{N}^d)$, which can then be viewed as a monomial in $k[t_1, \dots, t_d]$. For each column of T, we denote the monomial associated to $\mathbf{a_i}$ as $\mathbf{t}^{\mathbf{a_i}}$. Notice that with this notation, for $x_i \in \mathcal{X}$, we have

$$\varphi_T(x_i) = \mathbf{t}^{\mathbf{a_i}},$$

because

$$T(\mathbf{x}_i) = (\mathbf{x}_i)_1 \mathbf{a_1} + \cdots + (\mathbf{x}_i)_n \mathbf{a_n} = (\mathbf{x}_i)_i \mathbf{a_i} = \mathbf{a_i}.$$

In this way, the ring homomorphism φ_T is a map of the form $\hat{\pi}$ discussed at the beginning of Chapter 4 in Sturmfels (1996), and so the kernel of φ_T is the *toric ideal* of $\{\mathbf{a_1}, \dots, \mathbf{a_n}\}$, which we call I_T.

Because I_T is the toric ideal of $\{\mathbf{a_1}, \dots, \mathbf{a_n}\}$, and because

$$T : (u_1, \dots, u_n) \in \mathbb{N}^n \mapsto u_1 \mathbf{a_1} + \cdots + u_n \mathbf{a_n} \in \mathbb{N}^d,$$

by Corollary 4.3 of Sturmfels (1996), we see that

$$I_T = \langle \mathcal{X}^{f^+} - \mathcal{X}^{f^-} : f \in \ker(T) \rangle.$$

Further, by Corollary 4.4 of Sturmfels (1996), given a monomial ordering on $k[\mathcal{X}]$, there is a finite set $G_d \subset \ker(T)$, such that the collection

$$\{\mathcal{X}^{f^+} - \mathcal{X}^{f^-} : f \in G_d\}$$

forms a reduced Gröbner basis for I_T.

We can find this set G_d, given the summary statistic matrix T, by using a special case of the implicitization algorithm that was discussed in Section 3.2. Sturmfels presents this special case of the implicitization algorithm in Algorithm 4.5 of Sturmfels (1996). We reproduce it here, using the simplification that Sturmfels discusses on page 32 of Sturmfels (1996), which arises from the fact that the entries of the summary statistic matrix T are each nonnegative.

Computing a first Gröbner basis of a toric ideal:

1. Introduce $n+d+1$ indeterminates $t_1, t_1, \ldots, t_d, x_1, x_2, \ldots, x_n$, and fix an elimination order with $\{t_i\} > \{x_j\}$.
2. Compute the reduced Gröbner basis G for the ideal
$$\langle x_i - \mathbf{t}^{\mathbf{a_i}} : i = 1, \ldots, n \rangle.$$
3. Output: The set $G_d = G \cap k[\mathcal{X}]$ is the reduced Gröbner basis for I_T with respect to the chosen elimination ordering.

It is this algorithm that we call the *Diaconis–Sturmfels algorithm*, and that we use to generate a Markov basis for $\mathcal{F}_t$. The connection between I_T and a Markov basis for $\mathcal{F}_T$ is established in Theorem 3.1 of Diaconis and Sturmfels (1998). Without this theorem, the above algorithm would be of no use in computing Markov bases.

Theorem 3.3.1. *(Theorem 3.1 of Diaconis and Sturmfels (1998)) A collection of functions $f_1, \ldots, f_L \in L_{\mathbb{Z}}(\mathcal{X})$ is a Markov basis for $\mathcal{F}_t$ if and only if the set*

$$\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-}, \quad \text{for } 1 \leq i \leq L$$

generates the ideal I_T.

For the proof of this theorem, see Diaconis and Sturmfels (1998). Both directions involve induction. The proof also relies upon the following proposition, which Diaconis and Sturmfels (1998) assumes without proof.

Proposition 3.3.1. *Let $f \in L_{\mathbb{Z}}(\mathcal{X})$. Then, $T(f) = 0$ if and only if $\mathcal{X}^f \in$* $\ker(\varphi_T)$.

Proof. Let $f \in L_{\mathbb{Z}}(\mathcal{X})$ with $T(f) = 0$. Then $f = f^+ - f^-$, and T is linear, so

$$0 = T(f) = T(f^+ - f^-) = T(f^+) - T(f^-).$$

Thus, we have $T(f^+) = T(f^-)$. Then, by definition this gives us

$$\varphi_T(\mathcal{X}^{f^+}) = \varphi_T(\mathcal{X}^{f^-}),$$

and because φ_T is a ring homomorphism, this gives us that

$$\varphi_T(\mathcal{X}^{f^+} - \mathcal{X}^{f^-}) = 0.$$

That is, this gives us that $\mathcal{X}^f = \mathcal{X}^{f^+} - \mathcal{X}^{f^-} \in \ker(\varphi_T)$. Because all of these implications are bidirectional, the reverse direction also holds, concluding the proof.

□

Theorem 3.3.1 tells us that, in order to find a Markov basis for $\mathcal{F}_t$, we need only find a collection of monomial differences that generate I_T. We can then use Algorithm 4.5 of Sturmfels (1996), which we have presented above, to find such a generating set for I_T.

We have attempted to make as clear as possible the intuition behind this algorithm and the notation used in defining it. We now present the Diaconis–Sturmfels algorithm in the following theorem, as it appears in Diaconis and Sturmfels (1998), with only some slight notational modifications.

Theorem 3.3.2. *(Theorem 3.2 of Diaconis and Sturmfels (1998)) Let $\mathcal{X}$ be a finite set. Let $T : \mathcal{X} \to \mathbb{N}^d$ be given. Let $\mathcal{T} = \{t_1, \ldots, t_d\}$. Given an ordering for $\mathcal{X}$, extend it to an elimination ordering for $\mathcal{X} \cup \mathcal{T}$ with $t > x$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$ in the polynomial ring $k[\mathcal{X}, \mathcal{T}]$. Define $\mathcal{I}_T = \langle x - \mathcal{T}^{T(\mathbf{x})} : x \in \mathcal{X} \rangle$. Then $\mathcal{I}_{T_d} = \mathcal{I}_T \cap k[\mathcal{X}]$ is the toric ideal I_T, and the reduced Gröbner basis for I_T can be found by computing a reduced Gröbner basis for $\mathcal{I}_T$ and taking those output polynomials which only involve $\mathcal{X}$.*

With this algorithm, given an election result v_0 and a summary statistic matrix T with $T(v_0) = t$, we can compute a Markov basis for $\mathcal{F}_t$.

Chapter 4

Putting Theory into Practice

With access to the theory of Markov processes from Chapter 2, and to implicitization and elimination theory from Chapter 3, we have been able to develop a clear presentation of the Diaconis–Sturmfels algorithm, and to explain its usefulness in determining the conditional significance of summary statistics that describe the result of a survey. In this chapter, we concretely present the approach to data analysis that arises from the theory presented so far in this paper. We also discuss some of the necessary considerations in implementing this approach.

4.1 The Choice of a Collection of Nested Statistics

In our example in Section 1.1, we considered a nested collection of summary statistics for approval voting data. Recall that for fixed k and $0 \leq i \leq k$, the *ith order effects statistic* counts the number of times that each different i-subset of the n candidates was chosen by voters who selected k candidates. Notice that we can use these counts to determine the number of times that each different $(i-1)$-subset of the n candidates was chosen by voters who selected k candidates. In this way, given the ith order effects statistic for an election result, we can determine the $(i-1)$st order effects statistic for that result. It is for this reason that we call these summary statistics *nested*.

In general, given a collection of summary statistics $t_0, \ldots, t_k$ that are computed, respectively, by the linear transformations $T_0, \ldots, T_k$, we say that this collection of summary statistics is *nested* if, for $0 \leq i \leq k$, the rowspace of T_i contains the rowspaces of $T_0, \ldots, T_{i-1}$.

The choice of a nested collection of summary statistics to describe the result of a survey enables us to find a single summary statistic that best describes the result. Given the ith order effects statistic for approval data, the first through $(i-1)$st order effects statistics for that data can immediately be determined. In this way, the smallest order effects statistic such that all of the larger order effects statistics are residual, contains the essential information of the data. Because the kth order effects statistic trivially satisfies the property that all larger order effects statistics are residual, the existence of a smallest order effects statistic satisfying this property is guaranteed.

This nice property of nested collections of summary statistics, that there exists a single summary statistic that best captures the information in a survey result, makes the use of such collections of summary statistics desirable. Thus, given the result of a survey on which we would like to perform an analysis, the first step is the choice of a nested collection of summary statistics that can be used to describe the data.

For typical choices of a collection of nested summary statistics in the analysis of the result of a survey, the zeroth summary statistic simply counts the number of survey responses, and the largest summary statistic acts as the identity on the result.

The Isotypic and Inversion Decompositions for Fully Ranked Data

The choice of a collection of nested summary statistics that are informative for a given survey is generally not unique. For example, in Section 2.6 of Marden (1995), Marden presents two different decompositions of the underlying vector space associated with fully ranked data, namely the spectral decomposition and the inversion decomposition. The spectral decomposition arises as a result of representation theory that deals with the action of the symmetric group on the associated underlying vector space, while the inversion decomposition arises from the use of summary statistics that count the number of times that particular orderings of the items in a survey show up in the survey responses. Both of these decompositions give rise to distinct, nested collections of summary statistics for fully ranked data.

In Section 6 of Diaconis and Sturmfels (1998), Diaconis and Sturmfels use the nested summary statistics associated with the spectral decomposition of the underlying vector space in their analysis of fully ranked data, examining an election with 4 candidates. This collection of summary statistics associated with the spectral decomposition is also used in Diaconis and Eriksson (2006) by Diaconis and Eriksson, in their exploration of the analysis of fully ranked data using the Diaconis–Sturmfels algorithm.

It does not appear that the use of the Diaconis–Sturmfels algorithm has been explored in the analysis of fully ranked election data using the nested summary statistics associated with the inversion decomposition. It is straightforward to generate the summary statistic matrices associated with the inversion decomposition, and we have preliminarily experimented with the computation of the Markov bases for conditioning on the associated summary statistics, without confronting any unexpected hurdles. It would be interesting to pursue this work further, and to see how the analysis of fully ranked data is changed under this different decomposition.

4.2 Nested Summary Statistics for Data Analysis

Consider a survey with underlying vector space V, and a result v_0 on which we would like to perform an analysis. Let $t_0, t_1, \ldots, t_k$ be a collection of nested summary statistics for v_0, that are computed, respectively, via the linear transformations $T_0, T_1, \ldots, T_k$. We take T_0 to be the single row matrix whose entries are all 1s, with respect the standard basis for V, so that T_0 simply counts the number of survey responses in v_0. We take T_k to act as the identity on V.

With this choice of nested summary statistics, linear algebra gives us a natural decomposition of V into orthogonal subspaces,

$$V = V_0 \oplus V_1 \oplus \cdots \oplus V_k.$$

Here, for each $0 \leq i \leq k$, $V_0 \oplus \cdots \oplus V_i$ forms the rowspace of T_i, and the kernel of T_i is $V_{i+1} \oplus \cdots \oplus V_k$. Recalling our motivating example from Section 1.1, of the analysis of approval voting data, notice that for each $0 \leq i \leq k$, the subspace V_i corresponds to the "pure ith order effects space".

It is the magnitudes of the projections of v_0 into these V_i that we use to determine which summary statistics in $t_0, t_1, \ldots, t_k$ best capture the information in v_0. For each $0 \leq i \leq k-1$, we would like to compute the projections of the elements of $\mathcal{F}_{t_i}$ into $V_{i+1}, \ldots, V_k$. Recall here, that $\mathcal{F}_{t_i}$ is the set of all possible survey results that share the statistic t_i with v_0. We compare the magnitudes of the projections of the elements of $\mathcal{F}_{t_i}$ with the magnitudes of the projections of v_0. If we find that none of the projections of v_0 into $V_{i+1}, \ldots, V_k$ is larger than expected, given the projections of the elements in $\mathcal{F}_{t_i}$ into these same subspaces, then we say that the statistic t_i captures the essential information of v_0, and that the summary statistics $t_{i+1}, \ldots, t_k$ are artifacts of t_i. That is, for each $j > i$, the summary statistic t_j is not significant, after conditioning on t_i. When such a t_i exists, we say that it is this statistic that best describes the result v_0.

Unfortunately, in general, the size of the set $\mathcal{F}_{t_i}$ prohibits the computation of the projections of the elements of $\mathcal{F}_{t_i}$ into $V_{i+1}, \ldots, V_k$. Thus, we instead use the Diaconis–Sturmfels algorithm to generate a Markov basis for $\mathcal{F}_{t_i}$, and then use this Markov basis in a Markov chain with underlying hypergeometric distribution to generate a representative sample of $\mathcal{F}_{t_i}$, as we have outlined in Section 2.3. We then proceed with the method described above for determining the conditional significance of higher order effects statistics, computing the projections of the elements in our representative sample of $\mathcal{F}_{t_i}$ into the subspaces $V_{i+1}, \ldots, V_k$.

Computational Approach

In this process, it is the computation of the reduced Gröbner basis for the toric ideal associated with a summary statistic T_i, involved in the Diaconis–Sturmfels algorithm, that is computationally intensive and limits the types of survey results on which we can perform such an analysis.

In Diaconis and Eriksson (2006), Diaconis and Eriksson report that they have been able to compute Markov bases associated with fully ranked data from elections involving five candidates. This computation took them 90 hours of CPU time on a 2 GHz machine, and they report that this is the current computational bound for the analysis of fully ranked data using summary statistics associated with the the spectral decomposition of the underlying vector space. In Hansen and Orrison (2008), Hansen and Orrison compute the Markov bases associated with approval voting data for elections involving up to eight candidates, but they report that this is the current computational bound for the analysis of approval data using their method.

Both Diaconis and Eriksson (2006) and Hansen and Orrison (2008) use `4ti2` in computing their Markov bases. This is a software package that has been designed to optimize the computation of a Markov basis for conditioning on a given summary statistic matrix.

Most of our data analysis has been carried out using Mathematica 7. Originally, we were using the code that I had written using Mathematica's `GroebnerBasis` command, to implement the Diaconis–Sturmfels algorithm for computing the Markov bases that we desired for our analysis. However, using this Mathematica code, we found the computation necessary for analyzing the result of a survey in which participants were asked to identify their top three choices among six items, to be infeasible. Whereas, using `4ti2` this computation took negligible time. In fact, with `4ti2`, the computations associated with surveys of this form do not begin to slow down until we look at surveys in which voters are asked to select their top three choices among eight items. Thus, I have written code in Mathematica to facilitate the transfer of matrices between Mathematica and `4ti2`, and we have used `4ti2` in computing the Markov bases for use in our analysis. The details of our analysis of partially ranked data are given in Section 5.2. In that section we provide an analysis of the result of a survey on Girl Scout cookie preferences.

Remark 4.2.1. Given our interest in nested summary statistics for survey results, and the computational intensity of computing Markov bases, it would be extremely nice if there were a method for turning the Markov basis conditioning on a summary statistic computed by a linear transformation T, with rowspace W, into a Markov basis conditioning on a summary statistic computed by a linear transformation T', whose rowspace W' contains W as a direct summand, so that $W' = W \oplus U$, for some vector subspace $U \subset V$, where V is the underlying vector space associated with the survey.

We might even expect it to be possible to find such a method for turning a Markov basis B for T into a Markov basis B' for T', because elements in B' must be in the kernel of T', and so they must also be in the kernel of T. Thus, elements in B' can be represented as integer linear combinations of the elements of B.

It is straightforward, though notationally involved, to prove that if the projections into U of the elements of B have integer entries with respect to the standard basis for V, then B can be turned into a Markov basis for T'. This new Markov basis for B' generated from B consists of the projections of the elements of B into U.

The necessity of the requirement in this result, for integer entries with respect to the standard basis for V, is clear, because each step along a Markov basis element starting at a legal survey result must itself be a legal survey result. Unfortunately, given a collection of nested summary statistics for a survey, in general, it is not the case that the projections of the elements of a Markov basis for one of the lower order summary statistics, into the rowspace of a higher order summary statistic matrix, have integer entries with respect to the standard basis for the associated underlying vector space, and so this result is not very useful in practice.

Chapter 5

Analysis of Partially Ranked Data Under the Isotypic Decomposition

Consider a survey in which participants are asked to identify their top k choices of n items. Let the underlying vector space associated with this survey be V, and let the set of all possible survey results be $\mathcal{F}$.

Further, consider a result v_0 of this survey, and a collection $T_0, \ldots, T_k$ of linear transformations on V that compute a collection of nested summary statistics $t_0, \ldots, t_k$ for v_0. Here, we take T_0 to simply count the number of survey responses, and we take T_k to be the identity on V. Our analysis of v_0, with respect to the these summary statistics, will produce a significant statistic t_S, with $0 \leq S \leq k$, that best describes v_0. We would like our analysis of v_0 to return the same significant t_S regardless of our choice of labeling for the n candidates. This acts as a restriction on our choice of $T_0, \ldots, T_k$.

Recall from Chapter 4 that, given a choice of nested summary statistics $t_0, \ldots, t_k$ for v_0, there is an associated natural decomposition of V into $k+1$ orthogonal subspaces $V_0, \ldots, V_k$, where for each $0 \leq i \leq k$, we have

$$\ker(T_i) = V_{i+1} \oplus \cdots \oplus V_k.$$

Noticing that S_n, the symmetric group on n elements, acts transitively on the k-subsets of the n items, we can be confident that if, for each $0 \leq i \leq k$, the space V_i associated with the statistic t_i is invariant under the action of S_n, then our analysis of v_0 with respect to this choice of nested summary statistics will be independent of the choice of labeling for the items in the survey.

In this way, a natural choice of $T_0, \ldots, T_k$ would be so that the resulting V_i are the isotypic subspaces of V under the action of S_n. This is the line of reasoning used by Diaconis and Eriksson (2006) and by Diaconis and Sturmfels (1998: Section 6), in their choices to use summary statistics associated with the spectral decomposition of V in their analysis of fully ranked data. In Hansen and Orrison (2008), the authors also chose to use a collection of nested summary statistics that yields the isotypic decomposition of the underlying vector space in their analysis of approval voting data.

In considering partially ranked data resulting from a survey in which participants are asked to identify their top k choices of n items, the choice of a nested collection of summary statistics for v_0, such that the resulting $V_0, \ldots, V_k$ are each isotypic subspaces of V under the action of S_n, yields a very specific collection of associated linear transformations T_i. Namely, we get the same *ith order effects statistics* that were discussed in Section 1.1, and that are used in Hansen and Orrison (2008). That is, for $0 \leq i \leq k$, T_i simply counts, for each i-subset of the n items, the number of responses in v_0 that include that i-subset in their chosen top k-items.

Under this definition, T_0 simply counts the number of responses in v_0, and T_k counts, for each k-subset, the number of responses in v_0 that include that k-subset in their chosen top k-items. In this way, T_k simply acts as the identity operator on V. Thus, according to the criteria discussed in Chapter 4, this collection of nested summary statistics is a desirable choice for use in the analysis of partially ranked data.

Notice that if we were to take $i > k$, then the resulting T_i would be the zero operator on V, because it is impossible in this survey for participants to choose an i-subset of the n items. Thus, is natural to bound i by k.

For each $0 \leq i \leq k$, we call the isotypic subspace V_i the *pure ith order effects space*, adopting this terminology from Hansen and Orrison (2008).

In this chapter, we first explore the use of this collection of summary statistics in studying partially ranked data, and then use this approach to analyze the result of a survey on Girl Scout cookie preferences.

5.1 Computational Considerations

We begin with an example.

Example 5.1.1. We use the setup from Examples 2.1.1 and 2.1.2. Consider a survey in which participants are asked to identify their preferred two choices of four items, so that $n = 4$ and $k = 2$. This gives us a total of $\binom{4}{2} = 6$ possible individual participant responses to this survey, and so the underlying vector space V associated with this survey is 6 dimensional. Let us order these individual participant responses according to the list

$$\texttt{Subsetsk2n4} = (\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}).$$

and let this ordering define the ordering of the elements in the standard basis for V. This gives us the zeroth order effects statistic matrix

$$T_0 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

the first order effects statistic matrix

$$T_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

and the second order effects statistic matrix (the identity operator)

$$T_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Given a survey result

$$v_0 = (a, b, c, d, e, f),$$

where the ℓth entry in this list indicates the number of survey responses that chose the ℓth element in `Subsetsk2n4`, we compute the summary statistics for v_0 to be

$$\begin{aligned} t_0 &= T_0 v_0 = (a + b + c + d + e + f), \\ t_1 &= T_1 v_0 = ((a + b + c), (a + d + e), (b + e + f), (c + e + f)), \\ t_2 &= T_2 v_0 = (a, b, c, d, e, f) = v_0. \end{aligned}$$

Notice that, given four items from which to choose in a survey, the choice of two most preferred items identifies an associated choice of two least preferred items. Thus, because our summary statistics are chosen so that our analysis is independent of labeling, the result of a survey in which participants are asked to identify their two least preferred choices among four items will yield the same analysis as the result of the survey in Example 5.1.1.

The same pairing exists between choices of three preferred items among four items, and the choice of a single least preferred item among four items, and the results of the analyses of surveys of these two types will similarly be the same.

This pairing generalizes to surveys in which participants are asked to identify their top k choices of n items. Given a survey with $k > n/2$, we can carry out its analysis by focusing on the "complementary" survey, in which participants are asked to identify their $k' = n - k$ least preferred items, with $k' < n/2$. Thus, in our analysis of partially ranked data associated with surveys of this form, *we need only focus our attention on cases with* $k \leq n/2$.

There is an immediate computational advantage in studying survey results of this form, compared to the study of fully ranked data. This is a direct result of the comparative sizes of the underlying vector spaces associated to these two types of surveys. In a survey in which participants are asked to rank their preference of n items, there are $n!$ possible responses, and so the associated underlying vector space has dimension $n!$. Whereas, in the type of survey that we have described involving n items, there are only $\binom{n}{k}$ possible responses, and so the underlying vector space has much smaller dimension.

The analysis of survey results of this form is even less computationally intensive than the analysis of approval data discussed in Hansen and Orrison (2008), because we know that each survey participant selects exactly k items of which they approve.

Isotypic Subspaces

Because the collection of nested summary statistics that we have chosen for use in our analysis is also associated with the analysis of approval data, we are able to take advantage of certain known efficient computational methods associated with approval data.

Given a survey involving n candidates, Maslen, Orrison, and Rockmore (2004: Section 5) present a distance transitive graph, with vertex set consisting of k-subsets of the n items, and with an edge between two vertices if and only if the associated k-subsets differ by exactly one item. By Lemma 5.1 of Maslen et al. (2004), the *eigenspaces* of the adjacency matrix A for this graph are precisely the the *isotypic subspaces* of the underlying vector space associated with this survey, under the action of S_n. Further, if we denote these

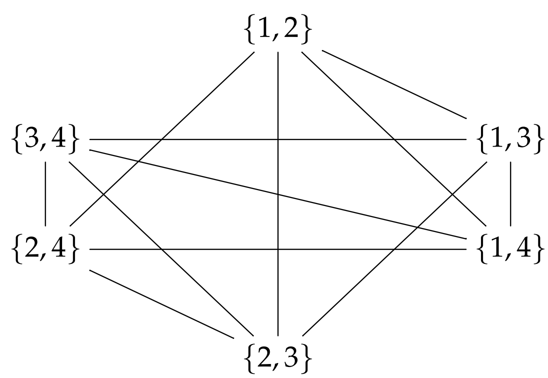

Figure 5.1 The distance transitive graph for $n = 4$, $k = 2$.

isotypic subspaces $V_0, \ldots, V_k$, where V_i is the pure ith order effects space, for $0 \le i \le k$, then the eigenspace of A with largest eigenvalue corresponds to V_0, and the eigenspace of A with smallest eigenvalue corresponds to V_k. More generally, the eigenspace of A with the $(i-1)$st-largest eigenvalue corresponds with the isotypic subspace V_i.

We clarify this approach to the computation of isotypic subspaces of V with the following example.

Example 5.1.2. Let $n = 4$ and $k = 2$. The vertex set of the associated transitive graph consists of the elements of `Subsetsk2n4` from Example 5.1.1. This graph is depicted in Figure 5.1. Ordering the vertices according to `Subsetsk2n4`, this graph has adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

We find the eigenvectors of A to be

$$\begin{aligned}
v_0 &= (1,1,1,1,1,1),\\
v_1 &= (1,0,-1,-1,0,1),\\
v_2 &= (0,1,-1,-1,1,0),\\
v_3 &= (-1,0,0,0,0,1),\\
v_4 &= (0,-1,0,0,1,0),\\
v_5 &= (0,0,-1,1,0,0),
\end{aligned}$$

with associated eigenvalues list

$$4,-2,-2,0,0,0.$$

In this way, we see that, in decomposing the underlying vector space V associated with a survey in which participants are asked to identify their top two choices of four items, the pure zeroth order effects space V_0 is spanned by the eigenvector v_0, with largest eigenvalue 4. The pure first order effects space V_1 is spanned by the eigenvectors with second largest eigenvalue 0, namely v_3, v_4, v_5. The pure second order effects space V_2 is spanned by the remaining eigenvectors v_1, v_2, with smallest eigenvalue -2.

Using the summary statistics T_i given in Example 5.1.1, notice that $T_0v_0 = 6$, while

$$T_0v_3 = T_0v_4 = T_0v_5 = T_0v_1 = T_0v_2 = 0,$$

and that $T_1v_0 = (3,3,3,3)$, $T_1v_3 = (-1,-1,1,1)$, $T_1v_4 = (-1,1,-1,1)$, $T_1v_5 = (-1,1,1,-1)$, while

$$T_1v_1 = T_1v_2 = (0,0,0,0),$$

and that T_2 acts as the identity, and so the images of $v_0, v_1, v_2, v_3, v_4, v_5$ under T_2 are each nonzero. In this way, we see that

$$\begin{aligned}
V_1 \oplus V_2 &= \operatorname{span}\{v_1,v_2,v_3,v_4,v_5\} = \ker(T_0),\\
V_2 &= \operatorname{span}\{v_1,v_2\} = \ker(T_1),
\end{aligned}$$

while $\ker(T_2) = \{0\}$. This is exactly what we would expect for a collection of nested summary statistic matrices T_0, T_1, T_2 associated with the decomposition of V into orthogonal subspaces V_0, V_1, V_2.

Computation for Analyzing Partially Ranked Data

We have written code that, when given n, k, and i, computes the ith order effects statistic matrix associated with a survey in which participants are asked to identify their k top choices of n items. We have also written code that, when given n and k, generates the adjacency matrix of the distance transitive graph described above, whose eigenspaces are the isotypic subspaces of the underlying vector space associated with a survey of this form. We use both of these methods heavily in our analysis of Girl Scout cookie preferences, which we discuss in Section 5.2. This code has been written in Mathematica 7, and is available online.

Once we have the eigenvectors that span each of the isotypic subspaces of V, it is straightforward to compute an orthonormal basis for each of these V_i, for $0 \leq i \leq k$. The use of an orthonormal basis to describe these V_i facilitates the computation of projections into these subspaces.

In Maslen, Orrison, and Rockmore (2004), the authors also present an efficient method for computing projections into the isotypic subspaces associated with this type of survey data. Their algorithm involves the Lanczos iteration, a modified version of the Gram–Schmidt orthogonalization process, discussed in Section 4.2 of the paper. Because the computation of Markov bases limits our analysis to small values of n and k, we have decided to carry out our analysis without implementing this more efficient algorithm for computing isotypic projections. This choice has, so far, not been a source of difficulty in our analysis.

Using `4ti2`, we have computed Markov bases for conditioning on the zeroth through kth order effects statistics for partially ranked data resulting from the described type of survey, for $n = 1, \ldots, 8$. These computations run very quickly up through $n = 7$. Then, for $n = 8$ and $k = 3$, the computation time dramatically increases, taking on the order of two days. The resulting Markov basis for conditioning on the second order effects statistic has $139,405$ elements.

The computation of the Markov basis for conditioning on the second order effects statistic when $n = 8$, $k = 4$ and $n = 9$, $k = 3$ seems to be at least similarly time consuming, but we have not yet gotten their computational results. This dramatic lengthening of computation time with respect to n, seems to be slightly delayed for conditioning on the third order effects statistics, with these computations running quickly for both $n = 8, k = 4$ and $n = 9, k = 3$. For a table describing the results of these computations, see Appendix A.

Given the rapid increase, with respect to n, in the amount of time that it takes to carry out these computations, it would be desirable to have a combinatorial description of the Markov basis elements that condition on the ith order effects statistic, for fixed i and arbitrary n and k.

5.2 Girl Scout Cookie Preferences

In this section, we consider the result of a survey in which participants were asked to identify their preferred three types of Girl Scout cookie. Because there are six different types of cookie, the data resulting from this survey is of the form we have been considering, with $n = 6$, $k = 3$.

The responses of participants were recorded according to the following labeling of the six types of Girl Scout cookie:

1=Do-Si-Dos,	2=Lemon Chalet Cremes,
3=Samoas,	4=Tagalongs,
5=Thin Mints,	6=Trefoils.

Lexicographically ordering the three-subsets of the six types of cookie, we get the list `Subsetsk3n6` of possible survey responses.

$$\begin{aligned}\texttt{Subsetsk3n6} = (&\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,2,6\},\{1,3,4\},\\ &\{1,3,5\},\{1,3,6\},\{1,4,5\},\{1,4,6\},\{1,5,6\},\\ &\{2,3,4\},\{2,3,5\},\{2,3,6\},\{2,4,5\},\{2,4,6\},\\ &\{2,5,6\},\{3,4,5\},\{3,4,6\},\{3,5,6\},\{4,5,6\}).\end{aligned}$$

The indicator functions on this length 20 list serve as the standard basis elements for the underlying vector space V associated with this survey.

5.2.1 Survey Result

This survey was posted to Facebook and the email discussion list for East dorm, and over the course of two days there were 132 responses. The resulting raw data is presented in Table 5.1.

We collect this result in the vector `DataVec`, encoding the survey responses with respect to the standard basis for the underlying vector space.

$$\texttt{DataVec} = (0,1,1,0,6,12,1,7,1,2,2,15,0,3,2,8,43,6,12,10).$$

Computing the eigenspaces of the distance transitive graph discussed in Section 5.1 for $n = 6$, $k = 3$, we determine the isotypic subspaces of the underlying vector space associated with this survey, under action of S_6.

$\{1,2,3\}$	0	$\{1,2,4\}$	1	$\{1,2,5\}$	1	$\{1,2,6\}$	0
$\{1,3,4\}$	6	$\{1,3,5\}$	12	$\{1,3,6\}$	1	$\{1,4,5\}$	7
$\{1,4,6\}$	1	$\{1,5,6\}$	2	$\{2,3,4\}$	2	$\{2,3,5\}$	15
$\{2,3,6\}$	0	$\{2,4,5\}$	3	$\{2,4,6\}$	2	$\{2,5,6\}$	8
$\{3,4,5\}$	43	$\{3,4,6\}$	6	$\{3,5,6\}$	12	$\{4,5,6\}$	10

Table 5.1 Girl Scout cookies data.

As we would expect for $n = 6$, $k = 3$, we find that this adjacency matrix has four distinct eigenspaces, each of which is associated with a different isotypic subspace of V.

Denoting the pure ith order effects space, for $0 \leq i \leq k$, as `spaceik3n6`, we find these isotypic subspaces to be

$$\texttt{space0k3n6} = \mathrm{span}\{(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)\},$$

$$\begin{aligned}\texttt{space1k3n6} = \mathrm{span}\{&(-1,0,0,0,0,0,0,1,1,1,-1,-1,-1,0,0,0,0,0,0,1),\\ &(0,-1,0,0,0,1,1,0,0,1,-1,0,0,-1,-1,0,0,0,1,0),\\ &(0,0,-1,0,0,-1,0,-1,0,-1,1,0,1,0,1,0,0,1,0,0),\\ &(0,0,0,-1,0,0,-1,0,-1,-1,1,1,0,1,0,0,1,0,0,0),\\ &(0,0,0,0,-1,-1,-1,-1,-1,-1,1,1,1,1,1,1,0,0,0,0)\},\end{aligned}$$

$$\begin{aligned}\texttt{space2k3n6} = \mathrm{span}\{&(1,0,0,0,0,0,0,0,0,-1,-1,0,0,0,0,0,0,0,0,1),\\ &(0,1,0,0,0,0,0,0,0,-1,-1,0,0,0,0,0,0,0,1,0),\\ &(0,0,1,0,0,0,0,0,0,-1,-1,0,0,0,0,0,0,1,0,0),\\ &(0,0,0,1,0,0,0,0,0,-1,-1,0,0,0,0,0,1,0,0,0),\\ &(0,0,0,0,1,0,0,0,0,-1,-1,0,0,0,0,1,0,0,0,0),\\ &(0,0,0,0,0,1,0,0,0,-1,-1,0,0,0,1,0,0,0,0,0),\\ &(0,0,0,0,0,0,1,0,0,-1,-1,0,0,1,0,0,0,0,0,0),\\ &(0,0,0,0,0,0,0,1,0,-1,-1,0,1,0,0,0,0,0,0,0),\\ &(0,0,0,0,0,0,0,0,1,-1,-1,1,0,0,0,0,0,0,0,0)\},\end{aligned}$$

Space	Norm of Projection
`space0k3n6`	29.5161
`space1k3n6`	32.5372
`space2k3n6`	24.0167
`space3k3n6`	12.8582

Table 5.2 Projections of `DataVec` into the isotypic subspaces.

$$\begin{aligned}\texttt{space3k3n6} = \operatorname{span}\{&(-1,0,0,1,0,0,1,1,-1,-1,1,1,-1,-1,0,0,-1,0,0,1),\\&(0,-1,0,1,0,0,0,1,0,-1,1,0,-1,0,0,0,-1,0,1,0),\\&(0,0,-1,1,0,0,0,1,-1,0,0,1,-1,0,0,0,-1,1,0,0),\\&(0,0,0,0,-1,0,1,1,0,-1,1,0,-1,-1,0,1,0,0,0,0),\\&(0,0,0,0,0,-1,1,1,-1,0,0,1,-1,-1,1,0,0,0,0,0)\}.\end{aligned}$$

Thus, we see that the pure zeroth order effects space `space0k3n6` is one-dimensional, the pure first order effects space `space1k3n6` is five-dimensional, the pure second order effects space `space2k3n6` is nine-dimensional, and the pure third order effects space is five-dimensional. As such, the direct sum of these orthogonal isotypic subspaces has dimension 20, and so this direct sum is precisely V.

Computing the norms of the projections of the result `DataVec` into these isotypic subspaces, we get Table 5.2. Notice that, while the projections of `DataVec` into `space0k3n6`, `space1k3n6`, `space2k3n6`, and `space3k3n6` each seem reasonably large, the last of these projections, into `space3k3n6`, seems comparatively small, and so we might initially conjecture that the significant information in `DataVec` is best captured by the second order effects statistic. However, we would like to determine the significance of each of these summary statistics after conditioning on the lower order effects, and so we proceed with the approach to election analysis introduced by Diaconis and Sturmfels in Diaconis and Sturmfels (1998).

5.2.2 Computation

We use `4ti2` to compute the Markov bases `mb0k3n6`, `mb1k3n6`, and `mb2k3n6`, for conditioning on the zeroth, first, and second order effects statistics, re-

Mar. basis:	Proj. into	`DataVec`	Mean Proj.	q 1/4	q 1/2	q 3/4	$\frac{\%}{100}$
`mb0k3n6`	`space0k3n6`	29.5161	29.5161	*	*	*	*
	`space1k3n6`	32.5372	5.39775	4.32049	5.1316	6.1101	1.
	`space2k3n6`	24.0167	7.50539	6.30872	7.4027	8.70632	1.
	`space3k3n6`	12.8582	5.60504	4.39697	5.41603	6.37704	1.
`mb1k3n6`	`space0k3n6`	29.5161	29.5161	*	*	*	*
	`space1k3n6`	32.5372	32.5372	*	*	*	*
	`space2k3n6`	24.0167	22.8867	21.5592	22.7772	24.162	0.74
	`space3k3n6`	12.8582	11.0887	9.45163	10.924	12.5831	0.78
`mb2k3n6`	`space0k3n6`	29.5161	29.5161	*	*	*	*
	`space1k3n6`	32.5372	32.5372	*	*	*	*
	`space2k3n6`	24.0167	24.0167	*	*	*	*
	`space3k3n6`	12.8582	12.3838	12.1381	12.4633	12.7017	0.87

Table 5.3 Result after $10,000$ steps of the Markov chain, starting with `DataVec`.

spectively, with $n = 6, k = 3$.

We find that

- The Markov basis `mb0k3n6`, for conditioning on the zeroth order effects statistic, consists of 19 vectors, each of which contain exactly two nonzero entries one of which is a 1 and the other of which is a -1;
- The Markov basis `mb1k3n6`, for conditioning on the first order effects statistic, consists of 69 vectors, each of which contain exactly four nonzero entries, two of which are 1 and two of which are -1; and,
- The Markov basis `mb2k3n6`, for conditioning on the second order effects statistic, consists of 30 vectors, two of which have twelve nonzero entries, and 28 of which have eight nonzero entries, with half of the nonzero entries being 1 and half being -1, in each of these vectors.

For each of these Markov bases, we run a Markov chain with hypergeometric stationary distribution 100 times, with each run starting at `DataVec` and running for $10,000$ steps. Each time, we take the resulting 100 data vectors and compute the norms of their projections into the four isotypic subspaces. The result of these Markov runs is summarized in Table 5.3.

In this table, for each Markov basis we first report the norms of the projections of `DataVec` into the isotypic subspaces of V. We then report the mean and the three quantiles of the norms of the projections of the results from the Markov chain associated with each Markov basis, followed by the percentiles of the norms of the projections of `DataVec` among the set of norms of projections resulting from this Markov chain.

For example, the second row of Table 5.3 provides information about the norms of the projections into `space1k3n6` of the results of a Markov chain with Markov basis `mb0k3n6`. We see that the norm of the projection of `DataVec` into this space is 32.5372. We also see that the mean norm projection of the results of this Markov chain into `space1k3n6` is 5.39775, and that

- 25 percent of the Markov chain results have norm projection less than 4.32049,
- 50 percent of the Markov chain results have norm projection less than 5.1316, and
- 75 percent of the Markov chain results have norm projection less than 6.1101.

The percentile of 1, reported in the last column of the second row, indicates that the projection of `DataVec` into `space1k3n6` has norm greater than or equal to the norm of this projection for each of the 100 results of the Markov chain with Markov basis `mb0k3n6`.

If the projection of `DataVec` into `space1k3n6` were typical of survey results that share their zeroth order statistic with `DataVec`, then we would expect the mean norm projection reported in the second row of Table 5.3 to agree closely with norm of the projection of `DataVec` into `space1k3n6`. We would also expect the norm of this projection of `DataVec` to fall somewhere between the lower an upper quantiles reported in this row, and for the percentile associated with this projection to fall somewhere between .25 and .75.

Recall that when we are conditioning on the ith order statistic, every result of the Markov chain with Markov basis `mbik3n6` will have projections into `space0k3n6,...,spaceik3n6` identical to the projections of `DataVec` into these spaces. Thus, for these projections, the quantiles and the percentile have no significance. In Table 5.3 we have replaced these values with the entry "$*$".

Remark 5.2.1. Our choice here, of running our Markov chain 100 times for 10,000 steps is modeled after the decision of Diaconis and Eriksson (2006: Section 5) to use these same numbers in generating their representative samples. In their paper, Diaconis and Eriksson report their choice of 10,000 steps to be arbitrary, and justify its use by the fact that wide variation in the run times did not change the result significantly enough to alter their conclusions.

In analyzing the results of our Markov chains, we have similarly found that running these processes for 20,000 steps as opposed to 10,000 steps does not significantly alter the resulting data, and that repeated runs of these processes for 10,000 steps produce similar results. By this we mean that, for each run of 10,000 steps, the values in Table 5.3 remain approximately the same, and that these values do not change appreciably when we run these processes instead for 20,000 steps. Thus, in our analysis we have chosen to adopt this Markov chain setup.

A more definite answer to the number of steps necessary for our Markov chains to converge to their desired representative samples would require a thorough analysis of the mixing times of the involved Markov processes. We are not aware that any such analyses have been carried out for the Markov processes involved in the Diaconis–Sturmfels algorithm for studying data of any kind. This would be a fruitful subject for future exploration.

5.2.3 Analysis of Conditional Significance

Conditioning on Zeroth Order Effects: Returning to the discussion above regarding the second row of Table 5.3, recall that the large percentile of 1 associated with the projections into `space1k3n6` indicates that the projection of `DataVec` into this space is atypical of survey results that agree with `DataVec` in their zeroth order effects statistic. The large percentiles in rows three and four of this table similarly indicate that the large norms of the projections of `DataVec` into the pure second and pure third order effects spaces should not be expected, given only the number of responses in `DataVec`.

The unexpected magnitudes of the projections associated with `DataVec`, compared with the magnitudes of the projections associated with conditioning on the zeroth order effects, is even more striking when we notice the great difference between the mean norm projections resulting from this Markov chain and the norms of these projections for `DataVec`. Further, considering the quantiles associated with the projections into `space1k3n6`, `space2k3n6`, and `space3k3n6`, we see that the mean norm projections that we get from our Markov chain lie very near the associated median norm projections. Additionally, in each case the lower and upper quantiles lie near their associated median norm projection, while the associated projection of `DataVec` lies well outside of these quantiles.

With this information, the zeroth order effects statistic does not seem to capture all of the significant information in `DataVec`, by any stretch. That is to say, that if we wish to meaningfully describe the result of our survey, we will want to use higher order effects statistics.

Conditioning on First Order Effects: Considering rows five through eight of Table 5.3, we explore the results of a Markov chain with Markov basis `mb1k3n6`, which conditions on the first order effects statistic (and so also on the zeroth order effects statistic) of `DataVec`. Here, it is only the projections into the pure second and pure third order effects spaces that are interesting. Notice that the percentiles associated with the norms of the projections resulting from this Markov chain are dramatically decreased from 1.

We see that the norm of the projection of `DataVec` into `space2k3n6` lies slightly below the upper quantile associated with this projection from this Markov chain, and so the norm of this projection of `DataVec` is not atypical among survey results that share their first order statistic with `DataVec`. From this we can conclude that the large projection of `DataVec` into the pure second order effects space `space2k3n6` is primarily an artifact of the first order effects statistic.

On the other hand, we see that the norm of the projection of `DataVec` into `space3k3n6` lies just above the upper quantile associated with this projection from this Markov chain, with associated percentile .78. From this, we gather that the projection of `DataVec` into the pure third order effects space is somewhat atypical of survey results that share their first order effects statistic with `DataVec`.

Notice here, that if we were only to consider the difference between the norms of the projections of `DataVec` into the isotypic subspaces of V and the mean norm projections into these subspaces that result from a Markov chain with Markov basis `mb1k3n6`, then we might consider the projection of `DataVec` into `space3k3n6` to be as typical as the projection of `DataVec` into `space2k3n6`, because

$$24.0167 - 22.8867 = 1.13 \quad \text{and} \quad 12.8582 - 11.0887 = 1.7695$$

are similarly valued. This could result in the classification of both the second and third order effects statistics for `DataVec` as residuals of the first order effects statistic. In this way, we see the importance of also considering the quantiles and the percentile associated with these projections, in determining the conditional significance of the different order effects statistics.

Conditioning on Second Order Effects: Considering rows nine through twelve of Table 5.3, we explore the results of a Markov chain with Markov basis `mb2k3n6`, which conditions on the second order effects statistic (and so also the zeroth and first order effects statistics) of `DataVec`. Here, it is only projections into `space3k3n6` that are interesting.

Notice that the norm of the projection of `DataVec` into this space is very close to mean norm projection into `space3k3n6` of the results of the associated Markov chain, and is even very close to the median norm of these projections. However, we see that the norm of this projection of `DataVec` lies well outside of the expected range, as defined by the associated lower and upper quantiles, when conditioning on the second order effects statistic for `DataVec`. In fact, the projection of `DataVec` into the pure third order effects space is as large or larger than the projections into this space of 87 out of the 100 vectors in our representative sample of survey results that share their second order effects statistic with `DataVec`.

In this way, the third order effects statistic for `DataVec` appears to be residual to none of the lower order effects statistics, but rather to capture its own significant information.

5.2.4 Information from Inner Products

From the discussion in Section 5.2.3, we can conclude that the single summary statistic that perhaps best describes the survey result `DataVec` is the third order effects statistic, if we are using the nested collection of summary statistics that arises from the isotypic decomposition of the underlying vector space under the action of S_6. We can also conclude that the second order effects statistic for `DataVec` is an artifact of the first order effects statistic. Further, while this first order effects statistic does not capture everything in `DataVec`, we see that a good deal of information is captured by this summary statistic.

This analysis, however, does not tell us the direction in which these more significant effects statistics lie. That is, while this analysis suggests that most survey participants felt strongly about a single type of Girl Scout cookie, or a triple of Girl Scout cookies, and that most survey participants did not care so much about individual pairs of Girl Scout cookies, this analysis does nothing to indicate which types of cookie or triples of cookies are driving participant preferences.

Given the third and first order effects statistics for `DataVec`, which we have found to best describe this survey result, we would like to be able to determine which 3-subsets and single types of Girl Scout cookies are driving the participant responses. In order to determine the direction in which the participant preferences most closely lie, we consider inner products of `DataVec` with certain interpretable vectors.

If we equip the underlying vector space V with the usual dot product, recall that for two vectors $a, b \in V$ with angle θ between a and b, we have

$$a \cdot b = \|a\| \, \|b\| \cos(\theta).$$

In this way, if a and b are unit vectors, then $a \cdot b = \cos(\theta)$.

If $\{b\}$ is a collection of easily interpretable survey results with unit norm, and a is a survey result that we would like to interpret, then the collection

$$\left\{ \frac{1}{\|a\|} a \cdot b \right\} = \{\cos(\theta)\}$$

is useful in interpreting the result a. If a lies strongly in the direction of b then $\cos(\theta)$ will be close to 1, and if a lies strongly in the opposite direction of b then this value will be close to -1. Whereas, if a lies independently of b, then this value will be close to 0. In this way, the elements of $\{b\}$ with largest associated values in $\left\{ \frac{1}{\|a\|} a \cdot b \right\}$ best summarize the result a.

For $0 \leq i \leq k$, there are $\binom{n}{i}$ rows in the ith order effects statistic matrix T_i associated with a survey in which participants are asked to identify their top k choices of n items. Each of these rows is associated with a specific i-subset of the n items in the survey, and can be viewed as a vector with entries consisting entirely of ones and zeros, and with an entry of 1 wherever this i-subset is included in the k-subset that gives the associated standard basis element. In this way, the rows of T_i form a collection of vectors that are easily interpretable.

Taking $\{b\}$ to be the set of rows of T_i, notice that the product $T_i a$ computes a sequence of dot products. Further, notice that each row of T_i has $\binom{n-i}{k-i}$ entries of 1, and so each row vector has the square root of this value as its norm. In this way, we see that the product

$$\frac{1}{\|a\| \sqrt{\binom{n-i}{k-i}}} T_i a$$

returns a vector whose entries are the values $\frac{1}{\|a\|} a \cdot b$ desired in association with the set $\{b\}$ for an analysis of a.

We carry out this analysis on the projections of `DataVec` into the pure first and pure third order effects spaces, using for $\{b\}$ the rows of the first and third order effects statistic matrices, respectively. Letting `Proj1DataVec` and `Proj3DataVec` be the projections of `DataVec` into the pure first and pure third order effects spaces, respectively, we find

$$\begin{aligned}
&\frac{1}{\|\texttt{Proj1DataVec}\|\sqrt{\binom{6-1}{3-1}}} T_1\,\texttt{Proj1DataVec} \\
&\qquad\qquad = \frac{1}{32.5372\sqrt{10}} T_1\,\texttt{Proj1DataVec}, \\
&\qquad\qquad = \{-0.34, -0.33, 0.30, 0.14, 0.45, -0.23\},
\end{aligned}$$

and also,

$$\begin{aligned}
&\frac{1}{\|\texttt{Proj3DataVec}\|\sqrt{\binom{6-3}{3-3}}} T_3\,\texttt{Proj3DataVec} \\
&\qquad\qquad = \frac{1}{12.8582} T_3\,\texttt{Proj3DataVec}, \\
&\qquad\qquad = \{0.10, 0.27, 0.09, -0.46, -0.22, -0.16, 0.28, \\
&\qquad\qquad\quad -0.07, 0.02, 0.15, -0.15, -0.02, 0.07, -0.28, \\
&\qquad\qquad\quad 0.16, 0.22, 0.46, -0.09, -0.27, -0.10\}.
\end{aligned}$$

Considering the first of these two computed lists, we see that when we consider only the effects of individual cookies on the survey result, there seems to be a preference in the data for cookie 5, with associated value 0.45. This cookie type happens to be Thin Mints.

Considering the second of these two computed lists, we see that when we consider only the effects of triples of cookies on the survey result, there seems to be a preference in the data for the cookie triple $\{3, 4, 5\}$, with associated value 0.46. This happens to be the cookie triple {Samoas, Tagalongs, Thin Mints}.

We can thus conclude our analysis of `DataVec` with the understanding that participants in this survey seem to care both about individual cookie types and about different triples of cookies, with the overall preferences tending towards Thin Mints and the cookie triple {Samoas, Tagalongs, Thin Mints}.

Chapter 6

Future Work

In the Chapters 1, 2, and 3 of this thesis, we have introduced the Diaconis–Sturmfels algorithm for generating a Markov basis, and we have explored the use of a Markov basis together with the Metropolis algorithm in sampling. We have also attempted to present all necessary associated background theory from Markov processes, Gröbner bases, implicitization, and elimination. The gathering of these large ideas and the clean presentation of this theory has required a significant effort, and I view it as one of the more substantial accomplishments of this thesis. The highlight of this background theory is our presentation of the Diaconis–Sturmfels algorithm in Theorem 3.3.2, which hopefully flows naturally from its supporting context in this document.

In Chapters 4 and 5 of this thesis, we have transitioned away from a discussion of the underlying theory of the Diaconis–Sturmfels algorithm and its use, to a discussion of the computational and procedural considerations in using this approach to data analysis. We have chosen to particularly focus on the application of this theory to the study of partially ranked data arising from a survey in which participants are asked to identify there top k choices of n items. This exploration culminates in Section 5.2, with a thorough example of the use of this approach to data analysis, focusing on the result of a survey on Girl Scout cookie preferences.

With the solid understanding of the underlying theory and its practical application presented in the first two portions of this thesis, we find ourselves at an excellent launching point for a variety of possible future explorations.

In Remark 5.2.1, we mentioned the desirability and existing lack of an analysis of the mixing times of the Markov chains involved in this type of data analysis. This would be one possible avenue for future exploration.

We have also mentioned, in Sections 4.2 and 5.1, the limitations that result from the computational intensity of finding a Markov basis for conditioning on a summary statistic. For this reason, given a nested collection of summary statistics, it would be desirable to find a combinatorial description of the associated Markov bases. It would be nice to complete the exploration of the analysis of partially ranked data present in this thesis, with a combinatorial description of the Markov bases for conditioning on the ith order effects statistics, $0 \leq i \leq k$. A conjecture for a degree bound on the Gröbner basis elements associated with these Markov bases for partially ranked data, along the lines of Conjecture 7 of Diaconis and Eriksson (2006) for fully ranked data, would be similarly desirable. However, we leave such a degree bound or combinatorial description open for future work.

At this point in our work, using a combination of Mathematica 7 and `4ti2`, we have developed the computational tools necessary to efficiently carry out the various algorithms associated with this type of data analysis. Thus, given a nested collection of summary statistic matrices and the necessary time to compute the associated Markov bases, carrying out the analysis of a data set associated with these summary statistics requires minimal effort. This brings our attention back to the importance in this analysis, of our choice of a nested collection of summary statistics, which was discussed in Section 4.1. It would be interesting to explore how the analysis of partially ranked data changes with the choice of a different collection of nested summary statistics.

As discussed in Section 4.1, we have carried out a preliminary exploration of the use of statistics associated with the inversion decomposition in the analysis of fully ranked data. It would be interesting, and most likely rather straightforward, to carry out a more thorough exploration of the use of the inversion decomposition for fully ranked data in this type of analysis. It would also be interesting to consider possible approaches for extending the inversion decomposition to partially ranked data.

Appendix A

Computed Markov Bases for Partially Ranked Data

In this Appendix, we summarize the results of our computation of Markov basis elements for studying partially ranked data. We consider surveys in which participants are asked to identify their top k choices of n items. We choose to carry out our analysis with respect to the ith order effects statistics, for $0 \leq i \leq k$, which simply count the number of times that each i-subset of the n items is chosen by the survey participants.

We group our computed Markov bases together into different tables according to i, and then according to the values of n and k. For each computed Markov basis, we report the degrees of the Markov basis elements, together with the number of elements of each degree.

n	k	Degree	# Elements
2	1	1	1
3	1	1	2
4	1	1	5
	2	1	5
5	1	1	4
	2	1	9
6	1	1	5
	2	1	14
	3	1	19
7	1	1	6
	2	1	20
	3	1	34
8	1	1	7
	2	1	28
	3	1	55
	4	1	69
9	1	1	8
	2	1	35
	3	1	83

Table A.1 Conditioning on zeroth order effects; $i = 0$.

n	k	Degree	# Elements
4	2	2	2
5	2	2	10
6	2	2	30
	3	2	69
7	2	2	70
	3	2	273
8	2	2	140
	3	2	812
	4	2	8
9	2	2	252
	3	2	2016

Table A.2 Conditioning on first order effects; $i = 1$.

n	k	Degree	# Elements
6	3	4	15
		6	15
7	3	4	105
		6	420
8	3	4	420
		6	4620
		8	68355
		10	58800
		12	4410
		14	2800

Table A.3 Conditioning on second order effects; $i = 2$.

n	k	Degree	# Elements
8	4	5	56

Table A.4 Conditioning on third order effects; $i = 3$.

Bibliography

4ti2 team. 2011. 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.

Cox, David, John Little, and Donal O'Shea. 1997. *Ideals, Varieties, and Algorithms*. Undergraduate Texts in Mathematics, Springer-Verlag.

Diaconis, Persi. 1988. *Group Representations in Probability and Atatistics*, *Institute of Mathematical Statistics Lecture Notes—Monograph Series*, vol. 11. Hayward, CA: Institute of Mathematical Statistics.

Diaconis, Persi, and Nicholas Eriksson. 2006. Markov bases for noncommutative fourier analysis of ranked data. *Journal of Symbolic Computation* 41:182–185.

Diaconis, Persi, and Bernd Sturmfels. 1998. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics* 26:363–397.

Drton, Mathias, Bernd Sturmfels, and Seth Sullivant. 2009. *Lectures on Algebraic Statistics*, *Oberwolfach Seminars*, vol. 39. Birkhauser Verlag AG.

Hansen, Michael W., and Michael E. Orrison. 2008. Spectral analysis and Markov bases for approval voting data. Preprint.

Karlin, Samuel, and Howard M. Taylor. 1975. *A First Course in Stochastic Processes*. New York, NY: Academic Press, Inc., 2nd ed.

Marden, John I. 1995. *Analyzing and Modeling Rank Data*. London: Chapman and Hall.

Maslen, David K., Michael E. Orrison, and Daniel N. Rockmore. 2004. Computing isotypic projection with the Lanczos iteration. *SIAM Journal on Matrix Analysis and Applications* 25:784–803.

Riccomagno, Eva. 2009. A short history of algebraic statistics. *Metrika* 69:397–418.

Ross, Sheldon M. 1996. *Stochastic Processes*. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, Inc., 2nd ed.

———. 1997. *Simulation*. Statistical Modeling and Decision Science, Hardcourt/Academic Press, 2nd ed.

———. 2002. *Probability Models for Computer Science*. Harcourt/Academic Press.

Sturmfels, Bernd. 1996. *Gröbner Bases and Convex Polytopes*, *University Lecture Series*, vol. 8. American Mathematical Society.