Abstract

BOCK, BRANDON WILLIAM. Algebraic and Combinatorial Properties of Statistical Models for Ranked Data. (Under the direction of Seth Sullivant.)

Statistical models involving the ranking of items have been around for over a century. These models can be somewhat cumbersome to use in practice. A total ranking on a list of n items has a natural correspondence with a permutation in S_n. This can be problematic when using models, as the number of permutations increase with size $n!$, which will also increase the necessary amount of observed data to yield accurate results. These models must also address the matter of contradictions within observations as well as provide methods for computing expected values of random variables when the observed data is a permutation. Models using partially ranked data alleviate some of these problems but create some new problems of their own. Both of these types of models can be sensitive to noise, another problem which makes using them in practice somewhat difficult. In, this paper, we seek to examine the algebraic and combinatorial of a few models which fall into these categories.

In Chapter 2 we examine the Mallows Model. The Mallows model is a discrete log-linear model which assigns a probability to every permutation in S_n, where n is the number of items being ranked. This probability corresponds to the probability of observing a given ranking. We analyze the algebraic and combinatorial aspects of this model. We then propose a Mallows mixture model, a simple mixture model with two underlying distributions, both of which are Mallows models. First we develop the tools necessary to analyze this model from an algebraic standpoint. We then analyze the combinatorial and algebraic aspects of this model, enabling us to compute a vanishing ideal on the model and greatly reduce Gröbner basis calculation time by eliminating extraneous equations which will always be true for this model. In the mixture model, we find the probability of observing a permutation depends on the distance of this permutation from the two “centers" of the underlying Mallows models.

In Chapter 3, we look at a generating function which will count the number of permutations which are distance i from a fixed permutation π and distance j from a fixed permutation σ. The generating function is necessary for any practical application of the Mallows mixture model introduced in Chapter 2. We analyze this generating function, which we call the bi-distance polynomial, and provide a closed form equation for calculating the number of permutations in S_n which are distance i from a fixed permutation π and distance j from a fixed permutation σ. We also discuss exactly when this bi-distance polynomial is factorable and give a set of guidelines which allow us to predict exactly how factorable the bi-distance polynomial is based on which permutations it is centered around.

In Chapter 4, we introduce a Thurstonian type model which has been adapted to be used in cases where the observed data is partially ranked data, as opposed to the traditional fully ranked data. After introducing the model and the underlying assumptions, we lay out how to implement different statistical methods–such as Maximum Likelihood Estimation and Bayesian posterior distribution

calculation via the EM algorithm and a Gibbs sampler respectively–on the model. We use the model and apply two different methods of parameter estimation to two different data sets and analyze the results.

Algebraic and Combinatorial Properties of Statistical Models for Ranked Data

by
Brandon William Bock

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Mathematics

Raleigh, North Carolina

2016

APPROVED BY:

Ernest Stitzinger	Agnes Szanto
Nathan Reading	Elena Jakubikova
Jeffrey Thorne	Seth Sullivant Chair of Advisory Committee

Dedication

To my parents and my loving wife.

Biography

The author was born in Battle Creek, Michigan. At five he and his family moved to Bryan, Ohio. This is where he grew up. He went to Wittenberg University in Springfield, Ohio, where he studied Math Education. He also met the most beautiful woman ever. He convinced her to marry him some 8 years later.

The author attended North Carolina State University for graduate school.

Acknowledgements

I would like to thank my advisor for his help. Without the help and support of my advisor, I would never have been able to finish. There is no shortage of praise I can give to my advisor, Dr. Sullivant. His patience and encouragement were seemingly limitless, and without it I surely would never have finished this thesis.

To my parents, who listened to me over the phone for endless hours and still had the grace to tell me that I was capable of completing what it is I had started, I give my unending gratitude.

To my beautiful and wonderful wife, I can never give enough praise. Her patience, love, support, and encouragement came without complaint, and she constantly put me before herself. I love her until the end of time, and can confidently say that without her I would never have made it out of graduate school.

Table of Contents

LIST OF TABLES . . . . . **vi**

LIST OF FIGURES . . . . . **vii**

Chapter 1 Introduction **1**
- 1.1 Background in Algebraic Geometry 2
- 1.2 Background in Combinatorics 8
- 1.3 Background in Statistics 14
- 1.4 Models for Ranked and Partially Ranked Data 25
- 1.5 Outline of Thesis 29

Chapter 2 The Mallows Mixture Model **30**
- 2.1 The Mallows Model 31
- 2.2 The Mallows Mixture Model 37
- 2.3 Set of Bi-Distance Pairs 40
- 2.4 Joins of Ideals 49
- 2.5 Vanishing Ideal of Mallows Mixture 57
- 2.6 Computations and Conjecture 61

Chapter 3 Characterizing the Bi-Distance Polynomial f_τ **64**
- 3.1 f_τ as a Generating Function 65
- 3.2 Factoring f_τ when $\tau = \pi \oplus \sigma$ and $\tau = \pi \ominus \sigma$ 67
- 3.3 Permutations with Contiguous Blocks 75

Chapter 4 Partial Rank Thurstonian Model **81**
- 4.1 Introduction 81
- 4.2 Model Description 83
- 4.3 Bayesian Methods 85
- 4.4 MLE For Thurstonian Model 87
- 4.5 Results 90
  - 4.5.1 HIV Dataset 90
  - 4.5.2 Prostate Cancer Cell Dataset 93

BIBLIOGRAPHY **96**

LIST OF TABLES

Table 2.1 Substrings of R_1 of the form $\alpha_j^{(1)} \cdots \alpha_m^{(1)}$ with $j = 1, \ldots, m$ and their corresponding element in $\mathscr{G}(\mathrm{id}, \tau)$ 46
Table 2.2 Comparing each point of the paths R_1, R_2 where R_2 is obtained by performing the braid transformation $s_1 s_3 = s_3 s_1$ to the word R_1. 48
Table 2.3 Sequence of braid transformations from $R_1, \ldots, R_8$ and the (i, j) pairs found in each $F_\tau(R_k)$. 49
Table 2.4 The number of degree 1, 2, 3, and 4 generators for m, r 62

Table 4.1 Paired comparison matrix for HIV data 92
Table 4.2 Paired comparison matrix for prostate cancer data 94

LIST OF FIGURES

Figure 1.1 Graph of the variety $V(y-x^3)$ in $\mathbb{R}^2$ 4
Figure 1.2 Graph of the variety $V(y-x^3)$ in $\mathbb{R}^3$ 4
Figure 1.3 Hasse diagram of the poset defined by all subsets of $\{x,y,z\}$ ordered by set inclusion. 9
Figure 1.4 Visualization of permutation 2431 in S_4 11
Figure 1.5 Visualization of composition of the permutation 2431on the left with 3241 in S_4 11
Figure 1.6 The graph of the probability density function for a Gaussian random variable X with $\mu=0, \sigma=1$ 17

Figure 2.1 The Cayley graph of S_4 32

Figure 3.1 The composition of $\tau\gamma$ when τ is a direct sum, with disjoint subsets highlighted 71
Figure 3.2 The composition of $\tau\gamma$ when τ is a skew sum, with disjoint subsets highlighted 73
Figure 3.3 The composition of $\tau\gamma$ when τ is a permutation consisting of contiguous blocks, with disjoint subsets highlighted 78
Figure 3.4 The composition of $\tau\gamma$ when τ is a permutation consisting of contiguous blocks, highlighting the permutation pattern of 123 in $\tau\gamma$ 79

Figure 4.1 Graphical representation of proposed model 84
Figure 4.2 The mutations corresponding to variables X_i and the maximum likelihood poset proposed in [3]. 91
Figure 4.3 Maximum Likelihood Poset for prostate cancer data 94

CHAPTER

1

INTRODUCTION

In this thesis, we examine statistical models where the observed data are rankings or partial rankings. Throughout the thesis, we we will refer to such models as statistical models for ranked data (or partially ranked data, though context should make clear which we mean). Examples of these models exist in many different fields of study. This is because there are many different situations in which studying ranked data is useful. As an example from the field of cognitive science, Steyvers et al. gave participants a list of 10 former U.S. presidents and asked them to order the presidents according to when they served [40]. They then attempted to reconstruct the most likely order based on all the participants responses and compared the reconstruction to the true ranking. The authors looked at different reconstruction methods to evaluate which performed best and most consistently. Or consider the model proposed by Beerenwinkel and Sullivant to predict mutation accumulation on a cellular level [3]. In this model, a cell was observed and tested to see which mutations is a specific set had occurred and which had not occurred, but there was no way to determine the order in which the mutations occurred. The first example is one that uses a statistical model for ranked data while the second is an example of a model for partially ranked data. These examples are just two out of many conceivable potential scenarios. While it is not hard to think of where models for ranked or partially ranked data can be used, such models present and interesting and unique set of complications. In this thesis, we seek to address some of the different aspects which can make these models difficult to work

with as well as examine their underlying structure from an algebraic and combinatorial perspective. Before we do, we examine the aspects of statistical models for ranked and partially ranked data in greater detail.

In this chapter, we introduce the underpinnings from algebraic geometry, combinatorics, and statistics which are used throughout the thesis. In Section 1.1, we will define the pertinent constructs from algebraic geometry, establish notation, and remind the reader of a few basic theorems from the field which will be used in this thesis. In Section 1.2, we provide the definitions from Combinatorics which will prove useful in this thesis. Furthermore, we explain the connection between ranked data and permutations, introduce theorems which will be useful while working with permutations and generating functions, and establish the notational conventions used for these concepts. In Section 1.3, we examine concepts from statistics to acquire the necessary background to understand the entirety of the thesis. This section is in no way intended to provide a comprehensive overview of all of statistics; as this is a thesis in Mathematics (as opposed to Statistics), we will be focusing only on the areas of statistics which will be relevant to the topics presented in this thesis. In Section 1.4, we will introduce some well known statistical models for ranked and partially ranked data and highlight some of the ways they have been used throughout various disciplines.

1.1 Background in Algebraic Geometry

In this thesis, we will make liberal use of the many tools afforded to us by algebraic geometry. We will cover some basic concepts of algebraic geometry which will be necessary for understanding the thesis. In order to do this, we first make explicit some of the notation we use as well as remind the reader of some important concepts in Algebraic Geometry. We let k denote a field, $k[x_1,\dots,x_n]$ denote the polynomials in variables $x_1,\dots,x_n$ with coefficients in k. We will often refer to k as the base field and shorten the notation $k[x_1,\dots,x_n]$ to $k[\mathbf{x}]$ in the places where it is clear that $\mathbf{x}$ is short form of $x_1,\dots,x_n$. A nonzero polynomial

$$f = \sum_{\alpha_1,\dots,\alpha_n} c_{\alpha_1,\dots,\alpha_n} x_1^{\alpha_1}\cdots x_n^{\alpha_n} \qquad c_{\alpha_1,\dots,\alpha_n} \in k$$

which we say has degree (or total degree) d if $c_{\alpha_1,\dots,\alpha_n} = 0$ whenever $\alpha_1+\cdots+\alpha_n > d$ and $c_{\alpha_1,\dots,\alpha_n} \neq 0$ for some index $\alpha_1+\cdots+\alpha_n = d$. We will frequently denote the degree of a polynomial f as $\deg(f)$. We can then define the polynomial as *homogeneous* if $c_{\alpha_1,\dots,\alpha_n} = 0$ for all $\alpha_1+\cdots+\alpha_n \neq d$. Whenever it is convenient to do so, we will use the multi-index notation

$$f = \sum_{\alpha} c_\alpha \mathbf{x}^\alpha$$

with $\alpha = (\alpha_1, \ldots, \alpha_n)$, $c_\alpha = c_{\alpha_1,\ldots,\alpha_n} \in k$ and $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_n$.

We will let $P_{n,d} \subset k[x_1, \ldots, x_n]$ denote the vector subspace of polynomials of degree $\leq d$. We know the monomials

$$x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$$

form a basis for $P_{n,d}$. Thus, we know

$$\dim P_{n,d} = \binom{n+d}{n} .$$

Furthermore, given distinct points $p_1, \ldots, p_N \in \mathbb{A}^n(k)$, we let $I_d(p_1, \ldots, p_N)$ be the vector space of polynomials of degree $\leq d$ which vanish at each of the points $p_1, \ldots, p_N$. We will also make use of some basic definitions found throughout the algebraic geometry literature. First, we define an affine space.

Definition 1.1.1. Given a field k and a positive integer n, we define the n-dimensional *affine space* over k to be the set

$$\mathbb{A}^n(k) := \{(a_1, \ldots, a_n) \mid a_i \in k\} .$$

We will sometimes denote $\mathbb{A}^n(k)$ simply as k^n.

The classic example of an affine space is the case where $k = \mathbb{R}$, in which case the n-dimensional affine space would simply be $\mathbb{R}^n$.

Definition 1.1.2. Given $S \subset \mathbb{A}^n(k)$, the number of conditions imposed by S on polynomials of degree $\leq d$ is defined as

$$C_d(S) := \dim P_{n,d} - \dim I_d(S) .$$

S is said to *impose independent conditions on* $P_{n,d}$ if $C_d(S) = |S|$. Otherwise, we say it *fails to impose independent conditions*.

Next we can define a hypersurface.

Definition 1.1.3. Given a field k and a polynomial $f \in k[x_1, \ldots, x_m]$ where $\deg(f) = d$. We define the *hypersurface* of degree d as

$$V(f) := \left\{(a_1, \ldots, a_m) \in \mathbb{A}^m \mid f(a_1, \ldots, a_m) = 0\right\} \subset \mathbb{A}^m(k) .$$

This is highly reminiscent of our definition of a variety.

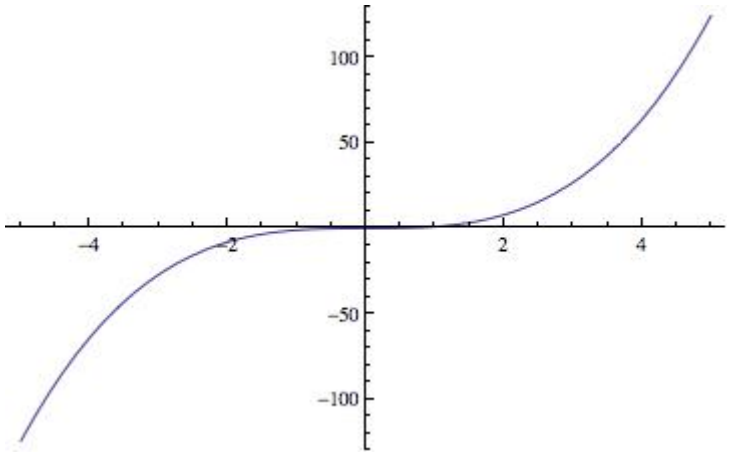

Figure 1.1 Graph of the variety $V(y-x^3)$ in $\mathbb{R}^2$

Definition 1.1.4. Let k be a field and $f_1,\ldots,f_s \in k[x_1,\ldots,x_n]$. Then the *affine variety* defined by $f_1,\ldots,f_s$ is the set

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

We consider two basic examples.

Example 1.1.5. Consider the variety in $\mathbb{R}^2$ given by the single polynomial $V(y-x^3)$. We can use any computer algebra system to plot this variety. Here, we use Mathematica.

Note that if we were to consider the same variety in $\mathbb{R}^3$ rather than $\mathbb{R}^2$, the variety would look very different.

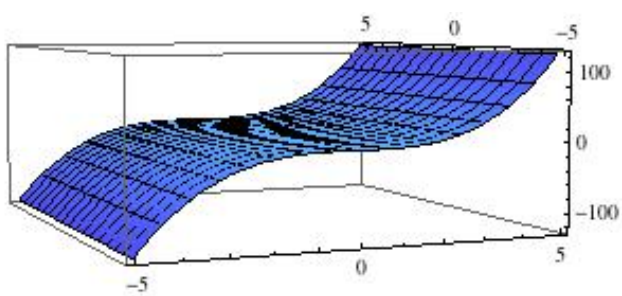

Figure 1.2 Graph of the variety $V(y-x^3)$ in $\mathbb{R}^3$

We assume the reader is familiar with ideals and monomial orderings and begin by defining the leading term of a polynomial.

Definition 1.1.6. Fix a monomial order on $k[x_1,\dots,x_n]$ and consider any nonzero polynomial

$$f = \sum_{\alpha} c_\alpha \mathbf{x}^\alpha \ .$$

The *leading term* of f, denoted $\mathrm{LT}(f)$, is the term $c_\alpha \mathbf{x}^\alpha$ such that $\mathbf{x}^\alpha$ is the largest monomial such that $c_\alpha \neq 0$.

Recall that an ideal is called *homogeneous* if it is generated entirely by homogeneous polynomials.

Definition 1.1.7. Given $I \subset k[x_1,\dots,x_n]$ an ideal, we define the *ideal of leading terms*

$$\mathrm{LT}(I) := \langle \mathrm{LT}(g) \mid g \in I \rangle$$

In a slight abuse of notation, if given a set of polynomials $G = \{g_1,\dots,g_s\}$, we let

$$\mathrm{LT}(G) := \{\mathrm{LT}(g_i) \mid g_i \in G\}$$

This definition allows us to define the idea of a Gröbner basis.

Definition 1.1.8. Fix a monomial order and let $I \subset k[x_1,\dots,x_n]$ be an ideal. A *Gröbner basis* for I is a set of nonzero polynomials $\{f_1,\dots,f_s\} \subset I$ such that $\mathrm{LT}(f_1),\dots,\mathrm{LT}(f_s)$ generate $\mathrm{LT}(I)$.

We know the following is true about Gröbner basis.

Theorem 1.1.9. *Fix a monomial order and let $I \subset k[x_1,\dots,x_n]$ be an ideal. Let $f_1,\dots,f_s$ be a Gröbner basis for I. Then $I = \langle f_1,\dots,f_s \rangle$.*

This well known result is just part of why Gröbner bases are such a powerful tool of algebraic geometry. It is also known that multivariate polynomial division using a Gröbner basis yields a unique remainder, thus answering the question posed by the ideal membership problem. It is well known that a Gröbner basis is not unique and depends largely on the choice of the monomial order.

We have already seen that an ideal can define a variety. We can also define the ideal defined by an affine variety.

Definition 1.1.10. Let $V \subset \mathbb{A}^n(k)$ be an affine variety. Then we define the set $\mathbf{I}(V)$ to be

$$\mathbf{I}(V) := \{f \in k[x_1,\dots,x_n] \mid f(a_1,\dots,a_n) = 0 \text{ for all } (a_1,\dots,a_n) \in V\}$$

More importantly, $\mathbf{I}(V)$ is an ideal (see, for example [10]). It is also true that an ideal is contained in the ideal of its variety.

Theorem 1.1.11. *If $f_1,\dots,f_s \in k[x_1,\dots,,x_n]$, then $\langle f_1,\dots,f_s\rangle \subset \mathbf{I}(V(f_1,\dots,f_s))$.*

It should be noted that this containment is an equality only when $\langle f_1,\dots,f_s\rangle$ is a radical ideal. The following propositions found in Hassett [22] are useful for understanding the relationship between a varieties generated by ideals and the ideals generated by that varieties.

Proposition 1.1.12. *For every collection of polynomials $F = \{f_j\}_{j\in J} \subset k[x_1,\dots,x_n]$ and each subset $F' \subset F$, we have that $V(F') \supset V(F)$.*

Proposition 1.1.13. *Given a collection of polynomials $F = \{f_j\}_{j\in J} \subset k[x_1,\dots,x_n]$ generating an ideal $I = \langle f_j\rangle_{j\in J}$, we have $V(F) = V(I)$.*

Proposition 1.1.14. *For any subsets $S' \subset S \subset \mathbb{A}^n(k)$ we have $\mathbf{I}(S') \subset \mathbf{I}(S)$.*

We also know that an arbitrary intersection of varieties is a variety and a finite union of varieties is a variety.

Because we will make use of the pull-back of morphisms, we will define it here.

Definition 1.1.15. Choose coordinates $x_1,\dots,x_n$ and $y_1,\dots,y_m$ on $\mathbb{A}^n(k)$ and $\mathbb{A}^m(k)$ respectively. Let $\phi : \mathbb{A}^n(k) \to \mathbb{A}^m(k)$ be a morphism given by the rule

$$\phi(x_1,\dots,x_n) = \big(\phi_1(x_1,\dots x_n),\dots,\phi_m(x_1,\dots,x_n)\big)$$

where $\phi_j \in k[x_1,\dots,x_n]$. Then for each $f \in k[y_1,\dots,y_m]$, the *pull-back* by ϕ is defined

$$\phi^* f = f\circ\phi = f\left(\phi_1(x_1,\dots x_n),\dots,\phi_m(x_1,\dots,x_n)\right) .$$

We then have the ring homomorphism

$$\begin{aligned} \phi^* : k[y_1,\dots,y_m] &\longrightarrow k[x_1,\dots,x_n] \\ y_j &\mapsto \phi_j(x_1,\dots,x_n). \end{aligned}$$

Furthermore, ϕ^* has the property $\phi^*(c) = c$ for all constants $c \in k$ and is therefore a k-algebra homomorphism.

It is true that there is a natural correspondence between morphisms and k-algebra homomorphisms (see, for instance, [22]).

Recall the definition of the sum of two ideals.

Definition 1.1.16. Given a ring R and a collection of ideals $\{I_j\}_{j\in J}$ in R. The *sum* of these ideals is the ideal

$$\sum_{j\in J} I_j := \left\{ f_1 + \cdots + f_s \mid f_j \in I_j \text{ for some } j \right\}$$

In other words, the ideal consisting of all finite sums of elements each taken from one of the I_j

Before we define the join of ideals and the join of varieties, we will introduce some notation. We will let Δ_N denote the variety

$$\Delta_N := \{(t_1, \ldots, t_N) \mid t_1 + \cdots + t_N = 1\} \subset \mathbb{A}^N(k)$$

We also know that for every finite set of points $S = \{p_1, \ldots, p_N\} \subset \mathbb{A}^n(k)$, there is a morphism

$$\begin{aligned} \sigma_S : \Delta_N &\to \mathbb{A}^n \\ (t_1, \ldots, t_N) &\mapsto t_1 p_1 + \ldots, t_N p_N \end{aligned}$$

where we add the p_j as vectors in k^n. The image is called the *affine span of* S in $\mathbb{A}^n(k)$ and is denoted affspan(S). The following proposition can be found in [22].

Proposition 1.1.17. *The set $S = \{p_1, \ldots, p_N\}$ imposes independent conditions on polynomials of degree ≤ 1 if and only if σ_S is injective. We say that S is in linear general position.*

We will use the definition presented by Sidman and Sullivant in [38].

Definition 1.1.18. Given a collection of ideals $I_1, \ldots, I_r \subset k[x_1, \ldots, x_n]$. The *join* of $I_1, \ldots, I_r$ is the ideal

$$I_1 * \cdots * I_r := \left(I_1(\mathbf{y}_1) + \cdots + I_r(\mathbf{y}_r) + \langle x_j - \sum y_{ij} \mid j \in [n] \rangle \right) \bigcap k[\mathbf{x}]$$

where $\mathbf{y}_i$ is a new set of variables $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$ and $I_i(\mathbf{y}_i)$ denotes the ideal obtained from I_i by substituting the variable y_{ij} for the variable x_j. It should be noted that the large ideal in the parentheses is contained in the ring $k[\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_m]$.

Definition 1.1.19. Let $V_1, \ldots, V_N \subset \mathbb{A}^n$ be affine varieties. The *join* of these varieties, denoted Join$(V_1, \ldots, , V_N) \subset \mathbb{A}^n$, is defined as the closure of the image

$$\begin{aligned} V_1 \times \cdots \times V_N \times \Delta_N &\to \mathbb{A}^n \\ (v(1), \ldots, v(N), (t_1, \ldots, t_N)) &\mapsto t_1 v(1) + \ldots + t_N v(N) \,. \end{aligned}$$

There is a nice relationship between the join of two varieties and the variety of the join of those two ideals. In short, given to ideals I, J and considering their varieties $V(I), V(J)$, we know that the variety $V(I * J)$ is the join of the varieties $V(I)$ and $V(J)$. Geometrically, the join of two varieties V, W is the union of all points which lie on a line which contains a point in V and a point in W. In other words, it is the union of all lines which pass through V and W.

1.2 Background in Combinatorics

Throughout the thesis, we will make us of the notation $[n]$ to denote the set $\{1, \ldots, n\}$.

Because we will be dealing extensively with partially ordered data, it will make sense to define a poset.

Definition 1.2.1. A *partially ordered set* (or *poset*) P is a set on which there is some binary relation $\leq$ which satisfies the following properties:

1. For all $x \in P$, $x \leq x$ (reflexivity)
2. If $x \leq y$ and $y \leq x$, then $x = y$ (anti-symmetry)
3. If $x \leq y$ and $y \leq z$, then $x \leq z$ (transitivity)

Within a poset P, we say that two items x and y are *comparable* if either $x \leq y$ or $y \leq x$. If neither of these is true, we say the items x, y are *incomparable*. We will say that a poset P *has an element* $\hat{0}$ if there exists an element $\hat{0} \in P$ such that for all $x \in P$, $\hat{0} \leq x$. Similarly P *has an element* $\hat{1}$ if there exists an element $\hat{1} \in P$ such that for all $x \in P$, $x \leq \hat{1}$. If $s, t \in P$, then we say t *covers* s (or s *is covered by* t) if $s < t$ and there does not exists an element $r \in P$ such that $s < r < t$. It is known that a locally finite poset is completely determined by such cover relations. The *Hasse diagram* of a finite poset P is the graph whose vertices are elements of P and whose edges are cover relations, where if $s < t$ then the vertex t is drawn with higher vertical coordinate than that of the vertex s.

Example 1.2.2. As a standard example, we can create a poset out of the $2^{[n]}$ subsets of any set $[n]$ with the order relation being the standard set inclusion (i.e., $S \leq T$ in the poset if $S \subset T$). We consider the set $\{x, y, z\}$ and the poset consisting of all subsets of this set. We can visualize this poset by creating a Hasse diagram. The Hasse diagram for this poset can be found in Figure 1.3.

We see that any two distinct elements in our poset which have the same set cardinality are incomparable. These are not the only pairs of incomparable elements; we know the elements $\{x\}$ and $\{y, z\}$ are also incomparable, for instance.

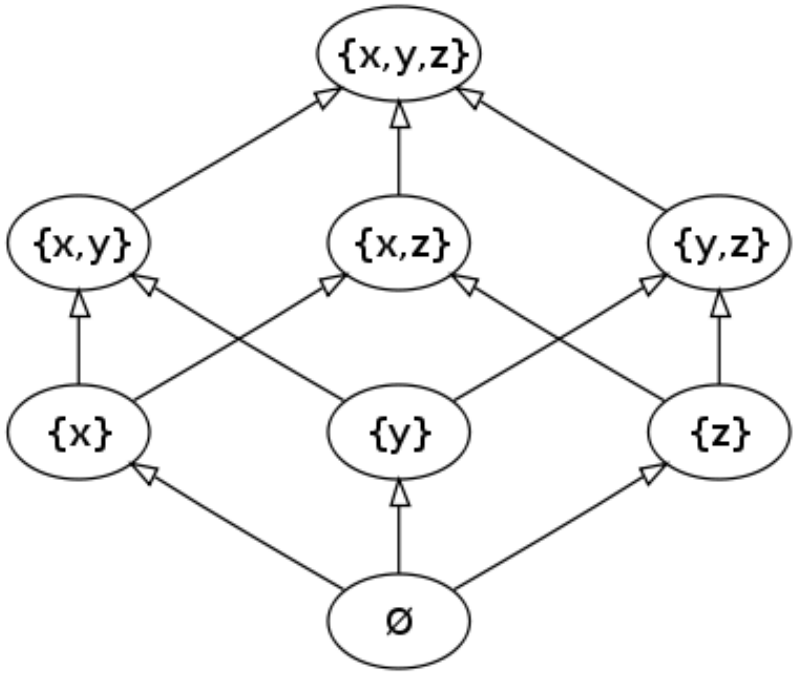

Figure 1.3 Hasse diagram of the poset defined by all subsets of $\{x, y, z\}$ ordered by set inclusion.

When dealing with ranked data, we will work with permutations rather than posets. We will now define the aspects of permutations we will use in this thesis. Recall that a permutation $\pi \in S_n$ is a bijective map from $[n]$ to $[n]$. While there are different ways to denote a permutation, we will use one line notation exclusively. To remind the reader, any permutation $\pi \in S_n$ can be written as $\pi = \pi_1 \cdots \pi_n$ which indicates the permutation maps i to π_i (again, both $i, \pi_i \in [n]$). We can define an inversion of a permutation $\pi = \pi_1 \cdots \pi_n$ to be a pair $(i, j) \in [n] \times [n]$ such that $i < j$ and $\pi_i > \pi_j$. Then we have the following

Definition 1.2.3. We denote the number of inversions of a permutation $\pi \in S_n$ as $\mathrm{inv}(\pi)$ and therefore

$$\mathrm{inv}(\pi) := \#\{(i, j) \in [n] \times [n] \mid i < j \text{ and } \pi(i) > \pi(j)\} .$$

We know that for any $\pi \in S_n$, $0 \leq \mathrm{inv}(\pi) \leq \binom{n}{2}$. We know that the symmetric group S_n is completely generated by the adjacent transpositions, and that the minimum number of adjacent transpositions required to generate an element $\pi \in S_n$ is the same as $\mathrm{inv}(\pi)$. Let π^{-1} denote the inverse of π in S_n (i.e. $\pi\pi^{-1} = \pi^{-1}\pi = \mathrm{id}$). Then for any $\pi \in S_n$ it is fairly simple to demonstrate that $\mathrm{inv}(\pi) = \mathrm{inv}(\pi^{-1})$ (as π can be minimally generated by a line of adjacent transpositions, reading these transpositions in

reverse would yield π^{-1}).

While we will denote a permutation in one line notation throughout the thesis, we will make use of permutation matrices. Given a permutation $\pi \in S_n$ with $\pi = \pi_1 \cdots \pi_n$, the permutation matrix of π, denoted M_π, will be the $n \times n$ matrix with $M_{i,j} = 1$ if $\pi_j = i$ (i.e. $\pi(j) = i$) and 0 otherwise. Permutation matrices always have exactly one 1 entry in every row and column (i.e. they are elementary matrices), and are therefore always of full rank.

Example 1.2.4. Consider the permutation $\pi = 2431 \in S_4$. The permutation sends 1 to 2, 2 to 4, 3 to itself, and 4 to one, as demonstrated in Figure 1.4. Furthermore, we we can find the permutation matrix of 2431 and it will have the form

$$M_{2431} = M_\pi = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

We know that a composition of permutations is the composition of the two bijective functions, so we examine a composition of permutations. Now suppose we consider the permutation $\sigma = 3241$ with permutation matrix

$$M_{3241} = M_\sigma = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We know that the composition $\pi\sigma$ will apply the mapping σ first and then apply the mapping π. The result is that first 1 will map to 3 (via σ) which will map to 3 (via π), 2 will map to 2 which will map to 4, etc., and the resulting composition will be the permutation 3412, as shown in Figure 1.5. The result can also be achieved by multiplying the permutation matrices:

$$M_\pi M_\sigma = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

We know the symmetric group S_n can be generated by the $n-1$ adjacent transpositions, those permutations which swap a pair of adjacent numbers. Any permutation can be written as a product of

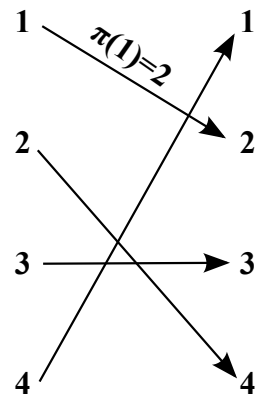

Figure 1.4 Visualization of permutation 2431 in S_4

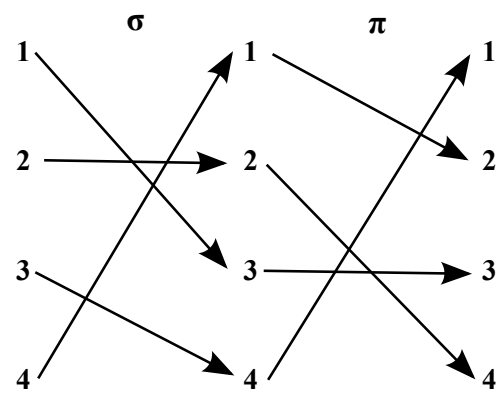

Figure 1.5 Visualization of composition of the permutation 2431on the left with 3241 in S_4

permutations which have one line notation $\epsilon_i = 1 \cdots i+1\ i \cdots n$. For all permutations $\pi \in S_n$, we can write π as a product of adjacent transpositions, $\prod_i \epsilon_{j_i}$. If we were to think of permutation matrices, these adjacent transpositions would correspond to permutation matrices of the form

$$\begin{pmatrix} 1 & 0 & \cdots & & \cdots & & 0 \\ 0 & \ddots & & & & & \\ \vdots & & 0 & 1 & & & \vdots \\ & & 1 & 0 & & & \\ \vdots & & & & 1 & & \\ & & & & & \ddots & 0 \\ 0 & \cdots & & & \cdots & 0 & 1 \end{pmatrix}.$$

The sign of a permutation can be defined in terms of the number of adjacent transpositions required to generate that permutations. If we let $\epsilon_1, \ldots, \epsilon_{n-1}$ be the $n-1$ adjacent transpositions which generate S_n. Then for any $\pi \in S_n$, $\pi = \epsilon_{i_1} \cdots \epsilon_{i_k}$ and we define the *sign of* π is defined as

$$\operatorname{sgn}(\pi) := (-1)^k \quad .$$

Permutations in S_n can generally be thought of as a specific ordering of the set $[n]$. While every element of $[n]$ is unique, we can extend the idea a permutation to be an ordering of a multiset. Recall that a multiset behaves in many ways like a set, with the exception that there can be repeated elements and elements that are repeated are indistinguishable from one another. For example, $M = \{1,1,2,3\}$ is a multiset where both of the elements 1 are treated as identical.

Definition 1.2.5. Let M be any multiset. Let $\mathfrak{S}_M$ denote the set of permutations of the multiset M.

Example 1.2.6. Consider the multiset $M = \{1,1,2\}$. Then we can list all the permutations of M:

$$\mathfrak{S}_M = \{112, 121, 211\} \, .$$

If we instead let $M = \{1,1,2,3\}$, we will have

$$\mathfrak{S}_M = \{1123, 1213, 1231, 1132, 1312, 1321, 2113, 2131, 2311, 3112, 3121, 3211\} \, .$$

Definition 1.2.7. The polynomial

$$1 + q + q^2 + \cdots + q^{n-1} = \frac{1-q^n}{1-q}$$

is denoted $(\mathbf{n})$ and is called the q-analogue of n.

Then we can define the q-analogue of $n!$ as

$$(\mathbf{n})! = (\mathbf{1})(\mathbf{2})\cdots(\mathbf{n}) = 1(1+q)(1+q+q^2)\cdots(1+q+\cdots+q^{n-1}) \, .$$

Similarly, the q-analogue of n choose k is

$$\binom{\mathbf{n}}{\mathbf{k}} = \frac{(\mathbf{n})!}{(\mathbf{n}-\mathbf{k})!(\mathbf{k})!}$$

Finally, we can consider the q-analogue of $n!$ evaluated at a function $g(\mathbf{x})$. If we let $f(q) = (\mathbf{n})!$,

then we let $(\mathbf{n})!_{g(\mathbf{x})}$ be the q-analogue of $n!$ evaluated when $q = g(\mathbf{x})$. That is

$$(\mathbf{n})!_{g(\mathbf{x})} = f\left(g(\mathbf{x})\right) .$$

Of course, we can extend this definition to the q-analogue of n choose k. That is

$$\binom{\mathbf{n}}{\mathbf{k}}_{g(\mathbf{x})} = \frac{(\mathbf{n})!_{g(\mathbf{x})}}{(\mathbf{n}-\mathbf{k})!_{\mathbf{g}(\mathbf{x})}(\mathbf{k})!_{\mathbf{g}(\mathbf{x})}}$$

Stanley shows that that q-analogue of $n!$ can be thought of as a generating function [39].

Proposition 1.2.8 (Stanley 2012)**.** *Let* $\mathrm{inv}(\omega)$ *denote the number of inversions of the permutation* $\omega \in S_n$*. Then*

$$\sum_{\omega \in S_n} q^{\mathrm{inv}(\omega)} = (1+q)(1+q+q^2)\cdots(1+q+q^2+\cdots+q^{n-1}) = (\mathbf{n})! .$$

A similar result works for permutations of a multiset.

Proposition 1.2.9 (Stanley 2012)**.** *Let* $M = \{1^{a_1}, \ldots, m^{a_m}\}$ *be a multiset with cardinality* $n = a_1 + \cdots + a_m$*. Then*

$$\sum_{\pi \in \mathfrak{S}_M} q^{\mathrm{inv}(\pi)} = \binom{\mathbf{n}}{\mathbf{a_1}, \ldots \mathbf{a_m}} .$$

We will also make use of a metric on S_n called Kendall's tau metric. First, as we have mentioned, S_n can be generated by the adjacent transpositions $\epsilon_1, \ldots, \epsilon_{n-1}$. That is, all permutations $\pi \in S_n$ can be written as a product of adjacent transpositions, $\prod_i \epsilon_{j_i}$. We define the metric in the following way: given any two permutations $\pi, \sigma \in S_n$, the distance between π, σ is given by

$$\mathrm{d}(\pi, \sigma) = \mathrm{inv}(\pi\sigma^{-1}) .$$

We will also make use of the fact that the symmetric group S_n is a Coxeter group. We will assume the reader is familiar with Coxeter groups. We will define those concepts we will use in this thesis. One of the things we will use is a Bruhat order on the symmetric group. We will define the terms we will use when thinking of S_n as a Coxeter group.

Definition 1.2.10. Let S_n be the set of all permutations of the set $[n]$ and consider any $\omega \in S_n$. We know that S_n is generated by the adjacent transpositions $s_1, \ldots, s_{n-1}$, where s_i is the permutation $1\ 2 \cdots i+1\ i \cdots n$ which switches i and $i+1$ (alternatively, this is the permutation $(i\ i+1)$ in cycle notation). We can represent ω as a *word* by $\omega = s_{i_1} \cdots s_{i_k}$. We say the word is a *reduced expression* for ω if there does not exist $\ell < k$ such that $\omega = s_{j_1} \cdots s_{j_\ell}$. That is, this is a reduce expression provided there is no

way to write ω as a product of fewer than k adjacent transpositions. In this case, we say the *length of* ω is k. It is true for all $\pi \in S_n$ that the length of π is the equal to $\mathrm{d}(\pi, \mathrm{id})$. Note that reduced expression are not in general unique.

We can now define the weak left Bruhat order on S_n.

Definition 1.2.11. Consider the symmetric group S_n and recall that it is generated by the adjacent transpositions $s_1, \ldots, s_{n-1}$ as defined above. The weak left (Bruhat) order on S_n is the partial order on the group S_n with the relation $\leq$ which. for any two elements $\pi, \sigma \in S_n$, is defined as $\pi \leq \sigma$ if there exists a reduced expression $\sigma = s_{i_1} \cdots s_{i_k}$ such that $s_{i_\ell} s_{i_{\ell+1}} \cdots s_{i_k} = \pi$ where ℓ is the length of π. That is, there is a reduced expression of for σ whose final substring is a reduced word for π.

Finally, we know that S_n with generators $\{s_1, \ldots, s_{n-1}\}$ is a Coxeter group as well as a braid group. We will define braid moves here.

Definition 1.2.12. Consider S_n generated by the adjacent transpositions $\{s_1, \ldots, s_{n-1}\}$. The relations $s_i s_j = s_j s_i$ for with $|i-j| > 1$ and $s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1}$ for $i, j \in [n-1]$ hold in any Coxeter group, including S_n, and a substitution of these forms in a word is called *braid moves* (or sometimes *braid transformations*). Not that because these are equalities, the word itself will not change due to a braid move; while the arrangement and frequency of the letters may change, the overall word will not.

Note that these relationships hold for any Coxeter group, including S_n.

1.3 Background in Statistics

Finally, we provide the background necessary to understand the majority of the statistical methods we use in this thesis. As a form of shorthand, we will denote $\Pr(A)$ to denote the probability of an event A. We will use $\Pr(AB)$ to denote the probability of the intersection of two events A and B, i.e. $\Pr(A \cap B) = \Pr(AB)$.

First, we assume the reader is acquainted with basic probability theory, including the concepts of a sample space, events, and outcomes. We will also assume that the reader has a basic understanding of the axioms of a probability measure as well some of the more basic concepts of the concept of two events being disjoint within a space and partitions of a sample space. For a more complete background on statistics and probability, the reader can consult [43].

We remind the reader that in statistics, random variables are usually divided into two classes: discrete random variables and continuous random variables. We will work with both in this thesis. At the risk of being pedantic, we take the time to define a random variable. Recall that a *random variable*

is a mapping

$$X : \Omega \longrightarrow \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω in the state space Ω. We also must add that a random variable must be measurable in some way.

Example 1.3.1. If we flip a fair coin 5 times, we can let $X(\omega)$ denote the number of heads we observe in the sequence ω. Thus, for $\omega = HHTHT$, $X(\omega) = 3$.

We will stick to the convention that using a capital letter X will denote a random variable, whereas a lower case x denotes a particular sample or value of the random variable X.

Now we remind the reader of a probability function and a probability density function.

Definition 1.3.2. Given X a discrete random variable (i.e. X can take on countably many values $\{x_1, x_2, \ldots\}$). We define the *probability function* (or sometimes a *probability mass function*) for X by $f_X(x) = \Pr(X = x)$.

This is exactly what we would hope it would be: a function whose input is an outcome and whose output is the probability of observing said outcome. When the random variable in question is discrete, this is a perfectly good definition. The analogue for continuous random variables is a probability density function:

Definition 1.3.3. For a continuous random variable X, there exists a function f_X such that:

1. $f_X(x) \geq 0$ for all x.
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.
3. For every $a \leq b$, $\Pr(a < X < b) = \displaystyle\int_a^b f_X(x)\, dx$.

The function f_X is called the *probability density function*.

As shorthand, we will generally use $f(x)$ to denote the probability density function, and it should be clear from context which probability density function we are referring to. We may sometimes write $\int f(x)\, dx$ to denote $\int_{-\infty}^{\infty} f(x)\, dx$. It is important to remember that for continuous random variables, we rely on integrals to obtain probabilities.

Recall the definition of independence of events in a sample space:

Definition 1.3.4. Given two events A, B on a sample space Ω with any probability distribution. Then the two events A and B are said to be *independent* if

$$\Pr(AB) = \Pr(A)\Pr(B)$$

and we write $A \perp\!\!\!\perp B$. A set of events $\{A_i \mid i \in I\}$ is independent if

$$\Pr\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \Pr(A_i)$$

for every finite subset J of I.

For the most part, we will assume two (or more) events are independent. Returning to our coin toss example, we usually would assume the two coin tosses are independent, which would reflect the fact that the coin has no memory of the first toss. We can also derive independence by verifying the definition and confirming that in fact $\Pr(AB) = \Pr(A)\Pr(B)$. For example, if we roll a fair 6-sided die twice, and let the event $A = \{2,4,6\}$ represent the event we roll an even number and the event $B = \{1,2,3,4\}$ represent that you roll a number less than 5. The we know that $AB = \{2,4\}$ and can compute $\Pr(AB) = 2/6 = 1/3$, as we know all outcomes are equally likely. We can also compute $\Pr(A) = 3/6 = 1/2$ and the $\Pr(B) = 4/6 = 2/3$ and verify that $\Pr(A)\Pr(B) = 1/2 \times 2/3 = 1/3$ and so $\Pr(AB) = \Pr(A)\Pr(B)$ and conclude $A \perp\!\!\!\perp B$. In this thesis, we will assume events are independent rather than derive they are independent. Note that disjoint events with positive probability are not independent (as $\Pr(AB) = \Pr(\emptyset) = 0$ and $\Pr(A), \Pr(B) > 0$).

Independence of events can also be conditional on other events. Given an event B with $\Pr(B) > 0$, we can define the conditional probability of an event A given B.

Definition 1.3.5. If $\Pr(B) > 0$, then the *conditional probability* of A given B is defined as

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)}$$

$\Pr(A|B)$ is the fraction of times that A occurs among those times in which B occurs. In other words, knowing that B has occurred, it is the fraction of times that A occurs, and thus we are restricting our sample space to the space in which B occurs. While it is true that $\Pr(\cdot|B)$ satisfies all three axioms of a probability (provided $\Pr(B) > 0$), it is not true in general that $\Pr(A|B \bigcup C) = \Pr(A|B) + \Pr(A|C)$. In general, the rules of probability apply to things on the left side of the condition bars, but not necessarily on the right. Another example is that, in general, $\Pr(A|B) \neq \Pr(B|A)$. As an example, the probability that you ate something cold given that you have a brain freeze is 1, whereas the probability

that you have a brain freeze given you ate something cold is not 1.

Before we examine the statistical techniques we will use in the thesis, we define a statistical model:

Definition 1.3.6. Given a sample space Ω, a *statistical model* is a set $\mathscr{P}$ of probability distributions on the sample space Ω.

In practice, a statistical model incorporates the set of assumptions germane to the generation the observed data from a larger population. A model represents the data-generating process, usually in extremely idealized forms.

Most statistical models that are used in practice are parametric models. A *parametric model* is a set of probability distributions $\mathscr{P}$ that can be parameterized by a finite number of parameters. We introduce one of the most famous parametric models in the following example:

Example 1.3.7. If we assume that the gathered data comes from a univariate Gaussian (or Normal) distribution, then the model is

$$\mathscr{P} = \left\{ f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} ; \mu \in \mathbb{R},\ \sigma^2 > 0 \right\} .$$

Here, we refer to μ as the mean and σ as the standard deviation. When a random variable follows this distribution, we refer to it as a Gaussian random variable (or Normal random variable). We denote that a random variable X follows a Gaussian distribution with mean μ and standard deviation σ by $X \sim \mathscr{N}(\mu, \sigma^2)$. The graph for the probability density function for a Gaussian random variable with $\mu = 0$ and $\sigma = 1$ is shown in Figure 1.6.

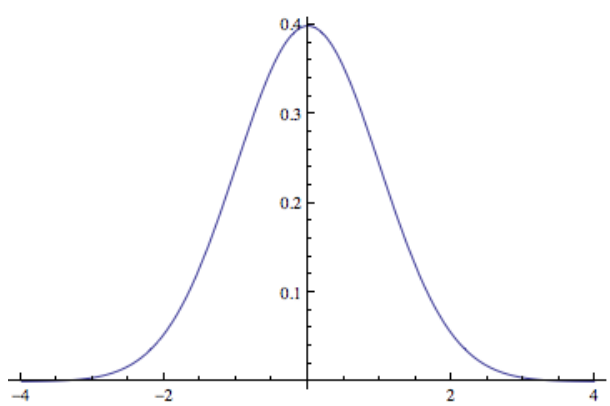

Figure 1.6 The graph of the probability density function for a Gaussian random variable X with $\mu = 0$, $\sigma = 1$

While it is true that there is a more general multivariate normal distribution, we will not use it within the scope of this thesis. In general, we denote that a random variable X is distributed to some parametric statistical model P with parameters θ with the notation $X \sim P(\theta)$.

Statisticians use statistical inferences to analyze data. The two most dominant types of statistical inference: *frequentist inference* and *Bayesian inference*. In this thesis, we are interested in recovering (or estimating) the parameters of a statistical model (as opposed to a probability density function or cumulative density function, for example). We will use both Bayesian and frequentist inference in Chapter 4 to do this. We will introduce the main idea of frequentist inference first.

Frequentist inference draws conclusions from sample data based on the frequency of that data. Statistical hypothesis testing and computation of confidence intervals are both frequentist techniques. We will make extensive use of maximum likelihood estimation throughout the thesis, which is another common frequentist technique.

Maximum likelihood estimation is a technique for estimating the parameters of a model when we do not have direct access to them. Since it is often the parameters of a statistical model that we are interested in sampling, this technique is rather common. During the general discussion of MLE techniques, we will refer to the set of parameters of a model as θ which comes from a parameter space Θ. In later sections, we will refer to parameters by their names rather than the general θ.

Before we can truly talk about MLE, we must first introduce the likelihood function. This requires us to have some knowledge of random variables being independent and identically distributed (iid). This concept is basically summed up in its name: given random variables $X_1, \ldots, X_n$, we say $X_1, \ldots, X_n$ are *independent and identically distributed* (or iid) if $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$ and each of the X_i follows a single distribution with parameter set θ_i. Furthermore, for each i, individual sample of the random variable X_i are independent of one another.

Example 1.3.8. Consider a fair or unfair six-sided die. Rolls of that die are iid, regardless of whether it is fair, as each roll is independent of the others. That is, even if you roll a 6 five times in a row, rolling a 6 on the next roll has the same probability as it did on all the previous rolls. The same would apply for rolling multiple fair or unfair dice, provided there is a way to denote which die is which.

We remind the reader of the expected value of a random variable. The mathematical definition of expected value of a random variable will be different depending on whether it is a discrete random variable or a continuous random variable, but in both cases, the general idea is that the expected value is, intuitively, the long-run average value the variable takes on with increasing repetitions of an experiment. We will only be using the expected value of continuous random variables in this thesis so we will only define the expected value for a continuous random variable. For a more general definition of the expected random variable, see [32, 43].

For a continuous random variable, we have the following:

Definition 1.3.9. Given a random variable X with probability space Ω and probability density function $f(x)$. The *expected value* (or *mean* or *first moment*) of X is given by

$$\mathbb{E}[X] = \int_{\Omega} x f(x)\, dx$$

assuming the integral is well defined.

Note that the expected value of a function requires integrating over the entire sample space. In cases where the sample space is more complicated, computing the expected value becomes more difficult. Later in this thesis, we will examine different methods to estimate the expected value of a random variable when computing the above integral is not straight-forward.

One fact worth noting at this point is if we have $g(X)$ a measurable function of X, we can compute

$$\mathbb{E}[g(x)] = \int_{\Omega} g(x) f(x)\, dx \quad .$$

This fact is sometimes referred to as the Rule of the Lazy Statistician [43].

We are now ready to define the likelihood function:

Definition 1.3.10. Let $X_1, \ldots, X_n$ be iid with probability density function $f_i(A_i|\theta_i)$. The *likelihood function* is defined by

$$\mathscr{L}_n(\theta|X) = \prod_{i=1}^{n} f_i(X_i|\theta_i)$$

The *log-likelihood function* is given by $\ell_n(\theta) = \log \mathscr{L}_n(\theta|X)$.

In this thesis, we will only use the log-likelihood function. We note that the log-likelihood function is just the log of the joint density function, but we are treating it as a function of the parameter θ, as indicated by the notation. Sometimes, to make this even more explicit, we denote the log-likelihood function as $\ell(\theta|X)$ where $X = (X_1, \ldots, X_n)$. It is worth noting that the log-likelihood function is not a density function, and therefore will not integrate to 1 (with respect to θ).

The definition of the MLE follows naturally

Definition 1.3.11. Let $X_1, \ldots, X_n$ be iid with probability density function $f_i(x_i; \theta_i)$. The *maximum likelihood estimator* MLE, denoted $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$, is the value of $\theta \in \Theta$ that maximizes $\mathscr{L}_n(\theta)$.

It should be clear that any value of θ which maximizes $\mathscr{L}_n(\theta)$ will also be the value of θ which maximizes $\ell_n(\theta)$. Note that by this definition, the MLE is not necessarily unique, although in practice

it usually is. A few notes about the MLE are that it is consistent–which means that it will converge to the true value of the parameter as sample sizes get larger and larger–and it is asymptotically normal. These results are contingent upon the model following certain well defined criteria; in this thesis, we will only use MLE in cases where the model meets the criteria necessary to guarantee that it is both consistent and asymptotically normal. We mention one theorem of note, which can be found in [43]:

Theorem 1.3.12. *Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be the MLE of θ. Then $\hat{\tau} = g(\hat{\theta}_n)$ is the MLE of τ.*

Now we look at a simple example.

Example 1.3.13. Suppose we have a Gaussian random variable X whose mean and standard deviation are unknown. We wish to find the values of μ, σ which maximize the log-likelihood function. Say we have n observations. Let $f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$. The log-likelihood function is

$$\begin{aligned}\ell(\mu,\sigma \mid X) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2}\right) \\ &= -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\end{aligned}$$

If we differentiate this equation with respect to μ we get

$$\frac{d}{d\mu}\left[-n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right] = \sum_{i=1}^{n} \frac{x^{(i)}-\mu}{\sigma^2}$$

When we set this equal to 0 we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{x^{(i)}-\mu}{\sigma^2} &= 0 \\ \left(\sum_{i=1}^{n} x^{(i)}\right) - n\mu &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} x^{(i)} &= \mu\end{aligned}$$

which means the value of μ which maximizes the log-likelihood function is the average value of the observed $X^{(i)}$.

Similarly, if we were to differentiate with respect to σ:

$$\frac{d}{d\sigma}\left[-n\log(\sigma\sqrt{2\pi})+\sum_{i=1}^{n}-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right]=0$$
$$\sum_{i=1}^{n}\frac{(x^{(i)}-\mu)^2}{\sigma^3}=\frac{n}{\sigma}$$
$$\frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\mu)^2=\sigma^2$$

and again we see that by definition, value of σ which maximizes the log-likelihood function is exactly the standard deviation of the sampled $X^{(i)}$. Thus, regardless of the number of samples taken, the maximum likelihood estimates for a univariate Gaussian random variable will be exactly the mean and the variance of the samples.

The Expectation-Maximization (EM) algorithm is an iterative method which inputs observed data to obtain the maximum *a posteriori* estimate for the parameters of a statistical model which has hidden (or latent or unobserved) variables. The algorithm has two steps, an Expectation step (E-step) and a Maximization step (M-step), hence its name. The EM algorithm has been applied to many different data sets and many different scenarios. We will focus on a rather straightforward application of the EM algorithm; we will assume the data is iid.

Before we give a formal outline for the EM algorithm, we will lay out the steps intuitively. First, we have observed data $Y^{(1)},\ldots,Y^{(N)}$ which we collect, which will be associated with hidden variables $X^{(1)},\ldots,X^{(N)}$. We will assume some initial estimate θ^0 for the true parameters θ. In the E step, we use the observed data to create a function for the expectation for the log-likelihood function $\ell(\theta|Y^{(\cdot)},X^{(\cdot)})$ using the current estimate for the parameters. The M step then computes the value of the parameters that will maximize the function created in the E step. Then we repeat the E step using the new parameter estimation computed in the previous M step. The algorithm will produce a sequence of estimates for the parameters θ^i which converge to locally maximum likelihood parameters.

Before we formally give the EM algorithm, we will introduce some shorthand notation. When working with models where we have observed values of a random variable X, we will let $X^{(i)}$ denote the ith observation. If we want to refer to all observations of the variable, we use $X^{(\cdot)}$. If our $X=(X_1,\ldots,X_n)$, then $X^{(i)}=(X_1^{(i)},\ldots,X_n^{(i)})$ and we can refer to all observations of the ith entry of X as $X_i^{(\cdot)}$. We will sometimes refer to all observations of $X=(X_1,\ldots,X_n)$ as simply $\mathbf{X}$.

Consider the following scenario: we are given a statistical model which generates a set of observed

data $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(N)})$ as well as unobserved data $\mathbf{X} = (X^{(1)}, \ldots, X^{(N)})$ with a vector of unknown parameters θ and a log-likelihood function $\ell(\theta \mid \mathbf{Y}, \mathbf{X}) = p(\mathbf{Y}, \mathbf{X} \mid \theta)$. The maximum likelihood estimate of the parameters is determined by the marginal likelihood of the observed data

$$\ell(\theta \mid \mathbf{Y}, \mathbf{X}) = p(\mathbf{Y} \mid \theta) = \sum_{\mathbf{X}} p(\mathbf{Y}, \mathbf{X} \mid \theta).$$

But because the hidden variable $\mathbf{X}$ cannot be observed, this quantity is almost always insoluble. This is where the EM algorithm can be used to recover a maximum likelihood estimate for the parameter.

Formally, the EM algorithm can be written as follows:

Algorithm 1.3.14. Given a statistical model which generates a set of observed data $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(N)})$ as well as unobserved data $\mathbf{X} = (X^{(1)}, \ldots, X^{(N)})$ with a vector of unknown parameters θ and a log-likelihood function $\ell(\theta \mid \mathbf{Y}, \mathbf{X})$ we initialize the algortihm with an initial value for the parameter vector θ^0. Then for $i = 1, 2, \ldots$, repeat steps one and two below

1. (The E Step) Calculate the function:

$$K(\theta | \theta^i) = \mathbb{E}_{\mathbf{X}|\mathbf{Y}, \theta^i}[\ell(\theta \mid \mathbf{Y}, \mathbf{X})]$$

   where the θ^i and the observed $\mathbf{Y}$ are fixed (θ is a variable).

2. Find the value of θ^{i+1} which maximizes $K(\theta | \theta^i)$. i.e.

$$\theta^{i+1} = \arg\max_{\theta} \{K(\theta | \theta^i)\}$$

The EM algorithm obtains an maximum likelihood estimate for the parameters of a model without computing $\ell(\theta \mid \mathbf{Y}, \mathbf{X})$, but computing the expected value $\mathbb{E}_{\mathbf{X}|\mathbf{Y}, \theta^i}[\ell(\theta \mid \mathbf{Y}, \mathbf{X})]$ can be equally difficult. We will see in Chapter 4 that we will need a way to estimate this expected value in order to make use of the EM algorithm. Still, the EM algorithm is a very powerful method for estimating parameters, especially in cases where the statistical model in question has hidden variables.

In the thesis which introduces the EM algorithm, Dempster, Laird, and Rubin modified the EM method to compute the maximum *a posteriori* estimates for Bayesian inference [11], making it a versatile tool in parameter estimation in both frequentist and Bayesian inference. For the model introduced in Chapter 4, we will be using the original EM algorithm along with methods for estimating the expected value in step 1 of Algorithm 1.3.14 to estimate the parameters of our model.

We also make use of Bayesian methods in Chapter 4. Bayesian inference is a kind of statistical inference which uses Bayes' Theorem to update the probability for a hypothesis as evidence (or data)

is collected. Because we are constantly updating the hypothesis, all Bayesian inference starts with a prior distribution on the value in question and seeks sample from the posterior distribution of that same value. This value is usually the true value of a parameter. Thus, we begin with an initial idea of that parameter might be by sampling from a prior distribution on that parameter. Then, we observe data, and finally, using that data and our parameter, we seek to sample from the posterior distribution on that parameter. We think of sampling from this posterior distribution as sampling from a distribution on the parameter of interest in light of the observed data.

We remind the reader of Bayes' Theorem:

Theorem 1.3.15 (Bayes' Theorem)**.** *Let A, B be events in the sample space where* $\Pr(B) > 0$. *Then*

$$\Pr(A \mid B) = \frac{\Pr(A)\ \Pr(B \mid A)}{\Pr(B)}$$

This formulation of Bayes' Theorem is for events in a sample space. If we have, instead, a sample space generated from random variables, we need to modify Bayes' theorem in order for it to be useful. The modification for random variables follows from the original statement of Bayes' theorem. We note that there are multiple formulations for Bayes' theorem as regarded to random variables based on whether the random variables are continuous or discrete. We present the formulation which will be used most frequently in this thesis.

Theorem 1.3.16 (Bayes' Theorem for Random Variables)**.** *Given continuous random variables X, Y which generate a sample space. Let f_X, f_Y denote the probability density functions of X, Y respectively. Then we have*

$$f_X(x \mid Y = y) = \frac{f_Y(y \mid X = x)\, f_X(x)}{f_Y(y)}$$

We know from the law of total probability that

$$f_Y(y) = \int_{-\infty}^{\infty} f_Y(y \mid X = \xi) f_X(\xi)\, d\xi$$

Bayesian techniques frequently employ Markov chain Monte Carlo algorithms. The reason for this is straight forward: Bayesian inference starts with a prior distribution on the parameters and seeks to find the posterior distribution on those parameters, given a sample of some sort. Sampling directly from this posterior distribution is often difficult if not impossible. Markov chain Monte Carlo (MCMC) methods are a class of algorithms used for sampling from a probability distributions by construction a Markov chain with the desired distribution as its equilibrium state. The state of the chain after some number of steps is used to sample the desired distribution; the quality of the sample obtained

improves as the number of steps increases. This is why MCMC algorithms are frequently employed in Bayesian techniques. Sampling from the posterior distribution is done by using a Markov chain, and taking enough of these samples allows for an accurate sampling from the posterior distribution of the parameters.

In Chapter 4, we will use a Bayesian technique known as a Gibbs sampler. Gibbs sampling is an MCMC algorithm used to obtain a sequence of observations which are approximated from the joint probability distribution of two or more random variables when direct sampling is difficult (or impossible). The sequence can be used to approximate the joint distribution, the marginal distribution of a single variable or a subset of the variables–including unknown parameters or hidden variables–or to compute integrals such as expected values. It is frequently used when the values of some of the variables are known, and therefore do not need to be sampled. It is a randomized algorithm, meaning it can be an alternative to deterministic algorithms, such as the EM algorithm. In its most basic form, Gibbs sampling is a special case of the Metropolis-Hastings algorithm.

Gibbs sampling is used in situations where the joint distribution of the random variables is not explicitly known or is difficult to sample directly, but the conditional distribution of each variable is known and is simple (or at the very least, easier) to sample from. The Gibbs sampling algorithm generates a sample from the distribution of each variable in turn, conditional on the current value of all the other variables. It has been shown that this sequence of samples is a Markov chain and the stationary distribution of this Markov chain is the joint distribution we are interested in.

The key idea behind Gibbs sampling is that if we are given a multivariate distribution, it is easier to sample from the conditional distribution than to marginalize by integrating over a joint distribution. The goal of a Gibbs sampler is to obtain a large number of samples of a random variable coming from a given joint distribution. The joint distribution, however, is either not explicitly known or not difficult to sample directly, so what a Gibbs sampler actually does is generate samples that approximate a joint distribution of all variables. Consider the most basic incarnation of the Gibbs sampler.

Algorithm 1.3.17 (Gibbs Sampler)**.** Given a random variable $\mathbf{X} = (x_1, \ldots, x_n)$. We wish to obtain k samples of $\mathbf{X}$ from a joint distribution $p(x_1, \ldots, x_n)$; denote the ith sample as $\mathbf{X}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$. Select some value of $X^{(0)}$ as an initial value of $\mathbf{X}$. Then

1. To obtain the $i+1^{\text{st}}$ sample, sample each component variable x_j^{i+1} (for $j = 1, \ldots, n$) from the distribution of that variable conditional on all other variables using the most recent value of each of the other variables. In other words, if we are updating the jth component, we update it according to the distribution $p(x_j \mid x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$

2. Repeat k times

When this kind of sampling takes place, we know that the samples approximate the joint distribution on the variables, the expected value of any variable can be approximated by taking the average over all samples, and the marginal distribution over a subset of variables can be approximated by considering the samples for that subset of variables and ignoring variables not in the subset. The initial value can be determined randomly or by another algorithm, such as the EM algorithm.

When using Gibbs sampling, it is fairly common to ignore a number of samples taken from the beginning of the algorithm. This is commonly referred to as a *burn in period*. It is also common to only "observe" every n samples after the burn in period. This prevents consecutive samples from being "trapped" in a particular part of the sample space, as well as ensures that each sample is sufficiently random. As an example, when running the Gibbs sampler in Chapter 4, we use a burn in value of 1000 (we discard the first 1000 full iterations of the Gibbs sampler) with 200 iterations between each sample afterward. In the context we use the Gibbs sampler in this paper, we know it converges to a true sampling of the posterior distribution [35].

Example 1.3.18. Suppose we have two Gaussian random variables, X, Y and our model dictates that both are distributed with $\mu = 0, \sigma = 1$ with the added stipulation that $X \leq Y$. We can use Gibbs sampling to sample points from this space. Start with initial values $X = Y = 0$. We first sample X, noting that $X \sim \mathcal{N}(0,1)$. We know, however, that $X \leq Y$ so we must preserve that relationship. There is any number of ways we could do this, but for now let's use a truncated normal distribution (a distribution which behaves like a normal distribution but does not allow us to sample any $X > 00$). With a random sample we get $X^{(1)} = -1.2$. Then we need to sample Y while preserving that $X \leq Y$. Mathematically, we look for $\Pr(y \mid X = -1.2)$. Again we can use a truncated normal to do this, and might see that we get $Y^{(1)} = -.45$. When we look to sample $X^{(2)}$, we use the value of $Y^{(1)}$ to ensure that we preserve the relationship $X \leq Y$. Therefore, when we find $p(x \mid Y = -.45)$ we might see that $X^{(2)} = -.82$. We continue in this manner until we get the desired number of samples.

As mentioned before, the Gibbs sampler can also be used to estimate expected values and parameter values. We will see more on this in Chapter 4.

1.4 Statistical Models for Ranked and Partially Ranked Data

A statistical model for ranked data is a model is a family of probability distributions on the symmetric group S_n. Such models are used in many different disciplines. Before we examine previously proposed statistical models for ranked data, we will use the convention of letting n be the number of items being ranked in our statistical model. It should be clear from context whether each model assigns a discrete probability function to S_n or whether the observed data is a ranking on n items. The first

model we consider in this thesis is based on the Mallows model. The Mallows model was originally proposed by Mallows in [29] in 1957. The paper contains many variations of statistical models for ranked data which come from different assumptions placed on the general model. The Mallows model is a location-scale model which assigns a probability to each permutation in S_n. The probability the model assigns each permutation is based on two parameters: a center permutation $\kappa \in S_n$ (which functions much like the mean of a normal distribution) and a parameter $c \in \mathbb{R}_+$ encoding spread (which behaves very similarly to the standard deviation of a normal distribution). If we let $p_\kappa(\pi)$ represent the probability of observing a permutation $\pi \in S_n$ where κ is the center permutation, we have

$$p_\kappa(\pi) = e^{-c\mathrm{d}(\pi,\kappa) - \log(\psi(c))}$$

where $c \in \mathbb{R}_+$, $e^{-\log\psi(c)}$ behaves as a normalizing constant, and $\mathrm{d}(\pi,\sigma)$ is Kendall's tau metric on S_n, defined by

$$\mathrm{d}(\pi,\sigma) = \mathrm{inv}(\pi\sigma^{-1}) \ .$$

We will talk more about this metric as well as the Mallows model itself in Chapter 2.

A ranking can arise through a series of sequential comparisons where a single item is preferred to all remaining items and, after it is selected, is removed from all future comparisons. This concept lies at the core of the Plackett-Luce model. The Plackett-Luce model (P-L model) is a statistical model for ranked data which has been adapted for partially ranked data as well. The model stems from the idea that For fully ranked data, we have n items to be ranked by k judges and assume no ties, we have a set of k observed rankings

$$\{y^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)}) \mid i = 1, \ldots, k\}$$

where $y_j^{(i)}$ is the position (or rank) assigned to item j by judge i. In other words, judge i ranks item j in position $y_j^{(i)}$. This ranking is naturally associated with a permutation $\pi^{(i)} = S_n$ where $\pi^{(i)} = y_1^{(i)} \cdots y_n^{(i)}$. The Plackett-Luce model (P-L model) is a distribution over all rankings which can be described entirely by a permutation σ. Thus, the probability assigned to σ is not the probability of the permutation associated directly with a ranking $y = (y_1, \ldots, y_n)$, but rather the probability assigned to the inverse of the permutation associated with y. The model has parameter vector $\theta = (\theta_1, \ldots, \theta_n)$ with $\theta_i \geq 0$ where θ_i is associated with item i. This model assigns probability

$$\Pr(\sigma \mid \theta) = \prod_{i=1,\ldots,n} \frac{\theta_{\sigma_i}}{\sum_{j=i}^{n} \theta_{\sigma_j}} \ .$$

Note that this is not the only formulation of this model.

The Plackett-Luce model can be used as a model for partially ranked data as well. In this case,

the model has a poset $\mathscr{P}$ associated with it. If $i < j$ is a relation of the poset $\mathscr{P}$, then item i is always ranked before item j in the corresponding model. Let Q be the maximal chains of the poset $\mathscr{P}$. The state space of this model is the set $\mathscr{L}(\mathscr{P})$ are the permutations $\pi \in S_n$ that respect the relations of $\mathscr{P}$, i.e. they are the permutations which are linear extensions of $\mathscr{P}$. Note that this model's state space is not all of S_n. The probability function can be obtained from the Plackett-Luce model for fully ranked data by normalizing over a subset of S_n (for more on this, see [41]. Then we see that for any $\pi \in \mathscr{L}(\mathscr{P})$, the probability of observing π is given by

$$\Pr(\pi \mid \theta) \;=\; \prod_{i=1}^{n-1} \frac{1}{\sum_{j=1}^{i} \theta_{\pi(j)}} \qquad \text{for} \quad \pi \in \mathscr{L}(\mathscr{P}) \;.$$

While we will not make use of the P-L model in this paper, several well known statistical tools can be used with the P-L model. The authors of [1] demonstrate how to use regression in a P-L model. Microsoft researchers Guiver and Snelson give an efficient method for inferring the parameters of P-L model in [21]. Mollica and Tardella develop methods for efficiently running the EM algorithm and a Gibbs sampler on a mixture of Plackett-Luce models [30] and use a mixture of P-L models to model epitope profiling [31]. The authors of [7] propose a Bayesian nonparametric extension of the P-L choice model capable of handling an infinite number of choice items.

Thurstonian models are a third class of statistical models for ranked data. Proposed in 1927, the Thurstonian model assumes that every item being ranked has an inherent, unobservable true value [42]. The model observes rankings on n items by a judge (or judges). The Thurstonian model assumes the each item is given a value X_i where $X_i \sim \mathscr{N}(\mu_i, \sigma_i^2)$. The mean μ_i is the true value of that item and σ_i is an unobservable parameter associated with the item. The ranking a judge assigns to the items depends entirely on the value of X_i's–the ranking assigned to an item i will be $\#\{j \mid X_j < X_i \text{ where j} \neq \text{i}\}$. Within the framework of this model, it is possible for the same judge to rank the same items in different ways. The notable aspects of this model assumes the value assigned during each ranking by each judge (where there are potentially multiple judges and multiple rankings from each judge) are continuous real number values which cannot be observed and these values are distributed according to a normal distribution.

Unlike the previous two models, Thurstonian models do not have a closed form for the probability of observing a ranking $\pi \in S_n$. This is in part due to the unobservable values of the X_i, the hidden variable which represents the value the judge assigns item i. As a result, estimating the most likely ranking order as well as the means of these normal distributions is not straight-forward. This makes working with this model somewhat more difficult than working with a Mallows or P-L model. This does not, however, mean it is avoided in practice. Thurstonian models are becoming more widely

used as methods for parameter estimation are developed and are made more efficient. In the sensory field, authors Bi and Kuesten claim that Torgerson's method of triads has been avoided due to the fact that the Thurstonian model that is a part of the method and "there are no published tables or available computer software for applications of the method[4]." In their paper, the propose a Thurstonian model for a special case of Torgeson's method of triads. Ennis and Rousseau develop a Thurstonian model for degree of differences methodology, a methodology where subjects are given pairs of samples and must indicate how different the are on a t-point scale [16]. The model can be used in many discrimination, rating, and ranking methodologies. The authors of [9] propose and alternative to the two-Alternative Forced Choice (2-AFC) model in which participants are presented a pair of items and asked which is preferred where the response of "no preference" is allowed. They then detail ways to extract estimates and standard error of the parameters in this two-alternative choice model. Gianola and Simianer introduce a fully Bayesian method for quantitative genetic analysis of data consisting of ranks which are scored at a series of events or experiments [20]. The rank observed is assumed to reflect the order of values of some unobserved variable which is distributed normally, and is therefore another application of the Thurstonian model. We will talk about more applications of Thurstonian models as well as methods for estimating values of interest in a Thurstonian model in Chapter 4.

These are some of the more well known statistical models for ranked data. As we mentioned before, statistical models for ranked and partially ranked data are used in many different disciplines. In the cognitive sciences, we mentioned Steyvers et al. proposed a model for reconstructing the true ranking of a series of events, such order of historical events or listing cities in the US from easternmost to westernmost, based on the responses (or guesses) of participants [40]. The authors of [27] models for ranked data to estimate the degree to which the responder is an expert, assigning a level of how much of an expert and therefore how likely their response is accurate a participant is based on the way they rank a number of different sets of events (again, including order of US presidents, rivers or the world from longest to shortest, etc.). We have seen a Thurstonian model for ranked data used in genetics to model the behavior of different biological and genetic processes, such as genotypes [20]. Beerenwinkel and Sullivant propose a model for partially ranked data which describes mutation accumulation in an organism [3]. The authors of [18] introduce a model for longitudinal partially ranked data and apply it to survey data recording the top two political concerns of citizens of the United Kingdom. This model takes more into account than simple paired-comparison tests. Models for ranked and partially ranked data have been proposed, adapted and used in signal detection [24, 25], food science [9, 16], sensory studies [4, 24, 25] and many other fields. The pervasiveness of these models is due to the myriad of conceivable instances where items can be ranked or partially ordered. Therefore studying statistical models for ranked or partially ranked data can have an impact on any number of academic disciplines.

We seek to study these models from an algebraic and combinatorial point of view.

1.5 Outline of Thesis

Statistical models for ranked data are studied and analyzed for many different reasons; they have been studied to compare their performance to other models, to test or improve their computational efficiency, to measure their tolerance to noise, and for many other reasons. In this thesis, we study these models from an algebraic perspective and look at the combinatorics associated with a vanishing ideal which describes the model. Sturmfels and Welker examined the algebraic properties of four different models for ranked data [41]. We will be examining the algebraic and combinatorial properties of statistical models for ranked data. In Chapter 2, we will examine the algebraic properties of the Mallows model and introduce a mixture of Mallows models. After describing the mixture model, we use combinatorial tools to simplify it and then develop the algebraic tools necessary to describe its vanishing ideal. In Chapter 3, we define a generating function which will count the number of permutations in S_n which are an equal distance from two fixed permutations π, σ. This function is necessary for any practical application of the Mallows mixture model developed in Chapter 2. Using the generating function, we give a closed form for the number of permutations in S_n which are an equal distance from two fixed permutations π, σ. In Chapter 4, we propose a Thurstonian model for partially ranked data. We then show how to use the EM algorithm and a Gibbs sampler to estimate the parameters of the model. Finally, we apply the model and the methods described to two different data sets and compare our results to other models in the literature.

CHAPTER

2

THE MALLOWS MIXTURE MODEL

The Mallows model is a statistical model for ranked data which gives a closed form for computing the probability of observing a particular permutation in S_n. It is a location scale model, much like the normal distribution, meaning the probability assigned to each permutation will decrease the further away it is from the “center" permutation. The Mallows model has been used in a number of different disciplines. Lebanon and Mao set up the framework for using the Mallows model specifically on permutations which are partition-preserving [26]. In this chapter, we build on this framework and examine a mixture of Mallows model. The interest in doing so comes in part from the work of Lebanon and Mao.

In this chapter, we introduce a mixture model based on a classic statistical model for ranked data, the Mallows model. The original model, proposed by Mallows[29], makes use of paired comparison techniques as well as Kendall’s tau metric on the symmetric group. The model is designed to be used for ranked data. That is, the observations are entire rankings on a set of n items. Because of this, it is natural to think of the observations as permutations in S_n. In Section 2.1 we introduce and examine the vanishing ideal of the original Mallows model. In Section 2.2 we will introduce the Mallows mixture model which we analyze and characterize in later sections. In Section 2.3 we introduce theorems to completely describe all (i, j) pairs in $\{0, \ldots, \binom{n}{2}\}^2$ for which there exists a permutation $\pi \in S_n$ such that $\mathrm{d}(\pi, \kappa_1) = i$ and $\mathrm{d}(\pi, \kappa_2) = j$. We then develop theorems which characterize the joins of ideals by their

degree in Section 2.4. Using the theorems for Section 2.3 and Section 2.4, we look at the vanishing ideal of the map of the Mallows mixture model in Section 2.5. In Section 2.6 we look at the number of generators of various degrees.

2.1 The Mallows Model

The original Mallows model was proposed by Mallows in 1957[29]. It is a location-scale model for ranked data for which uses paired comparisons to assign a probability to every possible ranking. Because the observed data points in this model are rankings on n items, it is natural to think of these observations as permutations in S_n. Thus, if we were to rank 4 items $\{1,2,3,4\}$, the permutation 4132 is equivalent to the ranking where item 4 is ranked first, item 1 is ranked second, etc. The model has a center, much like a mean. That is, the closer a permutation is to the center, the more likely it will be observed. A well defined concept of "closeness" requires a metric on S_n. We choose Kendall's tau distance as our metric. Recall the symmetric group is generated by the $n-1$ adjacent transposition of S_n. Under Kendall's tau metric, the distance between any two permutations is the minimum number of adjacent transpositions necessary to compose with one of the permutations to transform it into the second. That is, $\mathrm{d}(\pi,\sigma) = \mathrm{inv}(\pi\sigma^{-1})$. For instance, the distance between 3142 and 1243 is 2, as seen in Figure 2.1. The Cayley graph is a visualization of distance between elements of S_n.

This is just one way to describe Kendall's tau distance, but it is the one we will use for the remainder of the paper. It should be noted that this is a right invariant metric. That is,

$$\mathrm{d}(\pi,\sigma) = \mathrm{d}(\pi\tau,\sigma\tau) \qquad \forall \pi,\sigma,\tau \in S_n$$

Furthermore, for all $\pi,\sigma \in S_n$, we know that $0 \le \mathrm{d}(\pi,\sigma) \le \binom{n}{2}$

Under the Mallows model which is centered about the permutation κ, the probability of observing any probability π is exactly

$$p_\kappa(\pi) = e^{-c\mathrm{d}(\pi,\kappa)-\log(\psi(c))}$$

where $\pi,\kappa \in S_n$, $c \in \mathbb{R}_+$, and $\psi(c)$ is the normalizing constant with $\psi(c) = \sum_{\pi\in S_n} e^{-c\mathrm{d}(\pi,\kappa)}$. It will be useful to us to clean up this notation with a few simple substitutions. First, we note that since Kendall's tau distance is right invariant, we have that $\mathrm{d}(\pi,\kappa) = \mathrm{d}(\pi\kappa^{-1},\kappa\kappa^{-1}) = \mathrm{d}(\pi\kappa^{-1},\mathrm{id})$. We know that $\mathrm{d}(\pi\kappa^{-1},\mathrm{id}) = \mathrm{inv}(\pi\kappa^{-1})$ where $\mathrm{inv}(\pi\kappa^{-1})$ is the number of inversions of the permutation $\pi\kappa^{-1}$. Recall that

$$\mathrm{inv}(\pi\kappa^{-1}) = \#\left\{(i,j)\in[n]\times[n] \mid i<j \text{ and } \pi\kappa^{-1}(i) > \pi\kappa^{-1}(j)\right\}.$$

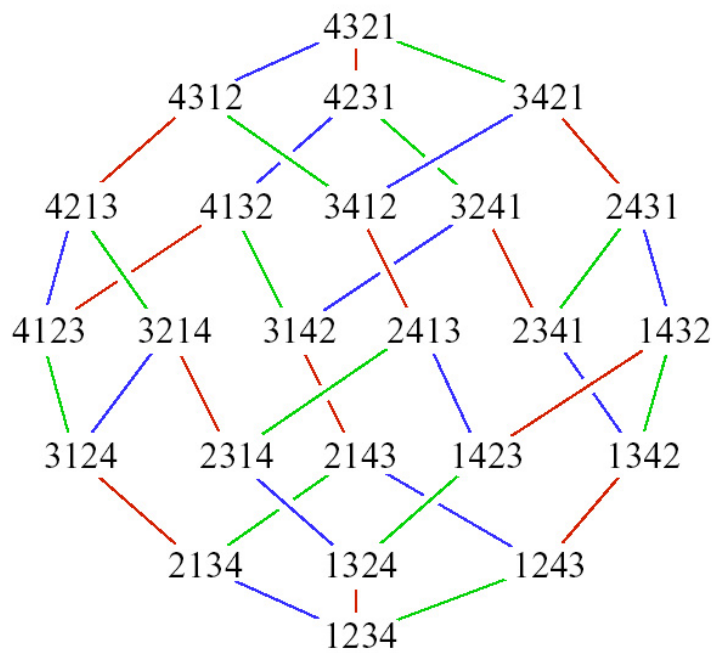

Figure 2.1 The Cayley graph of S_4

We can then rewrite

$$p_\kappa(\pi) = e^{-c\mathrm{d}(\pi,\kappa)-\log(\psi(c))} = e^{-c\mathrm{inv}(\pi\kappa^{-1})-\log(\psi(c))}$$

To further simplify notation, we can set $e^{-c} = q$. Then we have

$$p_\kappa(\pi) = e^{-c\mathrm{inv}(\pi\kappa^{-1})-\log(\psi(c))} = \left(e^{-c}\right)^{\mathrm{d}(\pi,\kappa)} e^{-\log(\psi(c))} = \frac{1}{\psi(c)} q^{\mathrm{inv}(\pi\kappa^{-1})}$$

Finally, we let $t = \dfrac{1}{\psi(c)}$ and write that

$$p_\kappa(\pi) = t\, q^{\mathrm{inv}(\pi\kappa^{-1})}$$

We will use this simplified notation throughout the paper.

One way to think of this model is a sort of “normal" distribution on the discrete set S_n. It has a center, much like the mean of a normal distribution, and permutations closer to this center will have higher probability. The parameter c encodes some information about the spread of this distribution, i.e. how quickly probabilities decrease as we move further from the center. The normalizing constant just makes everything sum to 1, making it a probability distribution. Thus, in many ways, this model

behaves somewhat like a normal distribution (as both are location-scale models) on a discrete set.

There are a few things to note here. First, while this is a discrete statistical model, and normally computing finitely many things is relatively simple, the number of permutations (and therefore the number of probabilities to compute) grows by $n!$. We might think this would hinder practical application of this model. But a closer examination shows that we need not do that many computations in order to make use of the model. Notice that every permutation that is a fixed distance from the center κ is assigned the same probability. That is, for every $\beta_1, \beta_2 \in S_n$ such that $\mathrm{d}(\beta_1, \kappa) = \mathrm{d}(\beta_2, \kappa)$, we have $p_{\beta_1} = p_{\beta_2}$ identically. Thus, in practice the number of probabilities necessary to compute in order to use the model is $\binom{n}{2}$, and therefore can be computed in polynomial time. To compute these $\binom{n}{2}$ probabilities we need to compute the normalizing constant $\psi(c)$. We note that while computing $\psi(c)$ may appear to have a non-trivial computation time, it is worth mentioning that by using the notation where $q = e^{-c}$, $\psi(c)$ is the q-analogue of $n!$ as we saw earlier in Proposition 1.2.8.

This same property of the Mallows model can be used to simplify other calculations involved with the Mallows model. Suppose we were to consider the map

$$\begin{aligned} \phi : \mathbb{R}[p_\pi \mid \pi \in S_n] &\to \mathbb{R}[t, q] \\ p_\pi &\mapsto t q^{\mathrm{inv}(\pi\kappa^{-1})} \end{aligned}$$

When seeking to understand the underlying structure of the Mallows model, we can ask about the underlying algebraic structure of this map. This would tell us something about the inter-relationships of the p_π. This map ϕ is actually the pullback of the ring homomorphism of the parameterization of this model. In other words, the map ϕ is the pullback of

$$\psi : \mathbb{K}^2 \to \mathbb{K}^{n!}$$

and we know that $\mathscr{I}\left(\mathrm{im}(\psi)\right) = \ker(\phi)$.

We can simplify the Gröbner basis calculation of this kernel by considering a ring in fewer variables, by taking the quotient of the ideal with the ideal defined by the relationships $p_{\beta_1} = p_{\beta_2}$ when $\mathrm{d}(\beta_1, \kappa) = \mathrm{d}(\beta_2, \kappa)$. Specifically, we consider a polynomial ring in fewer variables, say p_i where $i \in \left\{0, \ldots, \binom{n}{2}\right\}$ and let p_i be the probability of observing a permutation that is distance i from the center. By doing this, we would greatly reduce the number of variables, simplifying the actual calculation of the kernel of ϕ.

We can easily classify this ideal.

Theorem 2.1.1. *Consider the map*

$$\begin{aligned}\hat{\phi} : \mathbb{R}[p_i \mid i \in \{0,\dots,m\}] &\to \mathbb{R}[t,q] \\ p_i &\mapsto tq^i\end{aligned}$$

where $m = \binom{n}{2}$. *Then the kernel of* $\hat{\phi}$ *is a toric ideal generated by*

$$G = \left\{ p_i p_j - p_k p_\ell \mid i,j,k,\ell \in \{0,\dots,m\} \text{ and } i+j = k+\ell \right\}.$$

Furthermore, under the lexicographic monomial ordering where $p_i \succ p_j$ *for any* $i < j$, G *is a Gröbner basis for* $\ker(\hat{\phi})$.

Proof. We begin by showing that $G \subset \ker(\hat{\phi})$. Consider any polynomial of the form $p_i p_j - p_k p_\ell$ with $i,j,k,\ell \in \{0,\dots,m\}$ and $i+j = k+\ell$. Then we have that

$$\begin{aligned}\hat{\phi}(p_i p_j - p_k p_\ell) &= (tq^i)(tq^j) - (tq^k)(tq^\ell) \\ &= t^2 q^{i+j} - t^2 q^{k+\ell} \\ &= t^2 q^{i+j} - t^2 q^{i+j} \\ &= 0\end{aligned}$$

and therefore any polynomial $p_i p_j - p_k p_\ell$ with $i+j = k+\ell$ is in the kernel of $\hat{\phi}$.

Before we show G is a generating set, first we note that $\ker(\hat{\phi})$ is a homogeneous ideal. This can be seen by considering any non-homogeneous polynomial f in the ideal $\ker(\hat{\phi})$. Now consider the degree 2 part of f. We can see that this must be in the ideal $\ker(\hat{\phi})$ with the following argument. Every degree 2 monomial will map to a monomial where the exponent of the t variable is 2. The only way for this monomial (after mapping) to sum to zero is if another monomial whose degree of t is also 2 to cancel it out. We know that the only way a monomial can contain a t^2 is to have come from a monomial of degree 2 prior to applying $\hat{\phi}$. Thus, if $\hat{\phi}(f) = 0$, we know that all the terms containing a t^2 cancel out, and thus all the degree 2 monomials of f, or the degree 2 part of f, must be in the ideal $\ker(\hat{\phi})$. This technique can be applied for any degree which appears in f and therefor $\ker(\hat{\phi})$ is a homogeneous ideal.

Now, consider any polynomial $f \in \ker(\hat{\phi}) \subset \mathbb{K}[p_i \mid i \in \{0,\dots,m\}]$. Then we can write f as

$$f(\mathbf{p}) = r(\mathbf{p}) + \sum_{g_i \in G} h_i(\mathbf{p}) g_i(\mathbf{p})$$

where $r(\mathbf{p}) \in \ker(\hat{\phi})$ and no term of r is divisible by any monomial in $\mathrm{Lt}(G)$. But the only monomials

which do not appear in $\mathrm{Lt}(G)$ are monomials of the form $c_i p_i^{a_i} p_{i+1}^{b_i}$. Thus, in general, r has the form

$$r(\mathbf{p}) = \sum_i c_i p_i^{a_i} p_{i+1}^{b_i}$$

Because $r \in \ker(\hat{\phi})$, we also have that

$$\hat{\phi}(r) = 0 = \sum_i c_i t^{a_{i_j}+b_{i_j}} q^{ia_i+(i+1)b_i}$$

where $a_i, b_i \in \mathbb{Z}_{\geq 0}$ and not simultaneously zero (as no constant will be in the kernel of $\hat{\phi}$). Now, suppose, without loss of generality, that the monomial $p_i^a p_{i+1}^b$ appears in r. Then there exists a monomial $p_j^d p_{j+1}^e$ such that

$$\hat{\phi}(p_i^a p_{i+1}^b - p_j^d p_{j+1}^e) = 0 = t^{a+b} q^{ia+(i+1)b} - t^{d+e} q^{jd+(j+1)e}.$$

We know this will only be true if the following conditions hold

$$\begin{aligned} a+b &= d+e \\ ia+(i+1)b &= jd+(j+1)e \\ a,b,d,e &\in \mathbb{Z}_{\geq 0} \end{aligned}$$

where again the pairs a, b and d, e cannot be simultaneously equally to zero. We note that if $i = j$, then we have $b = e$ and $a = d$ and therefore these two monomials would cancel prior to mapping by $\hat{\phi}$.

We will manipulate the second equation by using the first in the following way:

$$\begin{aligned} ia+(i+1)b &= jd+(j+1)e \\ (a+b)i+b &= (d+e)j+e \\ (a+b)i+b &= (a+b)j+e \\ (a+b)(i-j)+b &= e \end{aligned}$$

Now we consider

$$(a+b)(i-j)+b = e \tag{2.1}$$

and use the original equation to do further manipulation in two ways. First, to simplify things, let $a+b = A$ and $d+e = D$. We note that then A, D are the degrees of the monomial and therefore $A = D$.

Without loss of generality, assume $i \geq j$. We can rewrite 2.1 as

$$A(i-j)+b=e \tag{2.2}$$

Then we see that

$$\begin{aligned} A(i-j)+b&=e \\ A(i-j)&=e-b\leq e\leq D=A \\ A(i-j)&\leq A \end{aligned}$$

Because $A \in \mathbb{Z}_{>0}$, we conclude that $i-j \leq 1$. Because $i-j \in \mathbb{Z}$, and we have assumed $i \geq j$, we have only two possibilities: either $i-j=0$ or $i-j=1$.

We have seen that $i-j=0$ implies that we would have two copies of the same monomial in r which would cancel each other out prior to mapping. We consider the remaining case: if $i-j=1$, we would have that

$$\hat{\phi}(p_{j+1}^a p_{j+2}^b - p_j^d p_{j+1}^e) = 0 = t^{a+b} q^{(j+1)a+(j+2)b} - t^{d+e} q^{jd+(j+1)e}.$$

This yields the following two equations

$$\begin{aligned} a+b&=d+e \\ (j+1)a+(j+2)b&=jd+(j+1)e \end{aligned}$$

and we can again manipulate the second of this two equations using the first to see that

$$\begin{aligned} (j+1)a+(j+2)b&=jd+(j+1)e \\ (j+1)a+(j+2)b+d&=jd+(j+1)e+d \\ (j+1)(a+b)+b+d&=(j+1)(e+d) \\ d+b&=0 \\ b&=-d \end{aligned}$$

and since $b, d \in \mathbb{Z}_{\geq 0}$ we have that $b = d = 0$ which implies that $a = e$. Thus we have

$$(p_{j+1}^a p_{j+2}^b - p_j^d p_{j+1}^e) = (p_{j+1}^a - p_{j+1}^a) = 0$$

and is therefore identically zero prior to mapping via $\hat{\phi}$.

Thus, we conclude that $r(\mathbf{p}) = 0$ and therefore every polynomial in the kernel of $\hat{\phi}$ can be written as the polynomial combination of polynomial in G and G is a generating set for $\ker(\hat{\phi})$.

□

Corollary 2.1.2. *Consider the map*

$$\begin{aligned} \phi : \mathbb{R}[p_\pi \mid \pi \in S_n] &\to \mathbb{R}[t, q] \\ p_\pi &\mapsto t q^{\mathrm{inv}(\pi\kappa^{-1})} \end{aligned}$$

Then the kernel of ϕ is the ideal generated by the union of the two sets

$$\{p_{\pi_1} p_{\pi_2} - p_{\sigma_1} p_{\sigma_2} \mid \pi_1, \pi_2, \sigma_1, \sigma_2 \in S_n \text{ and } \mathrm{d}(\pi_1, \kappa) + \mathrm{d}(\pi_2, \kappa) = \mathrm{d}(\sigma_1, \kappa) + \mathrm{d}(\sigma_2, \kappa)\}$$

$$\cup \{p_\pi - p_\sigma \mid \pi, \sigma \in S_n \text{ and } \mathrm{d}(\pi, \kappa) = \mathrm{d}(\sigma, \kappa)\} \ .$$

2.2 The Mallows Mixture Model

Mixture models are very common when modeling the behavior of a population. In this chapter, we will examine a model for a population whose behavior is described by the mixture of two Mallows models. The main goal in this section is to introduce this model and describe its vanishing ideal. We saw in the previous section that the vanishing ideal of the Mallows model could be simplified by simply using a single variable to represent the probability of observing any permutation which is distance i from the center of the first model and distance j from the center of the second.

Before we examine the Mallows mixture model and its characteristics, we recall some things about mixture models. Mixture models are used to describe populations which have one or more notable subpopulations, particularly when the subpopulations tend to behave in different ways. Formally, a mixture model has k base models each with parameter vector $\theta^{(i)}$ for $i = 1, \ldots, k$. If we let ω_i represent the weight of the ith model where $0 \leq \omega_i \leq 1$ and $\sum_{i=1}^k \omega_i = 1$, the probability assigned to an event A in the mixture model is given by

$$\Pr(X = A \mid \theta^{(1)}, \ldots, \theta^{(k)}) = \sum_{i=1}^k \omega_i \Pr(X = A \mid \theta^{(i)})$$

Random samples of the population of interest will have roughly the same proportions of the respective subpopulations as the population considered as a whole. Each subpopulation follows a particular distribution. The distribution of the overall population is a weighted sum of the distributions of the individual subpopulations. Because of their appeal and simplicity, mixture models are common in many different disciplines.

In general, if a population follows a mixture of k Mallows models with centers $\kappa_1, \ldots, \kappa_k$ which have scale parameters $c_1, \ldots, c_k$, then if we let p_π denote the probability assigned to a permutation π by the mixture model, p_π will have the form

$$p_\pi = \sum_{i=1}^{k} \frac{\omega_i}{\psi(c_i)} e^{-c_i \mathrm{d}(\pi, \kappa_i)}$$

where $0 \leq \omega_i \leq 1$ and $c_i \in \mathbb{R}_+$ for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} \omega_i = 1$.

In this chapter, we assume that the overall population we wish to model consists of two different sub-populations, each of which follows a Mallows model. As a result, the probability of observing a particular ranking will be the weighted sum of the probabilities of the ranking assigned by the two sub-distributions.

Definition 2.2.1. Given permutations $\kappa_1, \kappa_2 \in S_n$, then the mixture of two Mallows models with centers κ_1, κ_2 is the family of distributions given by

$$\mathscr{M}_{\kappa_1,\kappa_2} = \left\{ \mathscr{P} \in \mathbb{R}^{n!} \mid p_\pi = \frac{\omega_1}{\psi(c_1)} e^{-c_1 \mathrm{d}(\pi,\kappa_1)} + \frac{\omega_2}{\psi(c_2)} e^{-c_2 \mathrm{d}(\pi,\kappa_2)} \text{ where } c_1, c_2 \in \mathbb{R}_+,\ 0 \leq \omega_1, \omega_2 \leq 1, \text{ and } \omega_1 + \omega_2 = 1 \right\}$$

where $0 \leq \omega_1, \omega_2 \leq 1$ and $\omega_1 + \omega_2 = 1$, $c_1, c_2 \in \mathbb{R}_+$. We say ω_1, ω_2 are the weights of the two sub-distributions and c_1, c_2 are the spreads associated with the two sub-distributions.

We can again simplify the notation by letting $q_1 = e^{-c_1}$, $q_2 = e^{-c_2}$ and then lumping the weight in with the normalizing constant to let $t_1 = \dfrac{\omega_1}{\psi(c_1)}$, $t_2 = \dfrac{\omega_2}{\psi(c_2)}$. Then we have

$$p_\pi = t_1 q_1^{\mathrm{d}(\pi,\kappa_1)} + t_2 q_2^{\mathrm{d}(\pi,\kappa_2)} = t_1 q_1^{\mathrm{inv}(\pi\kappa_1^{-1})} + t_2 q_2^{\mathrm{inv}(\pi\kappa_2^{-1})}.$$

Similar to what we did with the Mallows model in the previous section, we examine the underlying algebraic structure of the mixture model. To do this, we look at the map

$$\phi : \mathbb{R}[p_\pi \mid \pi \in S_n] \to \mathbb{R}[t_1, t_2, q_1, q_2]$$
$$p_\pi \mapsto t_1 q_1^{\mathrm{inv}(\pi\kappa_1^{-1})} + t_2 q_2^{\mathrm{inv}(\pi\kappa_2^{-1})}$$

and consider the kernel of this map. The kernel of ϕ is the homogeneous vanishing ideal of the of the Mallows mixture model.

We can again simplify any calculations we may wish to do by noticing a relationship to those in the original Mallows model. We notice that any permutations $\beta_1, \beta_2 \in S_n$ such that $\mathrm{d}(\beta_1, \kappa_1) = \mathrm{d}(\beta_2, \kappa_1)$ and $\mathrm{d}(\beta_1, \kappa_2) = \mathrm{d}(\beta_2, \kappa_2)$ satisfy that $p_{\beta_1} = p_{\beta_2}$ by definition. Thus, instead of using a variable for each of the $n!$ permutations, we consider all possible pairs (i, j) such that there exists a $\beta \in S_n$ with $\mathrm{d}(\beta, \kappa_1) = i$ and $\mathrm{d}(\beta, \kappa_2) = j$. We do this by using the following definition:

Definition 2.2.2. Given any two permutations $\kappa_1, \kappa_2 \in S_n$. Define the *set of bi-distance pairs of the permutations* κ_1, κ_2, denoted $\mathscr{G}(\kappa_1, \kappa_2)$, as follows:

$$\mathscr{G}(\kappa_1, \kappa_2) = \left\{(i, j) \in \{0, 1, \ldots m\}^2 \ \mid \ \exists\, \beta \in S_n \text{ such that } d(\beta, \kappa_1) = i \text{ and } d(\beta, \kappa_2) = j\right\}.$$

where $m = \binom{n}{2}$.

We can then simplify our calculations by considering the pairs $(i, j) \in \mathscr{G}(\kappa_1, \kappa_2)$ instead of all the permutations in S_n. So, instead of considering the polynomial ring $\mathbb{R}[p_\pi \mid \pi \in S_n]$, we can use the polynomial ring $\mathbb{R}[p_{i,j} \mid (i, j) \in \mathscr{G}(\kappa_1, \kappa_2)]$. If we let $m = \binom{n}{2}$ and $\psi_1 : \mathbb{R}[p_\pi \mid \pi \in S_n] \longrightarrow \mathbb{R}[p_{i,j} \mid (i, j) \in \mathscr{G}(\kappa_1, \kappa_2)]$ be the projection map $\psi_1(p_\pi) = p_{i,j}$ where $i = \mathrm{d}(\pi, \kappa_1)$ and $j = \mathrm{d}(\pi, \kappa_2)$ and $\psi_2 : \mathbb{R}[p_{i,j} \mid (i, j) \in\in \mathscr{G}(\kappa_1, \kappa_2)] \longrightarrow \mathbb{R}[t_1, t_2, q_1, q_2]$ defined by $\psi_2(p_{i,j}) = t_1 q_1^i + t_2 q_2^j$, then the map ϕ is the composition of maps $\phi = \psi_2 \circ \psi_1$. We can think of ϕ as:

$$\phi : \mathbb{R}[p_\pi \mid \pi \in S_n] \xrightarrow{\psi_1} \mathbb{R}\left[p_{i,j} \mid (i, j) \in \mathscr{G}(\kappa_1, \kappa_2)\right] \xrightarrow{\psi_2} \mathbb{R}[t_1, t_2, q_1, q_2]$$

This polynomial ring has significantly fewer variables than the ring $\mathbb{R}[p_\pi \mid \pi \in S_n]$. Note that $\psi_1(\ker(\phi)) = \ker(\psi_2)$ and therefore we have that $\ker(\psi_2)$ is isomorphic to the quotient of $\ker(\phi)$ by

$$\left\langle p_\pi - p_\sigma \mid \pi, \sigma \in S_n \text{ and } \mathrm{d}(\pi, \kappa_1) = \mathrm{d}(\sigma, \kappa_1) \text{ and } \mathrm{d}(\pi, \kappa_2) = \mathrm{d}(\sigma, \kappa_2)\right\rangle \ .$$

The map ϕ is not the vanishing ideal of a statistical model, as $p_{i,j}$ do not satisfy $\sum_{(i,j)} p_{i,j} \neq 1$. If we were to know exactly, how many permutations have bi-distance (i, j) from a particular permutation, we could create a probability distribution using these $p_{i,j}$. We look at such a generating function in

Chapter 3.

2.3 The Set of Bi-Distance Pairs of Two Permutations

In the previous section, we defined $\mathscr{G}(\kappa_1,\kappa_2)$ as the set of bi-distance pairs of two permutations κ_1,κ_2. This set consists of all pairs (i,j) such that there exists a permutation in S_n which is distance i from κ_1 and distance j from κ_2. We have seen how this is may be able to simplify the computation of the vanishing ideal of the model. In this section, we will characterize exactly which (i,j) pairs will appear in $\mathscr{G}(\kappa_1,\kappa_2)\subset\{0,\dots,m\}^2$, where $m=\binom{n}{2}$. We will fix the notation $m=\binom{n}{2}$ and use it throughout this section. Before we can make use of the simplified computations available to us by considering $\phi=\psi_2\circ\phi_1$, we need to fully characterize $\mathscr{G}(\kappa_1,\kappa_2)$. We do this by using properties of the sign of a permutation, the triangle inequality (as Kendall's tau metric is a metric and therefore is subject to the triangle inequality), and Braid relations. We will need to introduce some new concepts which will be used in the section in order to facilitate the characterization of $\mathscr{G}(\kappa_1,\kappa_2)$. We will start by using the right-invariant property of Kendall's tau metric to show the relationship between the set of bi-distance pairs $\mathscr{G}(\kappa_1,\kappa_2)$ with the set of bi-distance paris $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$.

Consider the symmetric group S_n for some fixed n with Kendall's tau distance $d(\pi,\sigma)$ as a metric. Recall this metric is right invariant, and therefore $\mathrm{d}(\pi,\sigma)=\mathrm{d}(\pi\sigma^{-1},\mathrm{id})$ (see Lebanon and Mao [26]). The maximum value for $\mathrm{d}(\pi,\sigma)$ is $\binom{n}{2}$ for any $\pi,\sigma\in S_n$. The following lemma will prove useful for simplifying the proofs in this section:

Lemma 2.3.1. *For any $\kappa_1,\kappa_2\in S_n$, we have that $\mathscr{G}(\kappa_1,\kappa_2)=\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$*

Proof. Consider any element $(x,y)\in\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$; there exists a $\beta\in S_n$ such that $\mathrm{d}(\kappa_1\kappa_2^{-1},\beta)=x$ and $\mathrm{d}(\mathrm{id},\beta)=y$. The element $\beta\kappa_2$ satisfies the same relationship in $\mathscr{G}(\kappa_1,\kappa_2)$ as $\mathrm{d}(\pi,\beta\kappa_2)=\mathrm{d}(\kappa_1\kappa_2^{-1},\beta)=x$ and $\mathrm{d}(\kappa_2,\beta\kappa_2^{-1})=\mathrm{d}(\mathrm{id},\beta)=y$. So $(x,y)\in\mathscr{G}(\kappa_1,\kappa_2)$. The converse is also true; if $(x,y)\in\mathscr{G}(\kappa_1,\kappa_2)$, then there exists $\beta\in S_n$ such that $\mathrm{d}(\kappa_1,\beta)=x$ and $\mathrm{d}(\kappa_2,\beta)=y$ which implies $\mathrm{d}(\kappa_1,\beta)=\mathrm{d}(\kappa_1\kappa_2^{-1},\beta\kappa_2^{-1})=x$ and $\mathrm{d}(\kappa_2,\beta)=\mathrm{d}(\mathrm{id},\beta\kappa_2^{-1})=y$ so $(x,y)\in\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$. Thus $\mathscr{G}(\kappa_1,\kappa_2)=\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$. □

Because the sets $\mathscr{G}(\kappa_1,\kappa_2)$ and $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$ are identical, we work to characterize $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$, as it is easier to do a number of computations in $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$. For instance, since S_n is a Coxeter group, we will use our knowledge of Coxeter groups to help characterize $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$. It is easier to consider $\mathscr{G}(\kappa_1\kappa_2^{-1},\mathrm{id})$ when working with the idea of a longest word and a reduced expressions for a permutation in terms of the adjacent transpositions. We will use this result to tell us exactly how to find all of the pairs $(i,j)\in\mathscr{G}(\kappa_1,\kappa_2)$ in Theorem 2.3.7.

We next define the reversal of a permutation.

Definition 2.3.2. For any permutation $\pi = \pi(1)\ \pi(2)\ldots\pi(n)$ given in one line notation, define *the reversal of* π as the permutation given by $\text{rev}(\pi) = \pi(n)\ldots\pi(2)\ \pi(1)$. In the simplest terms, if you are given π in one line notation, reading π backwards will give you $\text{rev}(\pi)$ in one line notation.

The reversal of the identity is significant as it corresponds to the longest word in S_n when considered as a Coxeter group. We denote $\text{rev}(\text{id}) = w_0$. We can then show that $\text{rev}(\pi) = \pi w_0 = \pi \circ n \ldots 1 = \pi(n)\cdots\pi(1)$.

Recall that S_n is generated by the adjacent transpositions $s_1, \ldots, s_m$ and we saw in Definition 1.2.10 that all elements of S_n can be represented as a product of these s_i. Definition 1.2.10 tells us that a word was reduced if it is written as a product $s_{i_1}\cdots s_{i_\ell}$ if there does not exist $k < \ell$ such that $s_{i_1}\cdots s_{i_\ell} = s_{j_1}\cdots s_{j_k}$. Every element of S_n can be represented as a reduced word, though in general this representation is not unique. The minimal number of adjacent transpositions required to represent an element (i.e., if $s_{i_1}\cdots s_{i_\ell}$, the ℓ) is unique, and is defined as a the length of the word (or element). It is also true that the length of a permutation π when considered as a word is the same as $\text{d}(\text{id}, \pi)$. We can also define a maximal chain in S_n, when it is considered as a Coxeter group. A maximal chain in S_n has the form $s_{i_1}\ldots s_{i_m}$ where $\text{d}(s_{i_1}\ldots s_{i_j},\ s_{i_1}\ldots s_{i_{j+1}}) = 1$ for all $j = 1, \ldots, n-1$, and $s_{i_1}\ldots s_{i_m} = w_0$ where again $m = \binom{n}{2}$. Because w_0 functions as the $\hat{1}$ element in S_n under the weak (Bruhat) left order from Definition 1.2.11 (see for instance [5, 37]), all maximal chains are a reduced expression of w_0 when considered as a word.

Lemma 2.3.3. *For any $\tau \in S_n$, if we let $m = \binom{n}{2}$, there exists a sequence of adjacent transpositions $s_{i_1}\cdots s_{i_m}$ such that $s_{i_1}\cdots s_{i_m} = w_0$ and $s_{i_{m-k}}\cdots s_{i_m} = \tau$ where k is the length of τ (i.e. $d(\text{id}, \tau) = k$).*

Proof. Take any $\tau \in S_n$ where $\text{d}(\text{id}, \tau) = \ell$ and let w_0 be the longest word. In Definition 1.2.11, we introduced the weak (Bruhat) left order on S_n, and we know w_0 functions as $\hat{1}$ in this poset. Then we have that $\tau \leq w_0$ under the weak Bruhat left order, which means that by definition there exists a reduced expression $s_{i_1}\cdots s_{i_m} = w_0$ such that $s_{i_k}\cdots s_{i_m} = \tau$ for some k which is also a reduced expression. Because the length of a reduced expression for a fixed word is unique, we know that $k = \ell$ and we have shown that for every element $\tau \in S_n$, there exists a maximal chain which passes through τ. □

Lemma 2.3.4. *For any $\sigma \in S_n$ with $d(\text{id}, \sigma) = k$ with $\alpha_1\cdots\alpha_k = \sigma$ and $\beta_1\cdots\beta_k = \sigma$ distinct reduced expressions of σ ($\alpha_i, \beta_i \in \{s_1, \ldots, s_{n-1}\}$ adjacent transpositions for all $i \in [k]$), there exists a sequence of braid moves which transforms $\alpha_1\cdots\alpha_k$ to $\beta_1\cdots\beta_k$.*

Proof. Recall in 1.2.12 that a braid move of a braid group generated by $\{s_1, \ldots, s_{n-1}\}$ is a substitution of the form $s_i s_j = s_j s_i$ where $|i - j| > 1$ or $s_i s_{i+1} s_i = s_{i+1} s_i s_{i+1}$. This theorem is central to all work

done in Coxeter groups and the proof can be found in many places, including [5] (labeled as Word Property Theorem and Theorem 3.3.1). □

Using the relations on S_n as a Coxeter group, we will get a concrete relationship between $\mathrm{d}(\mathrm{id},\beta)$ and $\mathrm{d}(\beta, w_0)$.

Lemma 2.3.5. *For any $\beta \in S_n$, we have that* $\mathrm{d}(\beta, w_0) = \mathrm{d}(\mathrm{id}, w_0) - \mathrm{d}(\mathrm{id},\beta) = m - \mathrm{d}(\mathrm{id},\beta)$.

Proof. Take any $\beta \in S_n$. We will continue to denote w_0 as the reversal of the identity element, the element furthest from the identity. Lemma2.3.3 tells us that for any β, there is a reduced expression $s_{i_1} \dots s_{i_m}$ for w_0 such that $s_{i_{m-\ell+1}} \dots s_{i_m} = \beta$ for $\ell = \mathrm{d}(\mathrm{id},\beta)$. Any initial (or final) substring of a reduced word is itself reduced, meaning that

$$\mathrm{d}(\mathrm{id}, w_0) = \mathrm{d}(\mathrm{id}, s_{i_{m-\ell+1}} \dots s_{i_m}) + \mathrm{d}(s_{i_{m-\ell+1}} \dots s_{i_m}, s_{i_1} \dots s_{i_m}) .$$

Similarly, $\mathrm{d}(\beta, w_0)$ will be found by the exact $m-\ell$ adjacent transpositions $s_{i_1} \dots s_{i_{m-\ell}}$. Again, this is a reduced expression so we see that $\mathrm{d}(\beta, w_0) = m - \ell = \mathrm{d}(\mathrm{id}, w_0) - \mathrm{d}(\mathrm{id},\beta)$. □

As a direct result of this, we can see that:

$$\mathrm{d}(\beta, w_0) = \mathrm{d}(\mathrm{id}, w_0) - \mathrm{d}(\mathrm{id},\beta) \;\Rightarrow\; \mathrm{d}(\beta\tau, w_0\tau) = m - \mathrm{d}(\tau, \beta\tau) \qquad \forall \tau, \beta \in S_n .$$

If we call $\beta\tau = \beta'$, we see that $\mathrm{d}(\beta', w_0\tau) = m - \mathrm{d}(\tau, \beta')$ for all $\tau, \beta' \in S_n$. With these definitions and lemmas, we are able to enumerate all pairs $(i,j) \in \mathscr{G}(\kappa_1, \kappa_2)$.

Because we are interested in characterizing exactly the (i,j) pairs which can be found in $\mathscr{G}(\kappa_1,\kappa_2)$, we introduce a function which will help us find all the (i,j) which arise from all the substrings of the reduced expression for an element of S_n.

Definition 2.3.6. Fix a permutation $\tau \in S_n$ and let $\mathrm{d}(\mathrm{id},\tau) = k$. Then for any reduced expression $\alpha_1 \cdots \alpha_\ell$ of a word in S_n (where here the $\alpha_i \in \{s_1, \dots, s_{n-1}\}$ for $i \in [k]$ are adjacent transpositions), define the function $F_\tau(\alpha_1 \cdots \alpha_\ell)$ as

$$F_\tau(\alpha_1 \cdots \alpha_\ell) :=$$
$$\big\{\big(\mathrm{d}(\mathrm{id},\mathrm{id}), \mathrm{d}(\mathrm{id},\tau)\big), \big(\mathrm{d}(\alpha_\ell,\mathrm{id}), \mathrm{d}(\alpha_\ell,\tau)\big), \big(\mathrm{d}(\alpha_{\ell-1}\alpha_\ell,\mathrm{id}), \mathrm{d}(\alpha_{\ell-1}\alpha_\ell,\tau)\big), \dots, \big(\mathrm{d}(\alpha_1\cdots\alpha_\ell,\mathrm{id}), \mathrm{d}(\alpha_1\cdots\alpha_\ell,\tau)\big)\big\}.$$

In other words, this is the function that recovers all paired distances from final substrings of a reduced word to id, τ. We refer to it as the *paired distance function*.

The first element of $F_\tau(\alpha_1 \cdots \alpha_k)$ will be $(0, r)$ where $\mathrm{d}(\mathrm{id},\tau) = r$, regardless of the length k of the reduced word $\alpha_1 \cdots \alpha_k$. Furthermore, if we consider elements of $F_\tau(\alpha_1 \cdots \alpha_k)$ as moves along a lattice,

we know that there are only two possible moves between elements: $(1,1)$ and $(1,-1)$. It should be clear that the first coordinate will always increase by one as all final substrings of a reduced word must themselves be reduced, and therefore the length of the word will correspond to its distance from the identity. Since the length of the word increases by one letter between elements of $F_\tau(\alpha_1 \cdots \alpha_k)$, the first coordinate will increase by one between each two consecutive elements by definition. To see that the second coordinate of any move must be either -1 or 1, we note that the coordinate must change by an odd number due to the sign of a permutation (for more on this, see the proof of Proposition 2.3.9).

Intuitively, if we add a single adjacent transposition to our word, we are either moving one step close to τ or one step further away from it. To see that it can only be a -1 or 1 step, consider the second coordinate of two consecutive elements of $F_\tau(\alpha_1 \cdots \alpha_k)$, $\mathrm{d}(\alpha_{k-i} \cdots \alpha_k, \tau) = \mathrm{d}(\tau, \alpha_{k-i} \cdots \alpha_k)$ and $\mathrm{d}(\alpha_{k-i+1} \cdots \alpha_k, \tau) = \mathrm{d}(\tau, \alpha_{k-i+1} \cdots \alpha_k)$. The triangle inequality tell us

$$\begin{aligned}
&\mathrm{d}(\tau, \alpha_{k-i} \cdots \alpha_k) \leq \mathrm{d}(\tau, \alpha_{k-i+1} \cdots \alpha_k) + \mathrm{d}(\alpha_{k-i+1} \cdots \alpha_k, \alpha_{k-i} \cdots \alpha_k) \\
&\mathrm{d}(\tau, \alpha_{k-i} \cdots \alpha_k) \leq \mathrm{d}(\tau, \alpha_{k-i+1} \cdots \alpha_k) + 1 \\
&\mathrm{d}(\tau, \alpha_{k-i} \cdots \alpha_k) - \mathrm{d}(\tau, \alpha_{k-i+1} \cdots \alpha_k) \leq 1
\end{aligned}$$

and the distance between the second coordinate of two consecutive elements of $F_\tau(\alpha_1 \cdots \alpha_k)$ is less than 1. We can get that $-1 \leq \mathrm{d}(\tau, \alpha_{k-i} \cdots \alpha_k) - \mathrm{d}(\tau, \alpha_{k-i+1} \cdots \alpha_k)$ by using a similar approach, and therefore we can conclude that the change in the second coordinate of two consecutive elements of $F_\tau(\alpha_1 \cdots \alpha_k)$ is either 1 or -1. When a reduced expression for the longest word is plugged into the paired distance function, $F_\tau(\alpha_1 \cdots \alpha_m)$ will have $(m, m-r)$ as its final element.

Theorem 2.3.7. *Let $\kappa_1, \kappa_2 \in S_n$ be any permutations in the symmetric group of size n. Let $r = \mathrm{d}(\kappa_1, \kappa_2)$ then*

$$\mathscr{G}(\kappa_1, \kappa_2) = \left\{(i,j) \mid (i+j) \equiv r \bmod 2 \text{ and } r \leq i+j \leq 2m-r \text{ and } |i-j| \leq r\right\}.$$

We define a set related to $\mathscr{G}(\kappa_1, \kappa_2)$ as follows

Definition 2.3.8. Let $r, m \in \mathbb{Z}$ such that $0 \leq r < m$. Define the set $\mathscr{H}(r,m)$ as

$$\mathscr{H}(r,m) := \{(i,j) \in \{0, \ldots, m\}^2 \mid (i+j) \equiv r \bmod 2 \text{ and } r \leq i+j \leq 2m-r \text{ and } |i-j| \leq r\}.$$

We refer to this set as the set of Coxeter relations imposed by r, m.

This definition will simply be used to ease the proofs of the Theorems which follow. The proof of this theorem is somewhat complex. We break the proof up into smaller, more manageable pieces.

Proposition 2.3.9. *Let $\kappa_1, \kappa_2 \in S_n$ and $m = \binom{n}{2}$. Then*

$$\mathscr{G}(\kappa_1, \kappa_2) \subseteq \mathscr{H}\big(\mathrm{d}(\kappa_1, \kappa_2), m\big).$$

Proof. Take any $\kappa_1, \kappa_2 \in S_n$ and call $r = \mathrm{d}(\kappa_1, \kappa_2) = \mathrm{d}(\kappa_1\kappa_2^{-1}, \mathrm{id})$. For the sake of simplicity, let $\tau = \kappa_1\kappa_2^{-1}$. Let $\mathrm{d}(\mathrm{id}, \tau) = r$. By the definition of Kendall's tau distance and the reduced expression of a word, there exists a string of r adjacent transpositions $\alpha_1, \ldots, \alpha_r$ such that $\tau = \alpha_1 \cdots \alpha_r$ is a reduced expression. Note that due to the large number of instances of strings of adjacent transpositions, we will not denote the adjacent transpositions exclusively as $s_1, \ldots, s_{n-1}$. This will allow us to avoid instances where we need very elaborate subscripts. As shown above, $\mathscr{G}(\kappa_1, \kappa_2) = \mathscr{G}(\kappa_1\kappa_2^{-1}, \mathrm{id}) = \mathscr{G}(\tau, \mathrm{id})$. Take any $(i, j) \in \mathscr{G}(\tau, \mathrm{id})$. Suppose $(i + j) \not\equiv r \bmod 2$. We know there exists a $\beta \in S_n$ such that $\mathrm{d}(\tau, \beta) = i$ and $\mathrm{d}(\mathrm{id}, \beta) = j$. This means there are adjacent transpositions $\gamma_1, \ldots, \gamma_i$ and $\rho_1, \ldots, \rho_j$ such that

$$\tau = \gamma_1 \ldots \gamma_i \beta \quad \text{and} \quad \mathrm{id} = \rho_1 \ldots \rho_j \beta \ .$$

Using the second equation, we see that $\beta = \rho_j \ldots \rho_1$. Upon substituting, we see that $\mathrm{sgn}(\tau) = \mathrm{sgn}(\gamma_1 \ldots \gamma_i \rho_j \ldots \rho_1) = (-1)^{(i+j)}$. We know that $\tau = \alpha_1 \ldots \alpha_r$ and therefore $\mathrm{sgn}(\tau) = (-1)^r$. Since we assumed $(i + j) \not\equiv r \bmod 2$, we get that the sign of τ is both positive and negative, which is a contradiction as the sign function is well defined on S_n. Thus, we conclude that $(i + j) \equiv r \bmod 2$ for all $(i, j) \in \mathscr{G}(\tau, \mathrm{id})$.

Using the triangle inequality, we deduce the lower bound on $i + j$. Suppose there exists a pair $(i, j) \in \mathscr{G}(\tau, \mathrm{id})$ such that $i + j < r$. Then there exists a $\beta \in S_n$ such that $\mathrm{d}(\tau, \beta) = i$ and $\mathrm{d}(\mathrm{id}, \beta) = j$. Since $\mathrm{d}(\pi, \sigma)$ is a metric, we use the triangle inequality to see

$$\mathrm{d}(\tau, \mathrm{id}) \leq \mathrm{d}(\tau, \beta) + \mathrm{d}(\mathrm{id}, \beta) \quad \Rightarrow \quad r \leq i + j \ .$$

This is a contradiction to $i + j < r$ so we conclude that $r \leq i + j$ for all $(i, j) \in \mathscr{G}(\tau, \mathrm{id})$.

Next we look at the upper bound for $i + j$. Using the above lemma, we see:

$$\mathrm{d}(\tau w_0, \beta) = \tbinom{n}{2} - \mathrm{d}(\beta, \tau) = \tbinom{n}{2} - i \quad , \quad \mathrm{d}(\beta, w_0) = \tbinom{n}{2} - \mathrm{d}(\beta, \mathrm{id}) = \tbinom{n}{2} - j$$
$$\Rightarrow \mathrm{d}(\tau w_0, w_0) \leq \mathrm{d}(\tau w_0, \beta) + \mathrm{d}(\beta, w_0) = 2\tbinom{n}{2} - (i + j) \Rightarrow i + j \leq 2m - \mathrm{d}(\tau w_0, w_0) = 2m - \mathrm{d}(\tau, \mathrm{id})$$

Finally, we will look at the bounds for $i - j$. Take any $(i, j) \in \mathscr{G}(\tau, \mathrm{id})$. Then there exists an element $\beta \in S_n$ such that $d(\tau, \beta) = i$ and $d(\mathrm{id}, \beta) = j$. Using the triangle inequality we see:

$$\mathrm{d}(\tau, \beta) \leq \mathrm{d}(\tau, \mathrm{id}) + \mathrm{d}(\mathrm{id}, \beta) \ \Rightarrow \ i \leq r + j \ \Rightarrow \ i - j \leq r$$

$$\mathrm{d}(\mathrm{id},\beta) \le \mathrm{d}(\tau,\mathrm{id}) + \mathrm{d}(\tau,\beta) \;\Rightarrow\; j \le r+i \;\Rightarrow\; -r \le i-j\ .$$

Combining these two inequalities, we see $|i-j| \le r$. Thus, the bounds for the entries $(i,j) \in \mathscr{G}(\tau,\mathrm{id})$ are accurate. □

It remains to show that every combination of $(i,j) \in \mathscr{H}(r.m)$ is realized in $\mathscr{G}(\tau,\mathrm{id})$. We introduce the one more lemma before proving Theorem 2.3.7.

Lemma 2.3.10. *Let $n \ge 4$ and let $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ be reduced expressions for $w_0 \in S_n$ which differ by a single braid move (where $\alpha_i, \beta_i \in \{s_1, \ldots, s_{n-1}\}$ adjacent transpositions with $m = \binom{n}{2}$). Then $F_\tau(\alpha_1 \cdots \alpha_m)$ and $F_\tau(\beta_1 \cdots \beta_m)$ differ by at most two elements.*

Proof. We have already seen that we can view elements of $F_\tau(\alpha_1 \cdots \alpha_m)$ as coordinates along a lattice path starting at $(0,r)$ with steps $(1,1)$ or $(1,-1)$ (where $\mathrm{d}(\mathrm{id},\tau) = r$ and $m = \binom{n}{2}$). Furthermore, since $\alpha_1 \cdots \alpha_m$ is a reduced expression for w_0, we know that $F_\tau(\alpha_1 \cdots \alpha_m)$ will be such a path from $(0,r)$ to $(m, m-r)$. We know $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ are bot reduced expressions for w_0 and they differ by a single braid move. Recall from Definition 1.2.12 that braid moves involve making a substitution of the form $s_i s_j = s_j s_i$ or $s_i s_{i+1} s_1 = s_{i+1} s_i s_{i+1}$ where $|i-j| > 1 (i \in [n-2]$ and $j \in [n-1])$. We know that as a word, $\alpha_1 \cdots \alpha_m = \beta_1 \cdots \beta_m = w_0$. Because the differ by s ingle braid move, we have two cases. In the first case, $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ differ by a braid move of the form $s_i s_j = s_j s_i$ where $|i-j > 1$. Then we have that

$$\alpha_1 \cdots \alpha_\ell s_i s_j \alpha_{\ell+3} \cdots \alpha_m = \beta_1 \cdots \beta_\ell s_j s_i \beta_{\ell+3} \cdots \beta_m$$

and $\alpha_t = \beta_t$ for all $t \in [m]$ with $t \ne \ell+1, \ell+2$. Then the only final substrings of $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ which will be different will be $s_j \alpha_{\ell+3} \cdots \alpha_m$ and $s_i \beta_{\ell+3} \cdots \beta_m$, since $s_i s_j \alpha_{\ell+3} \cdots \alpha_m = s_j s_i \beta_{\ell+3} \cdots \beta_m$ according to our braid relations. Since only one final substring is different, $F_\tau(\alpha_1 \cdots \alpha_m)$ and $F_\tau(\beta_1 \cdots \beta_m)$ differ by at most 1 element.

Consider the case where $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ differ by a braid move of the form $s_i s_{i+1} s_1 = s_{i+1} s_i s_{i+1}$. Similar to before, we have that

$$\alpha_1 \cdots \alpha_\ell s_i s_{i+1} s_i \alpha_{\ell+4} \cdots \alpha_m = \beta_1 \cdots \beta_\ell s_{i+1} s_i s_{i+1} \beta_{\ell+4} \cdots \beta_m$$

and in this case there are at most two final substrings which differ in $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$, namely $s_i \alpha_{\ell+4} \cdots \alpha_m$ differs from $s_{i+1} \beta_{\ell+4} \cdots \beta_m$ and $s_{i+1} s_i \alpha_{\ell+4} \cdots \alpha_m$ differs from $s_i s_{i+1} \beta_{\ell+4} \cdots \beta_m$ (as, again, due to the equality of braid relationships, $s_i s_{i+1} s_i \alpha_{\ell+4} \cdots \alpha_m = s_{i+1} s_i s_{i+1} \beta_{\ell+4} \cdots \beta_m$). Thus, because only 2 final substrings of $\alpha_1 \cdots \alpha_m$ and $\beta_1 \cdots \beta_m$ differ, $F_\tau(\alpha_1 \cdots \alpha_m)$ and $F_\tau(\beta_1 \cdots \beta_m)$ differ by at most 2 elements. In both cases, $F_\tau(\alpha_1 \cdots \alpha_m)$ and $F_\tau(\beta_1 \cdots \beta_m)$ differ by at most 2 elements and we are done. □

Table 2.1 Substrings of R_1 of the form $\alpha_j^{(1)} \cdots \alpha_m^{(1)}$ with $j = 1, \ldots, m$ and their corresponding element in $\mathscr{G}(\mathrm{id}, \tau)$

Element in S_n	Element in $\mathscr{G}(\tau, \mathrm{id})$
id	$(0, r)$
$\alpha_m^{(1)}$	$(1, r-1)$
$\alpha_{m-1}^{(1)} \alpha_m^{(1)}$	$(2, r-2)$
$\vdots$	$\vdots$
$\alpha_{m-r}^{(1)} \cdots \alpha_m^{(1)}$	$(r, 0)$
$\alpha_{m-r-1}^{(1)} \cdots \alpha_{m-1}^{(1)} \alpha_m^{(1)}$	$(r+1, 1)$
$\vdots$	$\vdots$
$\alpha_1^{(1)} \cdots \alpha_m^{(1)}$	$(m, m-r)$

The following proposition will be the last part of proving Theorem 2.3.7.

Proposition 2.3.11. *Let $n \geq 4$ with $\tau \in S_n$, $d(\mathrm{id}, \tau) = r$, and $m = \binom{n}{2}$. Then*

$$\mathscr{G}(\mathrm{id}, \tau) \supseteq \mathscr{H}(r, m) .$$

Proof. By Lemma 2.3.3, there exist reduced expressions $\alpha_1 \cdots \alpha_m, \beta_1 \cdots \beta_m$ for w_0 such that $\alpha_{m-r} \cdots \alpha_m = \tau$ and $\beta_r \cdots \beta_m = \tau w_0$ are reduced expressions. To see that the length of τw_0 is in fact $m - r$, we use Lemma 2.3.5 to see that $\mathrm{d}(\tau w_0, w_0) = m - \mathrm{d}(\mathrm{id}, \tau w_0)$ and since $\mathrm{d}(\tau w_0, w_0) = \mathrm{d}(\tau, \mathrm{id}) = r$ (recall w_0 is self inverse), we have that $\mathrm{d}(\mathrm{id}, \tau \mathrm{w}_0) = \mathrm{m} - \mathrm{r}$. Note that $\mathrm{d}(\tau, \tau w_0) = \mathrm{d}(\tau^{-1}\tau, \tau^{-1}\tau w_0) = \mathrm{d}(\mathrm{id}, w_0) = m$. Thus, τw_0 is the element of S_n which is furthest from the element τ.

If we consider these reduced expressions as a path, we see that

$$F_\tau(\alpha_1 \cdots \alpha_m) = \{(0, r), (1, r-1), \ldots, (r-1, 1), (r, 0), (r+1, 1), \ldots, (m, m-r)\}$$
$$F_\tau(\beta_1 \cdots \beta_m) = \{(0, r), (1, r+1), \ldots, (m-r-1, m-1), (m-r, m), (m-r+1, m-1), \ldots, (m, m-r)\}$$

which, when considered together, form the boundaries of the set $\mathscr{G}(\mathrm{id}, \tau)$. By Lemma 2.3.4 there exists a series of braid moves which transforms $\alpha_1 \cdots \alpha_m$ to $\beta_1 \cdots \beta_m$. Let $R_1, \ldots, R_p$ be the sequence of reduced words for w_0 where R_i differs from R_{i+1} by a single braid move, $R_1 = \alpha_1 \cdots \alpha_m$ and $R_p = \beta_1 \cdots \beta_m$. We claim that

$$\bigcup_{i=1}^{k} F_\tau(R_i) = \mathscr{H}(r, m) .$$

We see in Table 2.1 the (i, j) pairs of $F_\tau(\alpha_1 \cdots \alpha_m)$ and the final substrings they correspond to.

Lemma 2.3.10 tells us that performing a single braid move on R_k changes at most two elements of

$F_\tau(R_k)$. Because we have the outermost extremes as our first and last reduced word, and we are allowed to change at most 2 elements of $F_\tau(R_k)$ for each braid transformation, and $F_\tau(R_k)$ can only make moves $(1,1)$ and $(1,-1)$, for a fixed i we must hit every j which satisfies the inequalities described by $\mathscr{H}(r,m)$. If we were to think of a braid move as deforming a lattice path (which indeed is an accurate way to describe a braid move), the a single braid move changes at most 2 points on the path. Since the path must start at $(0,r)$ and end at $(m,r-m)$, it is impossible for this series of braid moves to miss a single j which is in the bounds of $\mathscr{H}(r,m)$ for a fixed i. Any sequence of braid moves that takes R_1 to R_p must hit every $(i,j)\in\mathscr{H}(r,m)$. □

While the proof of Theorem 2.3.7 is very technical, the actual concept is very easy to follow. We demonstrate this with an example.

Example 2.3.12. Let $n=4$ and let $\tau = 1\,4\,3\,2 \in S_4$. We can calculate $\mathrm{d}(\tau,\mathrm{id})=3$, and in this case $m=\binom{n}{2}=6$ and $w_0=4\,3\,2\,1$. For any $\beta\in S_4$, there exists a sequence of reduced expressions $R_1,\dots,R_p$ where $R_j = s_{i_1^{(j)}}\cdots s_{i_m^{(j)}}$ for $j=1,\dots,p$, R_{i+1} comes from applying a single braid transformation to the expression R_i for all $i=1,\dots,p-1$, $s_{i_{m-r}^{(1)}}\cdots s_{i_m^{(1)}}=\tau$ where $\mathrm{d}(\tau,\mathrm{id})=r$ and $s_{i_{m-b}^{(p)}}\cdots s_{i_m^{(1)}}=\beta$ where $\mathrm{d}(\beta,\mathrm{id})=b$. For the purposes of this example, choose $\beta = 4\ 1\ 2\ 3$.

We know there exists a reduced expression for the longest word w_0 which passes through τ. In fact, there are multiple such expressions. We choose the expression $R_1 = s_3\,s_2\,s_1\,s_3\,s_2\,s_3$ which is a reduced expression for w_0 and see that $s_3\,s_2\,s_3=\tau$. If we apply the braid transformation $s_1s_3=s_3s_1$ to this path, we obtain the word $R_2 = s_3\,s_2\,s_3\,s_1\,s_2\,s_3$. Similar to what we did in Table 2.1, in Table 2.2 we compare each of the substrings obtained by successively adding adjacent transpositions on the left. We see that this corresponds to walking along the path from the identity, through τ, to the longest word w_0. It should be noted that computing the distance from each element in both Table 2.2a and Table 2.2b to the identity element as simple as counting the number of s_i in the expression. Note that there is only one permutation which differs in the two paths, and therefore there is only one new point $(i,j)\in\mathscr{G}(1\,4\,3\,2,\mathrm{id})$, namely the pair $(2,3)$.

We can also visualize this process using a matrix. First, define

$$S_{R_k} := \{s_j^{(k)}\cdots s_m^{(k)} \mid j=1,\dots,p\}$$

the set of all permutations are vertices of R_k when R_k is considered as a path from id to w_0 in S_n.

Let M_{τ,R_k} be the matrix given by

$$\left(M_{\tau,R_k}\right)_{i,j} = \begin{cases} X & \text{if } (i-1,j-1)\in F_\tau(R_k) \\ 0 & \text{otherwise} \end{cases}$$

We can then describe the braid transformation which takes W_1 to W_2 and see

$$M_{\tau,R_1} = \begin{pmatrix} 0 & 0 & 0 & X & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 0 & 0 & 0 \\ X & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & X & 0 & 0 & 0 \end{pmatrix} \xrightarrow{s_1 s_3 = s_3 s_1} M_{\tau,R_2} = \begin{pmatrix} 0 & 0 & 0 & X & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & X & 0 & 0 & 0 \end{pmatrix}.$$

Table 2.3 shows the sequence $R_1, \ldots, R_8$ along with the braid transformations and the associated $F_\tau(R_1), \ldots, F_\tau(R_8)$. The reader may observe that, in fact, R_7 is a reduced expression through $\beta = 4\ 1\ 2\ 3$. We choose to make one final braid move to create R_8 as doing so allows us to say that $\cup_{i=1}^{8} S_{R_i} = S_4$, and therefore we have hit every permutation and can say that $\mathscr{G}(1\ 4\ 3\ 2, \mathrm{id}) = \cup_{i=1}^{8} F_\tau(R_i)$. Using Table 2.3, we can see

$$\mathscr{G}(1\ 4\ 3\ 2, \mathrm{id}) =$$
$$\{(0,3), (1,2), (1,4), (2,1), (2,3), (2,5), (3,0), (3,2), (3,4), (3,6), (4,1), (4,3), (4,5), (5,2), (5,4), (6,3)\}$$

We see that any $(i, j) \in \{0, \ldots, 6\}^2$ which satisfy the conditions that $(i + j) \equiv 3 \bmod 2$ and $3 \leq i + j \leq 12 - 3 = 9$ and $|i - j| \leq 3$ is in fact an element of $\mathscr{G}(\ 1\ 4\ 3\ 2, \mathrm{id})$.

Table 2.2 Comparing each point of the paths R_1, R_2 where R_2 is obtained by performing the braid transformation $s_1 s_3 = s_3 s_1$ to the word R_1.

Table 2.2a The elements of S_4 along the path described by R_1 and their distance from τ

$\pi \in S_n$	π as product of s_i	d(τ, π)
1234	id	3
1243	s_3	2
1342	$s_2\, s_3$	1
$1432 = \tau$	$s_3\, s_2\, s_3$	0
2431	$s_1\, s_3\, s_2\, s_3$	1
3421	$s_2\, s_1\, s_3\, s_2\, s_3$	2
4321	$s_3\, s_2\, s_1\, s_3\, s_2\, s_3$	3

Table 2.2b The elements of S_4 along the path described by R_2 and their distance from τ

$\pi \in S_n$	π as product of s_i	d(τ, π)
1234	id	3
1243	s_3	2
1342	$s_2\, s_3$	1
2341	$s_1\, s_2\, s_3$	2
2431	$s_3\, s_1\, s_2\, s_3$	1
3421	$s_2\, s_3\, s_1\, s_2\, s_3$	2
4321	$s_3\, s_2\, s_3\, s_1\, s_2\, s_3$	3

Table 2.3 Sequence of braid transformations from $R_1, \ldots, R_8$ and the (i, j) pairs found in each $F_\tau(R_k)$.

R_{k-1}	Braid Transformation	R_k	$F_\tau(R_k)$
		$R_1 = s_3s_2s_1s_3s_2s_3$	{(0,3), (1,2), (2,1), (3,0), (4,1), (5,2), (6,3)}
$s_3s_2s_1s_3s_2s_3$	$\xrightarrow{s_1s_3=s_3s_1}$	$R_2 = s_3s_2s_3s_1s_2s_3$	{(0,3), (1,2), (2,1), (3,2), (4,1), (5,2), (6,3)}
$s_3s_2s_3s_1s_2s_3$	$\xrightarrow{s_3s_2s_3=s_2s_3s_2}$	$R_3 = s_2s_3s_2s_1s_2s_3$	{(0,3), (1,2), (2,1), (3,2), (4,3), (5,4), (6,3)}
$s_2s_3s_2s_1s_2s_3$	$\xrightarrow{s_2s_1s_2=s_1s_2s_1}$	$R_4 = s_2s_3s_1s_2s_1s_3$	{(0,3), (1,2), (2,3), (3,4), (4,3), (5,4), (6,3)}
$s_2s_3s_1s_2s_1s_3$	$\xrightarrow{s_1s_3=s_3s_1}$	$R_5 = s_2s_3s_1s_2s_3s_1$	{(0,3), (1,4), (2,3), (3,4), (4,3), (5,4), (6,3)}
$s_2s_3s_1s_2s_3s_1$	$\xrightarrow{s_3s_1=s_1s_3}$	$R_6 = s_2s_1s_3s_2s_3s_1$	{(0,3), (1,4), (2,3), (3,4), (4,5), (5,4), (6,3)}
$s_2s_1s_3s_2s_3s_1$	$\xrightarrow{s_3s_2s_3=s_2s_3s_2}$	$R_7 = s_2s_1s_2s_3s_2s_1$	{(0,3), (1,4), (2,5), (3,6), (4,5), (5,4), (6,3)}
$s_2s_1s_2s_3s_2s_1$	$\xrightarrow{s_2s_1s_2=s_1s_2s_1}$	$R_8 = s_1s_2s_1s_3s_2s_1$	{(0,3), (1,4), (2,5), (3,6), (4,5), (5,4), (6,3)}

2.4 Properties of Joins of Ideals

The vanishing ideal of a mixture model can be written as the join of the vanishing ideals of the underlying models (see, for instance [14]). As such, we can characterize the vanishing ideal of the mixture of two Mallows models introduced in Section 2.2 as the join of the vanishing ideals for two Mallows models. We have characterized the vanishing ideal of a Mallows model in Corollary 2.1.2 and Theorem 2.1.1. At this point, the reader may notice in Theorem 2.1.1 the vanishing ideal of the individual Mallows model is the kernel of $\hat{\phi} : \mathbb{R}[p_i \mid i \in \{0, \ldots, m\}] \longrightarrow \mathbb{R}[t, q]$ whereas the vanishing ideal of the mixture model defined in Section 2.2 is the kernel of $\phi : \mathbb{R}[p_\pi \mid \pi \in S_n] \longrightarrow \mathbb{R}[t_1, t_2, q_1, q_2]$. We recall from Definition 1.1.18 that the join of two ideals is defined only when the ideals live in the same polynomial ring, but $\hat{\phi}$ and ϕ are maps on different polynomial rings. This is easily remedied by considering the vanishing ideals of the Mallows model to be the kernel of a map analogous to that of the one introduced in Theorem 2.1.1 but in higher dimension. Specifically, let $\phi_1.\phi_2$ be the maps

$$\begin{aligned}
\phi_1 : \mathbb{R}[p_{i,j} \mid (i,j) \in \mathscr{G}(\kappa_1, \kappa_2)] &\longrightarrow \mathbb{R}[t_1, t_2, q_1, q_2] \\
p_{i,j} &\mapsto t_1 q_1^i \\
\phi_2 : \mathbb{R}[p_{i,j} \mid (i,j) \in \mathscr{G}(\kappa_1, \kappa_2)] &\longrightarrow \mathbb{R}[t_1, t_2, q_1, q_2] \\
p_{i,j} &\mapsto t_2 q_2^j
\end{aligned}$$

each of which describes the vanishing ideal of a Mallows model where now $\ker(\phi_1), \ker(\phi_2) \subset \mathbb{R}[p_{i,j} \mid (i,j) \in \mathscr{G}(\kappa_1, \kappa_2)]$ are in the same polynomial ring, allowing us to consider the join of the two ideals.

We conjecture that the reduced Gröbner basis of the vanishing ideal of the mixture of two Mallows models is generated entirely by degree 1, degree 2, and degree 3 polynomials (for $n \geq 4$). As such, in this section, we will develop the theorems to characterize the degree 1, degree 2, and degree 3 parts of the join of two ideals. Results of the computations of some of these vanishing ideals can be found in Section 2.6.

We will be looking at the joins of varieties in our analysis of the ideals generated from two separate maps. Because the ideals (as we will show in Section 2.5) are homogeneous, the join of the varieties will be the variety of the join of the ideals (see [38]). That being said, it will be useful for us to first examine the properties of joins of general homogeneous ideals. We start with the following definition:

Definition 2.4.1. Let J be a homogeneous ideal. Define

$$(J)_k = \{f \in J \mid f \text{ is homogeneous with } \deg(f) = k\} .$$

Our goal is to characterize the degree one and degree two parts of the join of two ideals. Using Definition 1.1.18, the join of two ideals $I, J \subset \mathbb{K}[\mathbf{x}] = \mathbb{K}[x_1, \ldots, x_t]$ is

$$I * J = \Big(I(\mathbf{r}) + J(\mathbf{s}) + \langle x_j - s_j - r_j \mid 1 \leq j \leq t\rangle\Big) \cap \mathbb{K}[\mathbf{x}] .$$

Now let us consider the degree one part of the join of two ideals $(I * J)$.

Lemma 2.4.2. *Let $I, J \subset \mathbb{K}[x_1, \ldots, x_t]$ be homogeneous ideals. Then*

$$(I * J)_1 = I_1 \cap J_1 .$$

Proof. We want to consider

$$(I * J)_1 = \Big(\Big(I(\mathbf{r}) + J(\mathbf{s}) + \langle x_j - s_j - r_j \mid 1 \leq j \leq t\rangle\Big) \cap \mathbb{K}[\mathbf{x}]\Big)_1 .$$

The degree one part of the join will be exactly the part of the sum of the three ideals which is linear after intersection with $\mathbb{K}[\mathbf{x}] = \mathbb{K}[x_1, \ldots, x_t]$. We need the sum to be linear in $\mathbf{x}$, which means whatever comes from $\langle x_j - s_j - r_j \mid 1 \leq j \leq t\rangle$ will be a linear polynomial. We also need it to be a polynomial in just the $\mathbf{x}$ terms, which means that any polynomial with some r_i or some s_j in any term of the polynomial will not be in the degree one part of the join. The only way for this to happen is for all of the r_i and s_j to cancel out within the sum. Now suppose the sum looks like $f(\mathbf{r}) + g(\mathbf{s}) + h(\mathbf{x}, \mathbf{r}, \mathbf{s})$ where $f \in I$, $g \in J$, and $h \in \langle x_j - s_j - r_j \mid 1 \leq j \leq t\rangle$ and suppose furthermore that the sum contains no r_i and no s_j. Since $g(\mathbf{s})$ cannot contain any terms r_i, this means that if the sum of the r_i terms is

zero, we have that if $f = \sum_{i=1}^{t} b_i r_i$, then we see that

$$h(\mathbf{x}, \mathbf{r}, \mathbf{s}) = \sum_{i=1}^{t} b_i (x_i - r_i - s_i) \ .$$

But now, we also assumed that the sum of the s_j is zero, so we have that if $g = \sum_{j=1}^{t} c_j s_j$, this forces

$$h(\mathbf{x}, \mathbf{r}, \mathbf{s}) = \sum_{i=j}^{t} c_j (x_j - r_j - s_j)$$

by similar constraints. This is only possible if $b_i = c_i$ for all $1 \leq i \leq t$. If this is the case, then $f(\mathbf{x}) = g(\mathbf{x})$ and then $f \in I_1 \cap J_1$ as it is a polynomial with degree one in both I and J.

The reverse implication holds as well: if there is a linear polynomial $f \in I_1 \cap J_1$, then it is clear that $f \in (I * J)_1$ as it can be expressed as $f(\mathbf{r}) + f(\mathbf{s}) + f(\mathbf{x} - \mathbf{r} - \mathbf{s})$ where

$$f(\mathbf{x} - \mathbf{r} - \mathbf{s}) = f(x_1 - r_1 - s_1, \ldots, x_t - r_t - s_t) \ .$$

Thus, we have that $(I * J)_1 = I_1 \cap J_1$. □

To characterize the degree 2 part of the join of two ideals we will make use of the polarization of a polynomial, which will be defined as follows:

Definition 2.4.3. Let $f \in \mathbb{K}[\mathbf{x}]$ be a homogeneous polynomial of degree d where $\mathbf{x} = (x_1, \ldots, x_n)$. For each $i = 1, \ldots, d$ introduce a new set of variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ and variables $\mathbf{t} = (t_1, \ldots, t_d)$. The *polarization* of f, denoted $\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ is the coefficient of $\mathbf{t}^{\mathbf{1}}$ in the expansion of $f(t_1\mathbf{x}_1 + \cdots + t_d\mathbf{x}_d)$ as a polynomial in $\mathbf{t}$.

Example 2.4.4. Consider the polynomial $f(x, y, z) = x^2 + yz$ where $f(x, y, z) \in \mathbb{K}[x, y, z]$. Let $\mathbf{x}_1 = (x_1, y_1, z_1)$ and $\mathbf{x}_2 = (x_2, y_2, z_2)$. Then the polarization of f, $\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2)$ is the coefficient of $t_1 t_2$ in the expression $f(t_1 x_1 + t_2 x_2,\ t_1 y_1 + t_2 y_2,\ t_1 z_1 + t_2 z_2)$. We see

$$\begin{aligned} f(t_1\mathbf{x}_1 + t_2\mathbf{x}_2) &= (t_1 x_1 + t_2 x_2)^2 + (t_1 y_1 + t_2 y_2)(t_1 z_1 + t_2 z_2) \\ &= t_1^2 x_1^2 + 2 t_1 t_2 x_1 x_2 + t_2^2 x_2^2 + t_1^2 y_1 z_1 + t_1 t_2 y_1 z_2 + t_1 t_2 y_2 z_1 + t_2^2 y_2 z_2 \end{aligned}$$

and therefore $\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2) = 2x_1 x_2 + y_1 z_2 + y_2 z_1$.

Using the concept of a polarization, we have the following theorem:

Theorem 2.4.5. *Let f be a homogenous polynomial and $I, J \in \mathbb{K}[x_1, \ldots, x_n]$ homogeneous ideals. Then $f \in (I * J)_2$ if and only if the following three conditions are true:*

1. $f \in I$
2. $f \in J$
3. $\mathbf{f}(\mathbf{r}, \mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$

where $\mathbf{f}(\mathbf{r}, \mathbf{s})$ is the polarization of f and $I(\mathbf{r})$ is the ideal I after replacing the variables x_i with r_i and similarly $J(\mathbf{s})$ is the ideal J after replacing the variables x_j with s_j.

Proof. We know $(I * J) = \Big(I(\mathbf{r}) + J(\mathbf{s}) + \langle x_j - s_j - r_j \mid 1 \leq j \leq n\rangle\Big) \cap \mathbb{K}[\mathbf{x}]$. If we assume $f(\mathbf{x}) \in I * J$, we want to consider

$$f(\mathbf{r} + \mathbf{s}) = f(r_1 + s_1, r_2 + s_2, \ldots, r_n + s_n)$$

as $f \in I * J \Leftrightarrow f(\mathbf{r} + \mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$. From Lemma 2.5 of [38], we have:

$$f(\mathbf{r} + \mathbf{s}) = \sum_{|\beta|=2} \frac{1}{\beta!} \mathbf{f}(\mathbf{r}^{\beta_1}, \mathbf{s}^{\beta_2})$$

where $\beta! = \beta_1! \cdot \beta_2!$ and $\mathbf{f}$ is the polarization of of f. Then

$$f(\mathbf{r} + \mathbf{s}) = \frac{1}{2}\mathbf{f}(\mathbf{r}, \mathbf{r}) + \mathbf{f}(\mathbf{r}, \mathbf{s}) + \frac{1}{2}\mathbf{f}(\mathbf{s}, \mathbf{s})$$

According to [38], we know $\mathbf{f}(\mathbf{r}, \mathbf{r}) = 2! f(\mathbf{r})$, and similarly for $\mathbf{f}(\mathbf{s}, \mathbf{s})$. So we have

$$f(\mathbf{r} + \mathbf{s}) = \frac{1}{2}\mathbf{f}(\mathbf{r}, \mathbf{r}) + \mathbf{f}(\mathbf{r}, \mathbf{s}) + \frac{1}{2}\mathbf{f}(\mathbf{s}, \mathbf{s}) = f(\mathbf{r}) + \mathbf{f}(\mathbf{r}, \mathbf{s}) + f(\mathbf{s}) \ .$$

Clearly, $f(\mathbf{r}) + f(\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$ implies $f \in I$ and $f \in J$. Then we know $f(\mathbf{r} + \mathbf{s}) \in I * J$ if $\mathbf{f}(\mathbf{r}, \mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$. Thus, whenever $f \in I * J$, $f \in I \cap J$ and $\mathbf{f}(\mathbf{r}, \mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$.

Since $I(\mathbf{r}) + J(\mathbf{s})$ is bihomogeneous, $g(\mathbf{r}, \mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$ if and only if all the bihomogeneous pieces of $g(\mathbf{r}, \mathbf{s})$ are in $I(\mathbf{r}) + J(\mathbf{s})$. Since all the bihomogenous pieces of $f(\mathbf{r} + \mathbf{s})$ are in $I(\mathbf{r}) + J(\mathbf{s})$, then the theorem holds. □

Without too much effort, we can construct a very similar proof for the degree three case.

Theorem 2.4.6. *Given a homogenous polynomial f and homogeneous ideals $I, J \in \mathbb{K}[x_1, \ldots, x_n]$. Then $f \in (I * J)_3$ if and only if the following conditions are true:*

1. $f \in I$

2. $f \in J$

3. $\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$

4. $\mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$

where $\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}), \mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s})$ *are the polarizations of* f *and* $I(\mathbf{r})$, $J(\mathbf{s})$ *are defined as before.*

Proof. As we did in Theorem 2.4.5, we will consider $f(\mathbf{r}+\mathbf{s})$. Recall in the proof of Theorem 2.4.5 we saw that $f \in I * J$ if and only if $f(\mathbf{r}+\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$. Again, from Lemma 2.5 of [38], we have:

$$f(\mathbf{r}+\mathbf{s}) = \sum_{|\beta|=2} \frac{1}{\beta!} \mathbf{f}(\mathbf{r}^{\beta_1}, \mathbf{s}^{\beta_2})$$

where $\beta! = \beta_1! \cdot \beta_2!$ and $\mathbf{f}$ is the polarization of of f. Know that f has degree 3, we can expand this to get

$$\begin{aligned} f(\mathbf{r}+\mathbf{s}) &= \frac{1}{6}\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{r}) + \frac{1}{2}\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}) + \frac{1}{2}\mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s}) + \frac{1}{6}\mathbf{f}(\mathbf{s},\mathbf{s},\mathbf{s}) \\ &= f(\mathbf{r}) + f(\mathbf{s}) + \frac{1}{2}\Big(\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}) + \mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s})\Big) \end{aligned}$$

As previously stated, $I(\mathbf{r}) + J(\mathbf{s})$ is bihomogenous, so $f \in I(\mathbf{r}) + J(\mathbf{s})$ if and only if all the bihomogeneous parts of f are in $I(\mathbf{r}) + J(\mathbf{s})$. Thus, we know that if $f \in (I * J)_3$, then $f \in I$ and $f \in J$ and $\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}), \mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$. The reverse implication is straight forward, so we have that $f \in (I * J)_3$ if and only if $f \in I$ and $f \in J$ and $\mathbf{f}(\mathbf{r},\mathbf{r},\mathbf{s}), \mathbf{f}(\mathbf{r},\mathbf{s},\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$. □

We can obtain a more specific result for degree 2 part of the join of homogeneous ideals. We restrict ourselves to ideals whose linear part is generated by monomials, i.e. the linear part is simply generated by single variables. After the proof, we will explain why this "restriction" is easily circumnavigated.

Theorem 2.4.7. *Let* $I, J \in \mathbb{K}[x_1, \ldots, x_n]$ *be homogeneous ideals such that* $(I)_1 = \langle x_{i_1}, \ldots, x_{i_t} \rangle$ *and* $(J)_1 = \langle x_{j_1}, \ldots, x_{j_\ell} \rangle$ *where* $\{x_{i_1}, \ldots, x_{i_t}\} \cap \{x_{j_1}, \ldots, x_{i_\ell}\} = \emptyset$. *Then the degree 2 part of the join is given by the formula*

$$(I * J)_2 = \Big((I)_2 \cap \mathbb{K}[x_{j_1}, \ldots, x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1}, \ldots, x_{i_t}]\Big) .$$

Proof. Let f be a polynomial with $\deg(f) = 2$. We know

$$f \in \Big((I)_2 \cap \mathbb{K}[x_{j_1}, \ldots, x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1}, \ldots, x_{i_t}]\Big)$$

if and only if $f \in I$ and $f \in J$ (as the first intersection requires $f \in I$ and $f \in \mathbb{K}[x_{j_1}, \ldots, x_{j_\ell}]$ which implies that $f \in J$ since $(J)_1 = \langle x_{j_1}, \ldots, x_{j_\ell} \rangle$). Similarly, the second intersection in the sum requires $f \in J$ and

$f \in I$. If we knew that $f(\mathbf{r},\mathbf{s}) \in I(\mathbf{r}) + J(\mathbf{s})$, then by Theorem 2.4.5, $f \in (I * J)_2$. Then all that needs to be shown is that $f \in \Big((I)_2 \cap \mathbb{K}[x_{j_1},\dots,x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1},\dots,x_{i_t}]\Big)$ implies that $\mathbf{f}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) \in I(\mathbf{x}^{(1)}) + J(\mathbf{x}^{(2)})$ where we have new sets of variables $\mathbf{x}^{(1)} = (x_1^{(1)},\dots,x_n^{(1)})$ and $\mathbf{x}^{(2)} = (x_1^{(2)},\dots,x_n^{(2)})$ in the polarization. Note that we use the variables $\mathbf{x}^{(1)},\mathbf{x}^{(2)}$ (as opposed to $\mathbf{r},\mathbf{s}$) to highlight these variables relation to the variables in our polynomial ring $\mathbb{K}[x_1,\dots,x_n]$.

Given the disjoint sets $\{x_{i_1},\dots,x_{i_t}\}$ and $\{x_{j_1},\dots,x_{i_\ell}\}$, let $\{x_{k_1},\dots,x_{k_p}\}$ be the pairwise disjoint subset of $\{x_1,\dots,x_n\}$ such that:

$$\{x_{i_1},\dots,x_{i_m}\} \cup \{x_{j_1},\dots,x_{i_\ell}\} \cup \{x_{k_1},\dots,x_{k_p}\} = \{x_1,\dots x_n\} \ .$$

We want to consider $f(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$, so let $\mathbf{x}^{(1)}$ be the first set of variables and $\mathbf{x}^{(2)}$ the second as described in Lemma 2.6 of [38].

The polynomials $f \in (I * J)_2$ will have a polarization of the form

$$\mathbf{f}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) = \sum_{i,j} c_{i,j}(x_i^{(1)}x_j^{(2)} + x_j^{(1)}x_i^{(2)})$$

because the polarization is symmetric and the $\mathbf{f}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ is bihomogeneous with bi-degree $(1,1)$. Because $\mathbf{f}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ must have bi-degree $(1,1)$ and the i,j can only appear if $i,j \in \{i_1,\dots,i_m,j_1,\dots,j_\ell\}$, we see that none of the terms containing any of $\{x_{k_1},\dots,x_{k_p}\}$ can appear, so we have already filled in the corresponding columns and rows with zeroes. We consider the coefficient matrix of $f(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$:

$$\begin{array}{c} \\ x_{i_1}^{(1)} \\ \vdots \\ x_{i_m}^{(1)} \\ x_{j_1}^{(1)} \\ \vdots \\ x_{j_\ell}^{(1)} \\ x_{k_1}^{(1)} \\ \vdots \\ x_{k_p}^{(1)} \end{array} \begin{array}{c} \begin{array}{ccc|ccc|ccc} x_{i_1}^{(2)} & \dots & x_{i_m}^{(2)} & x_{j_1}^{(2)} & \dots & x_{j_\ell}^{(2)} & x_{k_1}^{(2)} & \dots & x_{k_p}^{(2)} \end{array} \\ \left(\begin{array}{ccc|ccc|ccc} & & & & & & & & \\ & & & & & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & & & & & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & 0 & & & 0 & & & 0 & \\ & & & & & & & & \end{array}\right) \end{array}.$$

Now, if $f \in I$ and $f \in J$ where $\mathbf{f}$ is bihomogeneous of bi-degree $(1,1)$, we see that every monomial term of $\mathbf{f}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ must contain either a coefficient from $\{x_{i_1},\dots,x_{i_m}\}$ or $\{x_{j_1},\dots,x_{i_\ell}\}$. Then we get the

coefficient matrix

$$\begin{array}{c} \\ x^{(1)}_{i_1} \\ \vdots \\ x^{(1)}_{i_m} \\ \hline x^{(1)}_{j_1} \\ \vdots \\ x^{(1)}_{j_\ell} \\ \hline x^{(1)}_{k_1} \\ \vdots \\ x^{(1)}_{k_p} \end{array} \begin{array}{c} \begin{array}{ccc|ccc|ccc} x^{(2)}_{i_1} & \dots & x^{(2)}_{i_m} & x^{(2)}_{j_1} & \dots & x^{(2)}_{j_\ell} & x^{(2)}_{k_1} & \dots & x^{(2)}_{k_p} \end{array} \\ \left(\begin{array}{ccc|ccc|ccc} & & & & & & & & \\ & & & & & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & 0 & & & & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & 0 & & & 0 & & & 0 & \\ & & & & & & & & \end{array}\right) \end{array}.$$

Finally, the polarization is symmetric, which forces the coefficient matrix to be symmetric as well, leaving us with

$$\begin{array}{c} \\ x^{(1)}_{i_1} \\ \vdots \\ x^{(1)}_{i_m} \\ \hline x^{(1)}_{j_1} \\ \vdots \\ x^{(1)}_{j_\ell} \\ \hline x^{(1)}_{k_1} \\ \vdots \\ x^{(1)}_{k_p} \end{array} \begin{array}{c} \begin{array}{ccc|ccc|ccc} x^{(2)}_{i_1} & \dots & x^{(2)}_{i_m} & x^{(2)}_{j_1} & \dots & x^{(2)}_{j_\ell} & x^{(2)}_{k_1} & \dots & x^{(2)}_{k_p} \end{array} \\ \left(\begin{array}{ccc|ccc|ccc} & & & & & & & & \\ & & & & 0 & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & 0 & & & & & & 0 & \\ & & & & & & & & \\ \hline & & & & & & & & \\ & 0 & & & 0 & & & 0 & \\ & & & & & & & & \end{array}\right) \end{array}.$$

Notice that the matrix forces that for every monomial, either both terms come from the variables $\{x_{i_1}, \dots, x_{i_m}\}$ or both terms come from the variables $\{x_{j_1}, \dots, x_{i_\ell}\}$. Now when we consider

$$\Big((I)_2 \cap \mathbb{K}[x_{j_1}, \dots, x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1}, \dots, x_{i_m}]\Big)$$

we see that a polynomial is in this sum if and only if both terms come from the variables $\{x_{i_1}, \dots, x_{i_m}\}$ or both terms come from the variables $\{x_{j_1}, \dots, x_{i_\ell}\}$. Suppose they did not; then we would have

a monomial with one term from the variables $\{x_{i_1},\dots,x_{i_m}\}$ and one from $\{x_{j_1},\dots,x_{i_\ell}\}$. But this is impossible for a single monomial as the intersection forces each monomial that appears to lie either completely in $\mathbb{K}[x_{i_1},\dots,x_{i_m}]$ or completely in $\mathbb{K}[x_{j_1},\dots,x_{i_\ell}]$. Thus, we have that

$$f \in \Big((I)_2 \cap \mathbb{K}[x_{j_1},\dots,x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1},\dots,x_{i_m}]\Big)$$

if and only if the coefficient matrix of $\mathbf{f}(\mathbf{r}+\mathbf{s})$ is of the form:

$$\begin{array}{c|c}
 & \begin{array}{ccc|ccc|ccc} x^{(2)}_{i_1} & \dots & x^{(2)}_{i_m} & x^{(2)}_{j_1} & \dots & x^{(2)}_{j_\ell} & x^{(2)}_{k_1} & \dots & x^{(2)}_{k_p} \end{array} \\
\begin{array}{c} x^{(1)}_{i_1} \\ \vdots \\ x^{(1)}_{i_m} \\ \hline x^{(1)}_{j_1} \\ \vdots \\ x^{(1)}_{j_\ell} \\ \hline x^{(1)}_{k_1} \\ \vdots \\ x^{(1)}_{k_p} \end{array} &
\left(\begin{array}{ccc|ccc|ccc}
 & & & & 0 & & & 0 & \\
 & & & & & & & & \\
\hline
 & 0 & & & & & & 0 & \\
 & & & & & & & & \\
\hline
 & 0 & & & 0 & & & 0 & \\
 & & & & & & & &
\end{array}\right)
\end{array}.$$

Since the coefficient matrix in this case is the same as in the case as when $f \in (I*J)_2$, we know that

$$f \in (I*J)_2 \quad \Longleftrightarrow \quad f \in \Big((I)_2 \cap \mathbb{K}[x_{j_1},\dots,x_{j_\ell}]\Big) + \Big((J)_2 \cap \mathbb{K}[x_{i_1},\dots,x_{i_t}]\Big)\ .$$

□

While this theorem is stated in a way such that the ideals I, J must be homogeneous with their degree 1 parts generated by disjoint monomials, it is possible to use Theorem 2.4.7 to help characterize the degree 2 part of the join of any two homogeneous ideals. The degree 1 part of ideals can always be written as monomials after a change of variables. It is not true that doing this for two ideals (where both are written using the same change of variables) ensures that the degree 1 part of the two ideals will have an empty intersection. However, we can simply consider a subset of the variables which do not overlap, knowing that the monomials which do overlap will then be identical. Thus, the monomials (or variables) which are in the intersection of the degree 1 part of the two ideals are removed and considered separately, allowing us to apply Theorem 2.4.7 twice and thereby characterizing the degree 2 part of the join of any two homogeneous ideals.

Knowing these properties about the join of two homogeneous ideals will allow us to easily characterize the ideals which will be introduced in Section 2.5.

2.5 The Vanishing Ideal of the Mixture of Two Mallows Models

The vanishing ideal of the mixture of two Mallows models is an ideal in the polynomial ring with the variables $p_{i,j}$ such that $(i,j) \in \mathscr{G}(\pi,\sigma)$ for a given $\pi,\sigma \in S_n$ for some n. As we showed before, $\mathscr{G}(\pi,\sigma)$ does not depend on the specific π or σ, but on $\mathrm{d}(\pi,\sigma)$ and n. We can work in a more general environment in which this polynomial ring is a special case. Recall in Proposition 2.3.9 and Proposition 2.3.11 tell us that $\mathscr{G}(\kappa_1,\kappa_2) = \mathscr{H}(r,m)$ in the specific care where $m = \binom{n}{2}$ and $r = \mathrm{d}(\pi,\sigma)$. If we remove the restriction that $m = \binom{n}{2}$, $\mathscr{H}(r,m)$ is still a well defined set. We would like a polynomial ring with variables coming from the (i,j) pairs in this more general $\mathscr{H}(r,m)$ with the knowledge that in the special case when we can write $m = \binom{n}{2}$, we can consider this as a polynomial ring from which the vanishing ideal of our Mallows mixture model might originate. Using this as inspiration, define the ring

$$P_{m,r} = \mathbb{Q}[p_{i,j} \mid (i,j) \in \mathscr{H}(r,m)],$$

and consider the map

$$\begin{aligned} \Phi_{m,r} \ : \ P_{m,r} &\longrightarrow \mathbb{Q}[t_1,t_2,q_1,q_2] \\ p_{i,j} &\mapsto t_1 q_1^i + t_2 q_2^j \ . \end{aligned}$$

We are interested in characterizing the kernel of this map. More specifically, we would like to be able to give an exact method for determining all the generators of $\ker(\Phi_{m,r})$. In this section, we will characterize the degree 1 generators of $\ker(\Phi_{m,r})$ for general m,r. Then, we will look at the specific case where we restrict ourselves to cases where $r = 1$ and describe the degree 2 generators for $\Phi_{m,1}$.

In the case where $r = 1$, we see certain desirable properties which arise, so we look at $\Phi_{m,1}$ in closer detail. First, we look at the structure of $\mathscr{H}(r,m)$. Recall that

$$\mathscr{H}(1,m) = \left\{(i,j) \mid (i+j) \equiv 1 \bmod 2 \text{ and } 1 \leq i+j \leq (2m-1) \text{ and } |i-j| \ \leq 1\right\}.$$

Using this, we notice that all the $p_{i,j}$ in our particular $P_{m,1}$ will have either form $p_{i,i+1}$ or $p_{i+1,i}$, regardless of the size n. Define $\mathscr{I} = \ker \Phi_{m,1}$. We will show $\mathscr{I}$ has a number of nice properties and its generators are easy to characterize. First, we claim there are no linear generators of $\mathscr{I}$. Next, we will show the exact number of quadratic generators of $\mathscr{I}$ to be $2\binom{m-2}{2} = (m-2)(m-3)$. Finally, we aim to show exactly what these generators will be in the case where $r = 1$.

In order to do this, we will first define the maps:

$$\phi_{m,r} \ : \ P_{m,r} \longrightarrow \mathbb{Q}[t_1, t_2, q_1, q_2] \qquad \psi_{m,r} \ : \ P_{m,r} \longrightarrow \mathbb{Q}[t_1, t_2, q_1, q_2]$$

$$p_{i,j} \mapsto t_1 q_1^i \qquad p_{i,j} \mapsto t_2 q_2^j \ .$$

Again, we will be particularly interested in the case where $r = 1$. The kernels of these maps form an ideal. Define $I_1 = \ker(\phi_{m,1})$ and $I_2 = \ker(\psi_{m,1})$. When $r = 1$, we have that every variables of $P_{m,1}$ will be all variables of the form $p_{i,i+1}$ and $p_{i+1,i}$ where $i = 0, \ldots, m-1$. Then Theorem 2.1.1 tells us that

$$I_1 = \ker(\phi_{m,1}) = \langle p_{i,i+1} - p_{i,i-1} \ , \ p_{k,k+1}p_{j+1,j} - p_{k+1,k}p_{j,j+1} \ \mid \ 0 \le k < j \le m-1 \text{ and } 0 \le i \le m-1 \rangle$$
$$I_2 = \ker(\psi_{m,1}) = \langle p_{i+1,i} - p_{i-1,i} \ , \ p_{k,k+1}p_{j+1,j} - p_{k+1,k}p_{j,j+1} \ \mid \ 0 \le k < j \le m-1 \text{ and } 0 \le i \le m-1 \rangle \ .$$

We can characterize all linear generators of $\ker(\Phi_{m,r})$ for any r, and we need not restrict ourselves to $\ker(\Phi_{m,1})$.

Theorem 2.5.1. *The linear part of* $\ker(\Phi_{m,r})$ *is generated by polynomials of the form*

$$p_{i,j} + p_{i+2,j+2} - p_{i+2,j} - p_{i,j+2}$$

where $0 \le i, j \le m-2$ *and* $(i,j),(i,j+2),(i+2,j),(i+2,j+2) \in \mathscr{H}(r,m)$. *These equations are linearly independent and there are* $(m-r-1)(r-1)$ *of them.*

Proof. Consider any linear polynomial $f \in \ker(\Phi_{m,r})$. Then f has the form

$$f = \sum_{(i,j)\in\mathscr{H}(r,m)} c_{i,j} p_{i,j} \ .$$

First we observe that $f \in \ker(\Phi_{m,r})$ implies that $\Phi_{m,r}(f) = \sum_{(i,j)\in\mathscr{H}(r,m)} c_{i,j}(t_1 q_1^i + t_2 q_2^j) = 0$. This implies that all the $t_1 q_1^i$ terms and all the $t_2 q_2^j$ terms must sum to zero. Now consider $\text{span}\{p_{i,j} + p_{i+2,j+2} - p_{i+2,j} - p_{i,j+2} \ \mid \ 0 \le i, j \le m-2\}$ and for ease of notation denote $B_{i,j} = p_{i,j} + p_{i+2,j+2} - p_{i+2,j} - p_{i,j+2}$. It is clear that each $B_{i,j} \in \ker(\Phi_{m,r})$ for all $0 \le i, j \le m-2$, so any linear combination of the $B_{i,j}$ is also in $\ker(\Phi_{m,r})$. Now suppose $f \notin \text{span}\{B_{i,j} \ \mid \ 0 \le i, j \le m-2\}$. Then we can rewrite

$$f = \sum_{(i,j)\in\mathscr{H}(r,m)} c_{i,j} p_{ij} = \sum_{(i,j)\in\mathscr{H}(r,m)} \alpha_{i,j} B_{i,j} + \sum_{(i,j)\in\mathscr{H}(r,m)} \beta_{i,j} p_{i,j}$$

where $\sum_{(i,j)\in\mathscr{H}(r,m)} \beta_{i,j} p_{i,j}$ is not divisible by any of the $B_{i,j}$ for $0 \le i, j \le m-2$. We apply a lexicographic monomial ordering where $p_{i_1,j_1} < p_{i_2,j_2}$ is $i_1 > i_2$ or $i_1 = i_2$ and $j_1 > j_2$. Doing so, we see that

$\sum_{(i,j)\in\mathscr{H}(r,m)}\beta_{i,j}p_{i,j}$ will not have any leading terms which look like p_{i_0,j_0} where $i_0+j_0=r$ unless $i=r$ or $j=r$. If we apply $\Phi_{m,r}$ to f we get:

$$\begin{aligned}\Phi_{m,r}(f) &= \sum_{(i,j)\in\mathscr{H}(r,m)}\alpha_{i,j}\Phi_{m,r}(B_{i,j}) + \sum_{(i,j)\in\mathscr{H}(r,m)}\beta_{i,j}\Phi_{m,r}(p_{i,j}) \\ &= \sum_{(i,j)\in\mathscr{H}(r,m)} 0 + \sum_{(i,j)\in\mathscr{H}(r,m)}\beta_{i,j}(t_1q_1^i+t_2q_2^j) = 0\,.\end{aligned}$$

Again, the $t_1q_1^i$ and the $t_2q_2^j$ must sum to zero. By definition $(0,r),(r,0)\in\mathscr{H}(r,m)$. We know $\beta_{0,r}=0$ as the $\Phi_{m,r}(\beta_{0,r}p_{0,r})=\beta_{0,r}(t_1+t_2q_2^r)$ since there are no other terms which could possibly contain a $t_1q_1^0=t_1$. In fact, it can be shown that for each $0\le i_0<r$, there is only one j_0 such that p_{i_0,j_0} is not a leading term of the $B_{i,j}$. Since this is the only term which can appear in $\sum_{(i,j)\in\mathscr{H}(r,m)}\beta_{i,j}(t_1q_1^i+t_2q_2^j)$ with that particular power of q_1 and we know the q_1 terms must sum to zero, we see that $\beta_{i,j}=0$ for all j where $0\le i<r$. After knowing that these terms are zero, we can repeat the process until you get to $2m-r-1\le i+j$. That is, you can continue until you get to the i,j where either $i+j=2m-r-1$ or $2m-r=i+j$. Again, $(m,m-r),(m-r,m)\in\mathscr{H}(r,m)$ by the definition of $\mathscr{H}(r,m)$. We know that none to these terms can be a leading term of the $B_{i,j}$ due to the monomial order and bounds on i,j . We see that $p_{m,m-r}$ is the only term which can contain any q_1^m so $\beta_{m,m-r}=0$. In a similar fashion, we know $\beta_{m-r,m}=0$. Then for $p_{n-1,n-r-1}$, we see that any other $p_{n-1,j}$ would be a leading term of some $B_{m-1,j}$ for all $j<m-r-1$ such that $(m-1,j)\in\mathscr{H}(r,m)$. Since this is the case, then $p_{m-1,m-r-1}$ is the only term which maps to $\beta_{m-1,m-r-1}q_1^{m-1}$ so $\beta_{m-1,m-r-1}=0$. Similarly $\beta_{m-r-1,m-1}=0$. In this way we continue until we see that all $\beta_{i,j}=0$ which means

$$f=\sum_{(i,j)\in\mathscr{H}(r,m)}c_{i,j}p_{ij}=\sum_{(i,j)\in\mathscr{H}(r,m)}\alpha_{i,j}B_{i,j}+\sum_{(i,j)\in\mathscr{H}(r,m)}\beta_{i,j}p_{i,j}=\sum_{(i,j)\in\mathscr{H}(r,m)}\alpha_{i,j}B_{i,j}\,.$$

Thus, for all $f\in\ker(\Phi_{m,r})$ where f is linear, we have

$$f\in\operatorname{span}\{B_{i,j}\ \mid\ 0\le i,j\le n-2\ \text{ and }\ (i,j)\in\mathscr{H}(r,m)\}\,.$$

Then the linear part of $\ker(\Phi_{m,r})$ is generated by

$$\begin{gathered}\operatorname{span}\{B_{i,j}\ \mid\ 0\le i,j\le n-2\ \text{ and }\ (i,j)\in\mathscr{H}(r,m)\}= \\ \operatorname{span}\{p_{i,j}+p_{i+2,j+2}-p_{i+2,j}-p_{i,j+2}\ \mid\ 0\le i,j\le m-2\ \text{ and }\ (i,j)\in\mathscr{H}(r,m)\}\,.\end{gathered}$$

□

Corollary 2.5.2. *The ideal $\mathscr{I}=\ker(\Phi_{m,1})$ contains no linear polynomials.*

Proof. From Theorem 2.5.1 we see that in the case of $\mathscr{I} = \ker(\Phi_{m,1})$, for all $(i,j) \in \mathscr{H}(r,m)$ Theorem 2.3.7 shows $(i+2,j) \notin \mathscr{H}(r,m)$. Then we know that $p_{i,j} + p_{i+2,j+2} - p_{i+2,j} - p_{i,j+2}$ is not something which can appear as at least one of the variables would not be an element of $P_{m,1}$. Thus, the space

$$\operatorname{span}\{p_{i,j} + p_{i+2,j+2} - p_{i+2,j} - p_{i,j+2} \mid 0 \le i,j \le 2m-2 \text{ and } (i,j) \in \mathscr{H}(r,m)\} = \{\emptyset\}$$

and $\mathscr{I} = \ker(\Phi_{m,1})$ contains no linear polynomials. □

Next we look at $(I_1 * I_2)_2$. To describe $(I_1 * I_2)_2$, we wish to employ 2.4.7. To do so, we need the degree one parts of I_1, I_2 to be generated by disjoint monomials. Consider the linear change of variables $X_i = p_{i,i+1} - p_{i,i-1}$ and $Y_i = p_{i+1,i} - p_{i-1,i}$. After the change of variables, we consider

$$I_1, I_2 \subset \mathbb{K}[p_{0,1}, p_{1,0}, X_1, \dots, X_{m-1}, Y_1, \dots, Y_{m-1}]$$

We then observe that $(I_1)_1 = \langle X_i \mid 1 \le i \le m-1 \rangle$ and $(I_2)_1 = \langle Y_i \mid 1 \le i \le m-1 \rangle$ according to Theorem 2.1.1 and Corollary 2.1.2.

Because $p_{0,1}$ and $p_{1,0}$ do not appear in the linear part of either I_1 or I_2, they will not appear in $(I_1 * I_2)_2$. Following directly from the proof of Theorem 2.4.7, since I_1 and I_2 both have linear parts which are generated by a set of monomials, we can create the coefficient matrix of the prolongation for any candidate degree two polynomial $f \in (I_1 * I_2)_2$ which must have the following form:

$$\begin{array}{c|c}
 & \begin{matrix} p_{0,1}^{(2)} & p_{1,0}^{(2)} & X_1^{(2)} & \dots & X_{n-1}^{(2)} & Y_1^{(2)} & \dots & Y_{n-1}^{(2)} \end{matrix} \\
\begin{matrix} p_{0,1}^{(1)} \\ p_{1,0}^{(1)} \\ X_1^{(1)} \\ \vdots \\ X_{n-1}^{(1)} \\ Y_1^{(1)} \\ \vdots \\ Y_{n-1}^{(1)} \end{matrix} &
\left(\begin{array}{cc|ccc|ccc}
0 & 0 & & & & & & \\
0 & 0 & & 0 & & & 0 & \\
0 & 0 & & & & & & \\
0 & 0 & & & & & 0 & \\
0 & 0 & & & & & & \\
0 & 0 & & & & & & \\
0 & 0 & & 0 & & & & \\
0 & 0 & & & & & &
\end{array}\right)
\end{array}.$$

Now if we restrict ourselves to polynomials which have a good polarization, we observe that

$$\begin{aligned}
(I_1)_2 \cap \mathbb{K}[X_1, \dots, X_{m-1} Y_1, \dots, Y_{m-1}] &= \operatorname{span}_{\mathbb{K}}\{X_i X_j, Y_i Y_j - Y_{\lfloor \frac{i+j}{2} \rfloor} Y_{\lceil \frac{i+j}{2} \rceil} \mid 1 \le i,j \le m-1\} \\
(I_2)_2 \cap \mathbb{K}[X_1, \dots, X_{m-1} Y_1, \dots, Y_{m-1}] &= \operatorname{span}_{\mathbb{K}}\{Y_i Y_j, X_i X_j - X_{\lfloor \frac{i+j}{2} \rfloor} X_{\lceil \frac{i+j}{2} \rceil} \mid 1 \le i,j \le m-1\}.
\end{aligned}$$

Since $(I_1)_1$ is generated entirely by the X's and $(I_2)_1$ is generated entirely by the Y's, Theorem 2.4.7

tells us to consider the ideals

$$\big((I_1)_2 \cap \mathbb{K}[Y_1,\dots,Y_{m-1}]\big) = \operatorname{span}_{\mathbb{K}}\{Y_iY_j - Y_{\lfloor\frac{i+j}{2}\rfloor}Y_{\lceil\frac{i+j}{2}\rceil} \quad | \quad 1 \le i,j \le m-1\}$$
$$\big((I_2)_2 \cap \mathbb{K}[X_1,\dots,X_{m-1}]\big) = \operatorname{span}_{\mathbb{K}}\{X_iX_j - X_{\lfloor\frac{i+j}{2}\rfloor}X_{\lceil\frac{i+j}{2}\rceil} \quad | \quad 1 \le i,j \le m-1\}.$$

Then we have that

$$(I_1 * I_2)_2 = \operatorname{span}_{\mathbb{K}}\left\{X_iX_j - X_{\lfloor\frac{i+j}{2}\rfloor}X_{\lceil\frac{i+j}{2}\rceil}\ ,\ Y_iY_j - Y_{\lfloor\frac{i+j}{2}\rfloor}Y_{\lceil\frac{i+j}{2}\rceil}\ \mid\ 1 \le i,j \le m-1\right\}.$$

by Theorem 2.4.7. We summarize these results in the following theorem:

Theorem 2.5.3. *Let $m > 3$ and $\Phi_{m,1}$ as defined above. Then the degree two part of* $\ker(\Phi_{m,1})$ *is*

$$\left\{X_iX_j - X_{\lfloor\frac{i+j}{2}\rfloor}X_{\lceil\frac{i+j}{2}\rceil}\ ,\ Y_iY_j - Y_{\lfloor\frac{i+j}{2}\rfloor}Y_{\lceil\frac{i+j}{2}}\ \mid\ 1 \le i,j \le m-1 \text{ and } i+1 < j\right\}.$$

Furthermore, there are $(m-2)(m-3)$ polynomials of this form and they are linearly independent.

Proof. The proof of this follows directly from the computations above. To see that there are exactly $(m-2)(m-3)$ (non-trivial) polynomials of this form, consider first the number of polynomials only in X_i's. There are $\binom{m-1}{2}$ choices for grabbing a distinct X_i, X_j. However, if $j = i+1$, the polynomial $X_iX_j - X_{\lfloor\frac{i+j}{2}\rfloor}X_{\lceil\frac{i+j}{2}\rceil}$ is $X_iX_{i+1} - X_iX_{i+1}$ and is trivial. There are $m-2$ possible cases where $j = i+1$. When we subtract $\binom{m-1}{2} - (m-2)$ and find a common denominator we are left with $\dfrac{(m-2)(m-3)}{2}$ nontrivial polynomials in just the X_i's. Similarly, there will be $\dfrac{(m-2)(m-3)}{2}$ polynomials in just the Y_j's, giving us a total of $(m-2)(m-3)$ degree two polynomials which generate the degree two part of $\ker(\Phi_{m,1})$. □

2.6 Computations and Conjectures on the Number of Generators of Each Degree

In this section, we examine the computational data acquired from finding a minimal generating set for $\ker(\Phi_{m,r})$, specifically the number of generators of each degree for $\ker(\Phi_{m,r})$ for different values of m, r. We then propose conjectures regarding specific patterns observed in this data.

We saw in Theorem 2.5.1 exactly what the linear generators of $\ker(\Phi_{m,r})$ would be and there are exactly $(m-r-1)(r-1)$ of them. We saw in Theorem 2.5.3 that there will be exactly $(m-2)(m-3)$ degree two generators. It should be noted that due to symmetry, the $(i,j) \in \mathscr{H}(r,m)$ if and only if

Table 2.4 The number of degree 1, 2, 3, and 4 generators for m, r

m	r	1	2	3	4
3	0	0	0	0	0
3	1	0	0	2	2
4	0	0	0	1	0
4	1	0	2	10	0
4	2	1	2	10	0
5	0	0	0	4	0
5	1	0	6	22	0
5	2	2	6	22	0
6	0	0	0	10	0
6	1	0	12	38	0
6	2	2	12	38	0
6	3	4	12	38	0
7	0	0	0	20	0
7	1	0	20	58	0
7	2	4	20	58	0
7	3	6	20	58	0
8	0	0	0	35	0
8	1	0	30	82	0
8	2	5	30	82	0
8	3	8	30	82	0
8	4	9	30	82	0

$(m-i, j) \in \mathscr{H}(m-r, m)$. Thus, we only need to consider values of r from $0, \ldots, \left\lfloor \frac{m}{2} \right\rfloor$. Note that in the case where $r = 0$, we can say that $\ker(\Phi_{m,r})$ will be a secant ideal, and therefore the number of generators will be $\binom{m-1}{3}$ (see, for instance, [34]).

Apart from the linear generators of $\ker(\Phi_{m,r})$, we do not prove the number of generators of each degree is fixed. However, we do conjecture the number of degree 2 and degree 3 generators of $\ker(\Phi_{m,r})$ based on computational evidence. In Table 2.4, we see the number of degree 1, degree 2, degree 3, and degree 4 generators for $m = 3, \ldots, 8$ for all $r = 0, \ldots, \left\lfloor \frac{m}{2} \right\rfloor$. We notice that there are no degree 4 generators for and $m > 3$. Based on these computations, we propose the following conjectures.

Conjecture 2.6.1. *Let $r, m \in \mathbb{Z}$. Then:*

1. *If $m > 3$ is fixed and $0 < r < m$ then the number of generators of $\ker(\Phi_{m,r})$ of degree two and three does not depend on r.*

2. *If $m > 3$ and $0 < r < m$ the number of degree 2 generators of $\ker(\Phi_{m,r})$ is given by $(m-2)(m-3)$.*

For $r = 1$, this number is worked out previously in the chapter.

3. *If $m > 3$ and $0 < r < m$ the number of degree 2 generators of $\ker(\Phi_{m,r})$ is given by $2(m-1)(m-2)-2$.*

4. *If $m > 3$, there are no minimum generators of degree ≥ 4.*

CHAPTER

3

CHARACTERIZING THE BI-DISTANCE POLYNOMIAL F_τ

In Chapter 2, we introduced a statistical model which was the mixture of two Mallows models. In this model, the probability of observing a permutation depends on the distance that permutation is from the centers of the two individual Mallows models. To effectively use such a model in practice, we would need a method to count the number permutations that are a fixed bi-distance from these two centers. In this chapter, we define a generating function which counts the number of permutations that are distance i from one fixed permutation and distance j from a second fixed permutation.

As we shall see, our choice of the fixed permutations greatly affects the the form of this generating function. In Section 3.1, we introduce the generating function we will study in the chapter. We examine how this generating function will factor when it is centered around a permutation which can be written as the direct sum or the skew sum of two permutations in Section 3.2. Finally, in Section 3.3 we will look at how this generating function factors when centered around more general permutations.

3.1 f_τ as a Generating Function

In Chapter 2, we examined a mixture model based on the Mallows Model. In this model, the probability of observing a particular permutation (or ranking) was dependent on both its distance from first "center" and its distance from the second "center". We will change the notation used in Chapter 2 using π, σ to denote the centers of the two Mallows models as opposed to κ_1, κ_2. Recall in the model we described in Chapter 2, we defined the $p_{i,j}$ as the probability of observing any permutation that was distance i from the first center (π) and distance j from the second center (σ). That is, for any $\beta \in S_n$ such that $\mathrm{d}(\pi, \beta) = i$ and $\mathrm{d}(\sigma, \beta) = j$, $p_\beta = p_{i,j}$. Suppose instead that we knew the marginal probability of observing any permutation $\beta \in S_n$ which satisfies $\mathrm{d}(\pi, \beta) = i$ and $\mathrm{d}(\sigma, \beta) = j$. Because the model assigns the same probability to all such permutations, the probability of observing a specific β is the probability of observing any permutation which is distance i from π and distance j from σ. If we let

$$Q_{i,j} = \left\{ \gamma \in S_n \mid \mathrm{d}(\pi, \gamma) = i \text{ and } \mathrm{d}(\sigma, \gamma) = j \right\}$$

then we have that the marginal probability of observing a permutation which is distance i from π and distance j from σ is given by $|Q_{i,j}| \cdot p_{i,j}$.

In order to be able to use this for general i, j, we propose a generating function. Consider the following function:

$$\begin{aligned} f_\tau(t, u) &= \sum_{\gamma \in S_n} t^{\mathrm{d}(\mathrm{id}, \gamma)} u^{\mathrm{d}(\tau, \gamma)} \\ &= \sum_{\gamma \in S_n} t^{\mathrm{d}(\mathrm{id}, \gamma)} u^{\mathrm{d}(\tau\gamma^{-1}, \mathrm{id})} \\ &= \sum_{\gamma \in S_n} t^{\mathrm{inv}(\gamma)} u^{\mathrm{inv}(\tau\gamma^{-1})} \end{aligned}$$

In this function, the coefficient in front of $t^i u^j$ is exactly the number of permutations γ that are distance i from the identity permutation and distance j from the permutation τ. We can see how this relates to our two centers π, σ from our Mallows mixture model: we saw before that Kendall's tau metric is right invariant. Furthermore, it is easy to verify that a permutation has the same number of inversions as its inverse. We know S_n is generated by adjacent transpositions, so if we list the product of adjacent transpositions which make up any element of S_n, its inverse will be the product of these adjacent transpositions in the reverse order. Kendall's tau metric counts the minimum transpositions

in this product, we know that $\text{inv}(\gamma) = \text{inv}(\gamma^{-1})$ for all $\gamma \in S_n$. Using these two facts, we see that

$$t^{\text{d}(\pi,\hat{\gamma})} u^{\text{d}(\sigma,\hat{\gamma})} = t^{\text{d}(\text{id},\hat{\gamma}\pi^{-1})} u^{\text{d}(\sigma\pi^{-1},\hat{\gamma}\pi^{-1})}$$

Since we are summing up over all permutations $\hat{\gamma}$, we let $\tau = \sigma\pi^{-1}$ and $\gamma = \hat{\gamma}\pi^{-1}$ and see that this is in fact the sum we want to consider.

We saw in Proposition 1.2.8 that

$$\sum_{\gamma \in S_n} q^{\text{inv}(\gamma)} = (\mathbf{n})!$$

is the q-analogue of $n!$. This is a very well studied combinatorial object and has a nice factorization. Thus, when we let $u = 1$, our polynomial $f_\tau(t,u)$ is exactly the $(\mathbf{n})!_t$ and therefore has a nice factorization. Does this hold true when u is not fixed? If it does not hold true in general, can we find exactly when $f_\tau(t,u)$ does have a nice factorization.

As described above, this factorization would be useful for counting how many permutations are distance i from the identity and distance j from τ. We saw earlier that this generating function would be necessary to create a probability distribution using the $p_{i,j}$ described in Section 2.2. While such a factorization would certainly be important if ever we were to try and use the Mallows mixture model in practice, $f_\tau(t,u)$ is an interesting object from a combinatorial perspective. As we described in Section 2.1, in the original Mallows model, the normalizing constant is the q-analogue of $n!$. That is, if we let $\Pr(\pi)$ be probability of observing a permutation π in the original Mallows model with center κ, then $\Pr(\pi) \propto q^{\text{d}(\kappa,\pi)}$. Just as when $\Pr(\pi) \propto q^{\text{d}(\kappa,\pi)}$, the normalizing constant depends on the q-analogue of $n!$, if we were to look at a model where $\Pr(\pi) \propto t^{\text{d}(\kappa,\pi)} u^{\text{d}(\gamma,\pi)}$, the normalizing constant would depend on the polynomial $f_\tau(t,u)$ (where $\tau = \gamma\kappa^{-1}$). Though we do not examine such a model, it is certainly one that may be of interest.

Question 3.1.1. How does $f_\tau(t,u)$ factor, if at all? When is it irreducible? Are there cases where we can explicitly write the factorization of $f_\tau(t,u)$?

Because τ plays a pivotal role in the structure of $f_\tau(t,u)$, it is natural to ask whether the structure of the permutation τ plays a role in whether this polynomial has a nice factorization.

Consider the following two examples. First we consider the permutation $23451 \in S_5$. When we compute the bi-distance polynomial $f_{23451}(t,u)$ in Mathematica, we see it has the following factorization

$$f_{23451}(t,u) = (1+tu)^2(1+t^2u^2)(1+tu+t^2u^2)(t^4+t^3u+t^2u^2+tu^3+u^4)$$

In contrast, f_{3142} is irreducible and is of the form

$$f_{3142}(t,u) = t^3 + t^2u + 2t^4u + 2tu^2 + 2t^3u^2 + t^5u^2 + u^3 + 2t^2u^3 + 2t^4u^3 + t^6u^3 + tu^4 + 2t^3u^4 + 2t^5u^4 + 2t^2u^5 + t^4u^5 + t^3u^6$$

Why does the first bi-distance polynomial factor so neatly, while the second cannot be factored at all? Computations show that of all the permutations in S_4, only the permutations 2413 and 3142 have bi-distance polynomials that are irreducible. These are, in fact, the only two permutations of S_4 which are not separable. Computations suggest that any permutation that is not separable will have a very large irreducible component. To better understand the way the bi-distance polynomial factors for a given τ, we will examine the properties of the particular τ.

In the next two sections, we examine how imposing various structures on the permutation τ leads to specific factorizations of the polynomial f_τ.

3.2 Factoring f_τ when $\tau = \pi \oplus \sigma$ and $\tau = \pi \ominus \sigma$

In this section we will examine specific structure of the permutation τ which make factoring $f_\tau(t,u)$ easy. In particular, we characterize the particular kind of factorization we are interested in with the following definition:

Definition 3.2.1. Given any permutation $\tau \in S_n$, the bi-distance polynomial $f_\tau(t,u)$ of τ is called collapsable if $f_\tau(t,u)$ can be written as a product containing two or more bi-distance polynomials from symmetric groups with size strictly less than n. In other words,

$$f_\tau(t,u) = g(t,u) \cdot \prod_{i=1}^{k} f_{\pi_i}(t,u)$$

where $\pi_i \in S_{a_i}$ (with $a_i < n$ for all $i \in [k]$), $k \geq 2$ and $g(t,u) \in \mathbb{R}[t,u]$ any polynomial in t and u. We call $f_\tau(t,u)$ completely collapsable if $f_\tau(t,u)$ can be written factored as

$$f_\tau(t,u) = \left(\prod_{i=1}^{\ell_1} \binom{\mathbf{n}_i}{\mathbf{k}_i}_{tu} \right) \left(\prod_{j=1}^{\ell_2} u^{k_j(n_j-k_j)} \binom{\mathbf{n}_j}{\mathbf{k}_j}_{t/u} \right)$$

With this definition, we can restate the earlier question:

Question 3.2.2. For which $\tau \in S_n$ is $f_\tau(t,u)$ collapsable? For which τ is $f_\tau(t,u)$ not collapsable?

To answer these questions, we start with two different kinds of permutation classes, both of which act on the subsets $\{1,\ldots,k\}$ and $\{k+1,\ldots,n\}$ disjointly. We will need the following definition:

Definition 3.2.3. Let $\sigma \in S_n$ and $\pi \in S_m$ elements of the symmetric groups of sizes n, m respectively. Then we can define two operations on π and σ. Let M_π, M_σ be the permutation matrices of π, σ respectively. The direct sum of the permutations lies in S_{n+m}, $\pi \oplus \sigma \in S_{m+n}$ has permutation matrix

$$M_{\pi\oplus\sigma} = \left(\begin{array}{c|c} M_\pi & 0 \\ \hline 0 & M_\sigma \end{array}\right) .$$

The skew sum lies in S_{m+n}, $\pi \ominus \sigma \in S_{m+n}$, with permutation matrix

$$M_{\pi\ominus\sigma} = \left(\begin{array}{c|c} 0 & M_\pi \\ \hline M_\sigma & 0 \end{array}\right) .$$

Any permutation β which can be written as a string of direct sums and skew sums of the trivial permutation in S_1 is called *separable*.

Based on this definition, it is clear that any permutation that can be written as a direct sum or skew sum of permutations will act on the subsets $\{1,\dots,k\}$ and $\{k+1,\dots,n\}$ disjointly (or $\{1,\dots,n-k\}$ and $\{n-k+1,\dots,n\}$ disjointly, if it is the skew sum). That is, given $\tau \in S_n$, $\pi \in S_k$, and $\sigma \in S_{n-k}$ with $\tau = \pi \oplus \sigma$, then tau would send items $\{1,\dots,k\}$ to ranks $\{1,\dots,k\}$ and items $\{k+1,\dots,n\}$ to the rankings $\{k+1,\dots,n\}$. Similarly, if $\tau = \pi \ominus \sigma$, τ would send items $\{1,\dots,k\}$ to ranks $\{n-k+1,\dots,n\}$ and items $\{k+1,\dots,n\}$ to the rankings $\{1,\dots,n-k\}$. Thus, if τ can be written either as a direct sum or a skew sum of two permutations, it will act disjointly on two disjoint subsets of $[n]$.

In order to make use of this disjoint action of τ, we will define a map to split any permutation into two permutations, one of size k and one of size $n-k$. To ensure we have a bijection, we will also need to include a multiset $\{1^k, 2^{n-k}\}$. We introduce a map from the symmetric group of size n to the Cartesian product of the symmetric group of size k (with $1 < k < n$), the symmetric group of size $n-k$, and the multiset with k 1's and $n-k$ 2's. This map will be very similar to the map defined in Stanley's proof of Proposition 1.7.1 in [39]. Define the map

$$\begin{aligned} \phi_{k,n-k} : S_n &\longrightarrow S_k \times S_{n-k} \times \mathfrak{S}_{\{1^k,2^{n-k}\}} \\ \tau &\longrightarrow (\tau_1, \tau_2, \tau_3) \end{aligned}$$

$$\tau = w_1 w_2 \cdots w_n$$
$$\tau_1 = w_{\ell_1} \cdots w_{\ell_k} \text{ such that } w_{\ell_r} \le k \text{ for } r \in [k] \text{ and } \ell_i \le \ell_{i+1} \text{ for } i \in [k-1]$$
$$\tau_2 = w_{j_1} \cdots w_{j_{n-k}} \text{ such that } w_{j_r} > k \text{ for } r \in [n-k] \text{ and } j_i \le j_{i+1} \text{ for } i \in [n-k-1]$$

$$\tau_3 = v_1 \cdots v_n \text{ where } v_i = \begin{cases} 1 & \text{if } \ w_i \leq k \\ 2 & \text{if } \ w_i > k \end{cases}$$

We will refer to $\phi_{k,n-k}$ simply as ϕ_k when it is clear that its domain is the symmetric group of size n. We see that τ_1 will be the permutation with permutation pattern of the first k elements of $[n]$ (i.e. the permutation pattern of numbers 1 through k in τ if we write τ in one line notation), τ_2 will be the permutation which has permutation pattern of the last $n-k$ elements of $[n]$ (the permutation pattern of numbers $k+1$ to n in τ when τ is in one line notation), and τ_{3_i} will tell you whether the ith letter in the word τ is less than or equal to k, or greater than k.

We would like to consider the mapping of a particular class of permutations τ. Consider the action of ϕ on a permutation $\tau \in S_n$ which is the direct sum of two permutations $\pi \in S_k, \sigma \in S_{n-k}$. As we mentioned above, the definition of $\tau = \pi \oplus \sigma$ tells us that τ permutes items $\{1,\ldots,k\}$ and $\{k+1,\ldots,n\}$ in disjoint manner, i.e. τ sends elements $\{1,\ldots,k\}$ to positions $\{1,\ldots,k\}$ and elements $\{k+1,\ldots,n\}$ to positions $\{k+1,\ldots,n\}$.

Then by the definition of our map ϕ_k, the permutation pattern of $1,\ldots,k$ in $\tau = \pi \oplus \sigma$ is π, as we know that τ sends elements $\{1,\ldots,k\}$ to positions $\{1,\ldots,k\}$. Similarly, the permutation pattern of $k+1,\ldots,n$ in $\tau = \pi \oplus \sigma$ is σ. Thus, when $\pi \in S_k$, $\sigma \in S_{n-k}$ and $\tau = \pi \oplus \sigma$

$$\phi(\tau) = (\pi, \sigma, 1 \cdots 12 \cdots 2)$$

We can also say something about the number of inversions of τ when it is the direct sum of two permutations. Recall from Definition 1.2.3, if $\beta = \beta_1\beta_2\cdots\beta_n \in S_n$ then

$$\text{inv}(\beta) = \#\Big\{(i,j) \in [n] \times [n] \mid i < j \text{ and } \beta_i > \beta_j\Big\}$$

This definition tells us that a pair (i,j) is an inversion of a permutation β if $i < j$ and $\beta(i) > \beta(j)$ where $\beta(i)$ is the action of the permutation on the element i. In the case where we have $\tau = w_1 \cdots w_n = \pi \oplus \sigma$, since τ must permute items $\{1,\ldots,k\}$ and $\{k+1,\ldots,n\}$ disjointly and it sends $\{1,\ldots,k\}$ to positions $\{1,\ldots,\text{k}\}$, we know

$$\text{inv}(\tau) = \text{inv}(\{w_1 \cdots w_k\}) + \text{inv}(\{w_{k+1} \cdots w_n\}) = \text{inv}(\pi) + \text{inv}(\sigma)$$

(where $\pi \in S_k$ and $\sigma \in S_{n-k}$).

Suppose we consider how ϕ_k acts on the product of two permutations $\tau\gamma$ for any $\gamma \in S_n$ with $\tau = \pi \oplus \sigma$ (with $\pi \in S_k, \sigma \in S_{n-k}$). Because we know τ will permute elements $\{\gamma(1),\ldots,\gamma(k)\}$ and $\{\gamma(k+1),\ldots,\gamma(n)\}$ disjointly, we can describe the action of ϕ_k on $\tau\gamma$.

Theorem 3.2.4. *Let $\tau,\gamma \in S_n$ where γ is any permutation in S_n and $\tau = \pi \oplus \sigma$ with $\pi \in S_k$ and $\sigma \in S_{n-k}$. Define the map ϕ_k as above. Let $\phi_k(\gamma) = (\gamma_1, \gamma_2, \gamma_3)$. Then*

$$\phi_k(\tau\gamma) = (\pi\gamma_1, \sigma\gamma_2, \gamma_3)$$

Proof. We know that τ permutes the items $\{1,\dots,k\}$ and $\{k+1,\dots,n\}$ disjointly.

Let $\gamma = v_1 v_2 \cdots v_n$. By definition we have that:
$\gamma_1 = v_{\ell_1} \cdots v_{\ell_k}$ such that $v_{\ell_r} \le k$ for $r \in [k]$ and $\ell_i \le \ell_{i+1}$ for $i \in [k-1]$
$\gamma_2 = v_{j_1} \cdots v_{j_{n-k}}$ such that $v_{j_r} > k$ for $r \in [n-k]$ and $j_i \le j_{i+1}$ for $i \in [n-k-1]$

$$\gamma_3 = u_1 \cdots u_n \text{ where } u_i = \begin{cases} 1 & \text{if } \ v_i \le k \\ 2 & \text{if } \ v_i > k \end{cases}$$

Using this, it is clear that $\gamma(i) \le k \Leftrightarrow i \in \{\ell_1,\dots,\ell_k\}$. Similarly for $\gamma(i) > k$. Furthermore, τ permuting items $\{1,\dots,k\}$ and $\{k+1,\dots,n\}$ disjointly means τ permutes the letters $\{v_{\ell_1},\dots,v_{\ell_k}\}$ and $\{v_{j_1},\dots,v_{j_{n-k}}\}$ of γ disjointly. Since $v_i = \gamma(i)$, if $v_i \le k$ then $\tau(v_i) = \tau\gamma(i) \le k$, and similarly if $v_i > k$ then $\tau(v_i) = \tau\gamma(i) > k$. Then by definition of the map ϕ_k, if $\phi_k(\tau\gamma) = (\beta_1, \beta_2, \beta_3)$, then we have just shown that $\beta_3 = \gamma_3$.

Not only do we know τ permutes $\{v_{\ell_1} \cdots v_{\ell_k}\}$ and $\{v_{j_1} \cdots v_{j_{n-k}}\}$ disjointly, we know that it applies the permutation π to $v_{\ell_1} \cdots v_{\ell_k} = \gamma_1$ and applies the permutation σ to $v_{j_1} \cdots v_{j_{n-k}} = \gamma_2$. By definition of our map ϕ_k we know that β_1 is the permutation with permutation pattern of $1,\dots,k$ in $\tau\gamma$ when $\tau\gamma$ is written in one line notation. But this must be exactly $\pi(v_{\ell_1} \cdots v_{\ell_k}) = \pi\gamma_1$ as γ will act on $1,\dots,k$ according to the permutation $v_{\ell_1} \cdots v_{\ell_k} = \gamma_1$ and then τ will permute these according to π. Thus, $\beta_1 = \pi\gamma_1$. By similar argument, $\beta_2 = \sigma\gamma_2$.

Thus, $\phi_k(\tau\gamma) = (\pi\gamma_1, \sigma\gamma_2, \gamma_3)$. □

To better understand the proof, we look at a specific example.

Example 3.2.5. Let $\tau = 3125647 = 312 \oplus 2314$. We can take any other permutation as our γ, so let $\gamma = 5361274$. We see that $\phi_{3,4}(\gamma) = (312, 2341, 2121122)$. We compute $\tau\gamma = 6273145$ and see that $\phi_{3.4}(\tau\gamma) = (231, 3412, 2121122)$. It is immediately clear that the mutliset permutation associated with $\phi_{3,4}(\gamma)$ is the same as the multiset permutation of $\phi_{3,4}(\tau\gamma)$. Furthermore, as we see in Figure 3.1, we know that this pattern will hold for any τ which can be written as the direct sum of two permutations.

We use Mathematica to compute the bi-distance polynomial

$$f_{3125647}(t,u) = (1+tu)^3(t^2+tu+u^2)^2(1+t^2u^2)(1-tu+t^2u^2)(1+tu+t^2u^2+t^3u^3+t^4u^4)$$
$$(1+tu+t^2u^2+t^3u^3+t^4u^4+t^5u^5+t^6u^6)$$

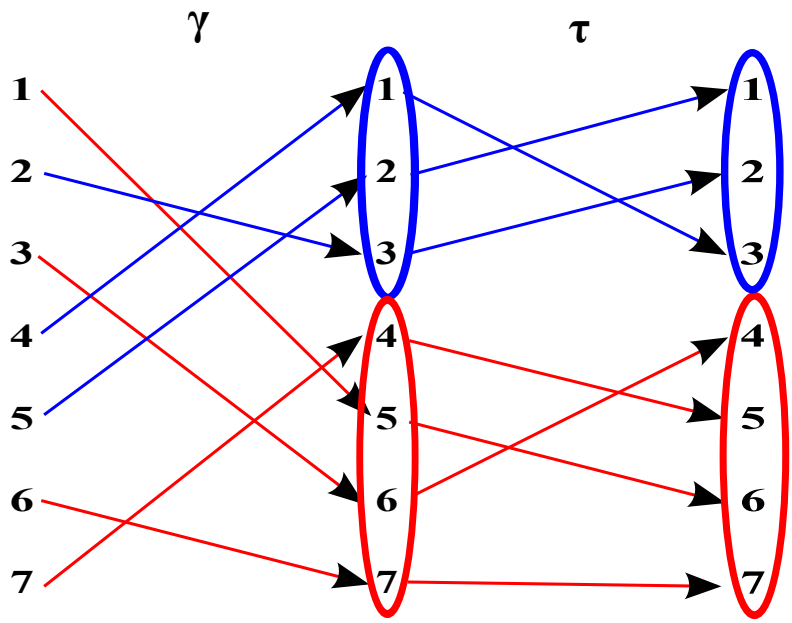

Figure 3.1 The composition of $\tau\gamma$ when τ is a direct sum, with disjoint subsets highlighted

Corollary 3.2.6. *Let $\tau,\gamma \in S_n$ where γ is any permutation in S_n and $\tau = \pi \oplus \sigma$ with $\pi \in S_k$ and $\sigma \in S_{n-k}$. Define the map ϕ_k as above. Let $\phi_k(\gamma) = (\gamma_1, \gamma_2, \gamma_3)$. Then*

$$\mathrm{inv}(\tau\gamma) = \mathrm{inv}(\pi\gamma_1) + \mathrm{inv}(\sigma\gamma_2) + \mathrm{inv}(\gamma_3)$$

Proof. In his proof of Proposition 1.7.1 in of *Enumerative Combinatorics* [39], Stanley showed that if $\phi_k(\alpha) = (\alpha_1, \alpha_2, \alpha_3)$, then $\mathrm{inv}(\alpha) = \mathrm{inv}(\alpha_1) + \mathrm{inv}(\alpha_2) + \mathrm{inv}(\alpha_3)$. Using this fact, this result follows directly from 3.2.4. □

Using this fact, we can simplify the expression for $f_\tau(t,u)$.

Corollary 3.2.7. *Let $\tau \in S_n$, $\pi \in S_k$ and $\sigma \in S_{n-k}$ with $\tau = \pi \oplus \sigma$. Then τ is collapsable. Moreover,*

$$f_\tau(t,u) = f_\pi(t,u) f_\sigma(t,u) \binom{\mathbf{n}}{\mathbf{k}}_{tu}$$

where $f_\pi(t,u)$, $f_\sigma(t,u)$ are polynomials considered in the symmetric group of size k, $n-k$ respectively and $\binom{\mathbf{n}}{\mathbf{k}}_{tu}$ is the q-analogue of $\binom{n}{k}$ evaluated at tu.

Proof. Recall

$$f_\tau(t,u) = \sum_{\gamma \in S_n} t^{\mathrm{inv}(\gamma)} u^{\mathrm{inv}(\tau\gamma)}$$

Using Corollary 3.2.7 we can rewrite this to be

$$\begin{aligned} f_\tau(t,u) &= \sum_{(\gamma_1,\gamma_2,\gamma_3)\in S_k\times S_{n-k}\times \mathfrak{S}_{\{1^k,2^{n-k}\}}} t^{\mathrm{inv}(\gamma_1)+\mathrm{inv}(\gamma_2)+\mathrm{inv}(\gamma_3)} u^{\mathrm{inv}(\pi\gamma_1)+\mathrm{inv}(\sigma\gamma_2)+\mathrm{inv}(\gamma_3)} \\ &= \Big(\sum_{\gamma_1\in S_k} t^{\mathrm{inv}(\gamma_1)} u^{\mathrm{inv}(\pi\gamma_1)}\Big)\Big(\sum_{\gamma_2\in S_{n-k}} t^{\mathrm{inv}(\gamma_2)} u^{\mathrm{inv}(\sigma\gamma_2)}\Big)\Big(\sum_{\gamma_3\in \mathfrak{S}_{\{1^k,2^{n-k}\}}} (tu)^{\mathrm{inv}(\gamma_3)}\Big) \\ &= \Big(f_\pi(t,u)\Big)\Big(f_\sigma(t,u)\Big)\binom{\mathbf{n}}{\mathbf{k}}_{tu} \end{aligned}$$

where $\binom{\mathbf{n}}{\mathbf{k}}_{tu}$ is the q-analogue of $\binom{n}{k}$ evaluated at $q = tu$. □

We have shown the polynomial $f_\tau(t,u)$ has a nice factorization when τ can be written as the direct sum of two permutations.

A similar technique is used when τ can be written as the skew sum of two permutations.

Theorem 3.2.8. *Given $\tau,\gamma\in S_n$ where $\tau=\pi\ominus\sigma$ and $\pi\in S_k$, $\sigma\in S_{n-k}$. Let $\phi_k(\gamma)=(\gamma_1,\gamma_2,\gamma_3)$. Then*

$$\phi_{n-k}(\tau\gamma)=(\sigma\gamma_2,\pi\gamma_1,\hat{\gamma_3})$$

where $\hat{\gamma_3}$ is the permutation of the multiset $\{1^{n-k},2^k\}$ obtained by reversing the roles of 1 and 2 in γ_3.

Proof. Recall that if $\tau=\pi\ominus\sigma$ with $\sigma\in S_{n-k}$ and $\pi\in S_k$, then τ will send items $\{1,\dots,k\}$ to positions $\{n-k+1,\dots,n\}$ and items $\{k+1,\dots,n\}$ to positions $\{1,\dots,n-k\}$. Thus, $\phi_{n-k}(\tau)=(\sigma,\pi,\{2,\dots,2,1,\dots,1\})$.

If we let $\phi_{n-k,k}(\tau\gamma)=(\beta_1,\beta_2,\beta_3)$, β_1 will be the permutation in S_{n-k} with permutation pattern of the elements $\{1,\dots,n-k\}$ of the composition $\tau\gamma$. We know γ_2 is the permutation of S_{n-k} with the permutation pattern of $\{k+1,\dots,n\}$ assigned by γ, and τ will send this set to the set $\{1,\dots,n-k\}$ and permute them according to σ. Thus it follows that $\beta_1=\sigma\gamma_2$. Similarly, $\beta_2=\pi\gamma_1$.

Consider $\beta_3=w_1\cdots w_n$ where $w_i=1$ if $\tau\gamma(i)\le n-k$ and $w_i=2$ if $\tau\gamma(i)>n-k$. By definition of the map ϕ_k we know that if we let $\gamma_3=u_1\cdots u_n$, then

$$u_i=\begin{cases}1 & \text{if } \gamma(i)\le k\\ 2 & \text{if } \gamma(i)>k\end{cases}.$$

Now, if $w_i=1$ we have that $\tau\gamma(i)\le n-k$. Then we know $\gamma(i)>k$, as τ will sends $\{1,\dots,k\}$ to $\{n-k+1,\dots,n\}$ and sends $\{k+1,\dots,n\}$ to $\{1,\dots,n-k\}$. This implies $u_i=2$ by the above definition.

Therefore

$$w_i = \begin{cases} 1 & \text{if } \ u_i = 2 \\ 2 & \text{if } \ u_i = 1 \end{cases}$$

and we have that $\beta_3 = \hat{\gamma_3}$ where $\hat{\gamma_3}$ is the permutation of the multiset $\{1^{n-k}, 2^k\}$ obtained by reversing the roles of 1 and 2 in γ_3. □

Again, the proof is more easily understood by looking at an example. Consider the following:

Example 3.2.9. Let $\tau = 6573241 = 213 \ominus 3241$. We can again, choose any permutation γ at random, but we will use $\gamma = 5361274$ from the previous example. We know $\phi_{3,4}(\gamma) = (312, 2341, 2121122)$ and when we compute $\tau\gamma = 2746513$, we see $\phi_{4,3}(\tau\gamma) = (2413, 321, 1212211)$. This is exactly as we described it above. As we can see in Figure 3.2, this is exactly what we would expect to happen for any such τ, γ when τ is a skew sum.

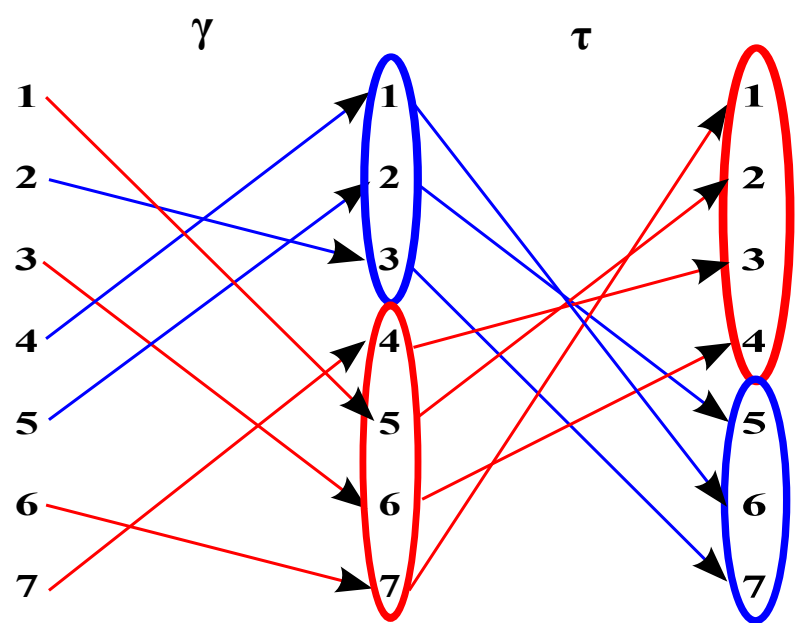

Figure 3.2 The composition of $\tau\gamma$ when τ is a skew sum, with disjoint subsets highlighted

Corollary 3.2.10. *Given $\tau, \gamma \in S_n$ where $\tau = \pi \ominus \sigma$ and $\pi \in S_k$, $\sigma \in S_{n-k}$. Let $\phi_k(\gamma) = (\gamma_1, \gamma_2, \gamma_3)$. Then*

$$\text{inv}(\tau\gamma) = \text{inv}(\sigma\gamma_2) + \text{inv}(\pi\gamma_1) + \text{inv}(\hat{\gamma_3})$$

where $\hat{\gamma_3}$ is the permutation of the multiset $\{1^{n-k}, 2^k\}$ obtained by reversing the roles of 1 and 2 in γ_3. Therefore, when τ can be written as the skew sum of permutations, τ is collapsable.

Corollary 3.2.11. *Given $\tau, \gamma \in S_n$ where $\tau = \pi \ominus \sigma$ and $\pi \in S_k$, $\sigma \in S_{n-k}$. Let $\phi_k(\gamma) = (\gamma_1, \gamma_2, \gamma_3)$. Then*

$$f_\tau(t,u) = f_\pi(t,u) f_\sigma(t,u) \left(u^{k(n-k)} \binom{\mathbf{n}}{\mathbf{k}}_{t/u} \right)$$

Proof. We know from the above theorem that $\phi_{n-k}(\tau\gamma) = (\sigma\gamma_2, \pi\gamma_1, \hat{\gamma_3})$. We first need to see the relationship between the number of inversions of γ_3 and the inversions of $\hat{\gamma_3}$. Any pair that is an inversion in γ_3 will not be an inversion in $\hat{\gamma_3}$, and vice versa. We also know that the maximum number of inversions of a permutation in $\mathfrak{S}_{\{1^k, 2^{n-k}\}}$ is $k(n-k)$. We can verify this quite quickly: we know that the permutation with the most inversions will be $2 \cdots 21 \cdots 1$. Each 2 in this permutation has exactly k 1's after it, corresponding to k inversions. If there are $n-k$ 2's and each has k inversions, there is a total of $k(n-k)$ inversions. Thus, we can say that $\mathrm{inv}(\hat{\gamma_3}) = k(n-k) - \mathrm{inv}(\gamma_3)$.

When we let

$$\begin{aligned}
f_\tau(t,u) &= \sum_{(\gamma_1,\gamma_2,\gamma_3) \in S_k \times S_{n-k} \times \mathfrak{S}_{\{1^k, 2^{n-k}\}}} t^{\mathrm{inv}(\gamma_1)+\mathrm{inv}(\gamma_2)+\mathrm{inv}(\gamma_3)} u^{\mathrm{inv}(\pi\gamma_1)+\mathrm{inv}(\sigma\gamma_2)+\mathrm{inv}(\hat{\gamma_3})} \\
&= \Big(\sum_{\gamma_1 \in S_k} t^{\mathrm{inv}(\gamma_1)} u^{\mathrm{inv}(\pi\gamma_1)} \Big) \Big(\sum_{\gamma_2 \in S_{n-k}} t^{\mathrm{inv}(\gamma_2)} u^{\mathrm{inv}(\sigma\gamma_2)} \Big) \Big(\sum_{\gamma_3 \in \mathfrak{S}_{\{1^k, 2^{n-k}\}}} t^{\mathrm{inv}(\gamma_3)} u^{\mathrm{inv}(\hat{\gamma_3})} \Big) \\
&= \Big(\sum_{\gamma_1 \in S_k} t^{\mathrm{inv}(\gamma_1)} u^{\mathrm{inv}(\pi\gamma_1)} \Big) \Big(\sum_{\gamma_2 \in S_{n-k}} t^{\mathrm{inv}(\gamma_2)} u^{\mathrm{inv}(\sigma\gamma_2)} \Big) \Big(\sum_{\gamma_3 \in \mathfrak{S}_{\{1^k, 2^{n-k}\}}} t^{\mathrm{inv}(\gamma_3)} u^{k(n-k)-\mathrm{inv}(\gamma_3)} \Big) \\
&= \Big(\sum_{\gamma_1 \in S_k} t^{\mathrm{inv}(\gamma_1)} u^{\mathrm{inv}(\pi\gamma_1)} \Big) \Big(\sum_{\gamma_2 \in S_{n-k}} t^{\mathrm{inv}(\gamma_2)} u^{\mathrm{inv}(\sigma\gamma_2)} \Big) \Big(\sum_{\gamma_3 \in \mathfrak{S}_{\{1^k, 2^{n-k}\}}} \left(\frac{t}{u}\right)^{\mathrm{inv}(\gamma_3)} u^{k(n-k)} \Big) \\
&= \Big(f_\pi(t,u) \Big) \Big(f_\sigma(t,u) \Big) \left(u^{k(n-k)} \binom{\mathbf{n}}{\mathbf{k}}_{t/u} \right)
\end{aligned}$$

□

We know that a permutation is separable if it can be written as a sequence of direct and skew sums of the trivial permutation in S_1. The following is a result of all that has been shown above

Corollary 3.2.12. *Let $\tau \in S_n$ be a separable permutation. Then $f_\tau(t,u)$ is completely collapsable.*

Proof. We know τ is separable. Then without loss of generality, $\tau = \pi_1 \oplus \pi_2$ where π_1, π_2 are in S_{n_1}, S_{n_2} respectively. Then $f_\tau(t,u) = f_{\pi_1}(t,u) f_{\pi_2}(t,u) \binom{\mathbf{n}}{\mathbf{n_1}}_{t\,u}$ by Corollary 3.2.11. But then we know that π_1, π_2 can both be written as the direct sum or the skew sum of two permutations. We can then split up

$f_{\pi_1}(t,u), f_{\pi_2}(t,u)$ into the product of two bi-distance polynomials and the appropriate remainder (depending on whether they are written as a direct sum or a skew sum). We continue in this manner until we have $\pi_{i_1}\dots\pi_{i_k}$ which are all the trivial permutation. Since $f_1(t,u)=1$, we see that

$$f_\tau(t,u)=\left(\prod_{i=1}^{k_1}\binom{\mathbf{n}_i}{\mathbf{k}_i}_{tu}\right)\left(\prod_{j=1}^{k_2}u^{k_j(n_j-k_j)}\binom{\mathbf{n}_j}{\mathbf{k}_j}_{t/u}\right)$$

and thus $f_\tau(t,u)$ is completely collapsable. □

Based on computational evidence, we propose the following conjecture.

Conjecture 3.2.13. *For any permutation $\tau\in S_n$, the bi-distance polynomial $f_\tau(t,u)$ is completely collapsable if and only if τ is a separable permutation.*

As we know, if a permutation is separable, then it is completely collapsable. We do not show the reverse implication, but rather consider the following computational evidence. We start by creating a function in Mathematica to find the bi-distance polynomial of any given permutation. When we run this function on every permutation, we see that the number of polynomials which are completely collapsable is the same as the number of permutations which are separable (that is, the number of completely collapsable bi-distance polynomials with permutations of size n is exactly $a(n)$, the large Schröder number, which is exactly the number of separable permutations in S_n [6]). This is true through S_{10}. Furthermore, it is true that, through S_8 every completely collapsable bi-distance polynomial comes from a separable permutation.

Next, we consider more general forms of permutations which permute subsets of $[n]$ in a disjoint manner.

3.3 Permutations with Contiguous Blocks

In this section, we consider permutations τ that are not necessarily a direct sum of skew sum of permutations, but can be separated into contiguous blocks. In other words, the permutation τ (when written in one line notation) can be separated into parts whose elements are a span of consecutive numbers. For instance, consider the permutation $\tau=3\,4\,5\,9\,10\,2\,1\,8\,7\,6$. We see that we can separate τ in as $\tau=345|910|21|876$. While this τ cannot be written as a series of skew and direct sums, we take advantage of its blocked structure. Notice that this particular τ will send elements $\{1,2,3\}$ to positions $\{3,4,5\}$. Similarly, we can see it sends $\{4,5\}$, $\{6,7\}$, and $\{8,9,10\}$ to positions $\{9,10\}$, $\{1,2\}$, and $\{6,7,8\}$ respectively. Thus, τ still acts disjointly on certain subsets of elements, although it does permute these subsets. Before we can take advantage of this structure, we need the following definition. We can define

a more generalize version of the ϕ we defined earlier. Let $\phi_{\alpha_1,\ldots,\alpha_k} : S_n \longrightarrow S_{a_1} \times \cdots \times S_{a_k} \times \mathfrak{S}_{\{1^{a_1},\ldots,k^{a_k}\}}$ where $\sum_{i=1}^{k} a_i = n$ be defined by $\phi_{\alpha_1,\ldots,\alpha_k}(\tau) = (\tau_1,\ldots,\tau_k,\tau_M)$ where τ_i is the permutation pattern of $\{a_1+\ldots+a_{i-1}+1,\ldots a_1+\ldots+a_{i-1}+a_i\}$ in τ for $i = 1,\ldots,k$ and τ_M is the multiset whose j^{th} element will be $(\tau_M)_j = i$ if $a_1+\ldots+a_{i-1}+1 \leq \tau(j) \leq a_1+\ldots+a_{i-1}+a_i$. This is a logical extension of the map ϕ_k defined previously and is again reminiscent of the map defined in [39].

We can apply this map to the previous example where $\tau = 3\ 4\ 5\ 9\ 10\ 2\ 1\ 8\ 7\ 6$. Consider $\phi_{2,3,3,2}(\tau)$. The first entry will be the permutation pattern of $\{1,2\}$ in τ, namely 2 1. The second entry will be 1 2 3, etc. We get that $\phi_{2,3,3,2}(\tau) = (21,123,321,12,2224411333)$. We see the multiset here is of a very particular form due to the contiguous block configuration of τ. The multiset can be any element of $\mathfrak{S}_{\{1^{a_1},\ldots,k^{a_k}\}}$ in general (as this is a bijection), and in general will not have this nice form (for instance, $\phi_{1,3,2}(263415) = (1,123,21,232213)$).

We would like to examine τ permute subsets of $[n]$ disjointly. To accomplish this, we introduce the following definition.

Definition 3.3.1. Given any $\tau \in S_n$ that satisfies $\phi_{a_1,\ldots,a_k}(\tau) = (\tau_1,\ldots,\tau_k,\tau_M)$ in $S_{a_1} \times \ldots \times S_{a_k} \times \mathfrak{S}_{\{1^{a_1},\ldots,k^{a_k}\}}$ with $\sum_{i=1}^{k} a_i = n$ where τ_M has the form $i_1 \cdots i_1 i_2 \cdots i_2 \cdots i_k \cdots i_k$. Then τ is called a *contiguous block permutation*. Furthermore, the permutation $\pi = i_1 i_2 \cdots i_k \in S_k$ is *the order permutation associated with* τ.

This characterizes permutations that cannot be written as a series of skew sums and direct sums, but still permute subsets of $[n]$ disjointly. We take advantage of this block structure in the following theorem.

Theorem 3.3.2. *Given any $\tau \in S_n$ that is a contiguous block permutation with $\phi_{a_1,\ldots,a_k}(\tau) = (\tau_1,\ldots,\tau_k,\tau_M)$ with $\sum_{i=1}^{k} a_i = n$, $\tau_M = i_1 \cdots i_1 i_2 \cdots i_2 \cdots i_k \cdots i_k$, and τ having order permutation $\pi = i_1 \cdots i_k$. Take any $\gamma \in S_n$ where $\phi_{a_{i_1},\ldots,a_{i_k}}(\gamma) = (\gamma_1,\ldots,\gamma_k,\gamma_M)$. Then*

$$\phi_{a_1,\ldots,a_k}(\tau\gamma) = (\tau_1\gamma_{\pi^{-1}(1)},\ldots,\tau_k\gamma_{\pi^{-1}(k)},\pi(\gamma_M))$$

where $\pi(\gamma_M)$ will be the multiset permutation in $\mathfrak{S}_{\{1^{a_1},\ldots,k^{a_k}\}}$ obtained by applying the permutation π to the multiset permutation γ_M (i.e., the ith entry of $\pi(\gamma_M)$ will be π applied to the ith entry of γ_M).

Proof. To see this is true, we again take advantage of the block structure of our permutation τ. Note that τ will send the elements $\{1,\ldots,a_1\}$ to the subset $\{a_{i_1}+\cdots+a_{i_{j-1}}+1,\ldots,a_{i_1}+\cdots+a_{i_j}\}$ where $j = \pi^{-1}(1)$, and so on. Thus, when looking at the permutation pattern of $\{1,\ldots,a_1\}$ in the product $\tau\gamma$, we will need to look at the elements that τ sends to items $\{1,\ldots,a_1\}$, which is exactly τ_1. But we know that τ_1 will be in the $\pi^{-1}(1)$ block of τ. Thus, we will want to find the permutation pattern of γ which

maps to the $\pi^{-1}(1)$ block of τ, which will be precisely $\gamma_{\pi^{-1}(1)}$ as defined by the map $\phi_{a_{i_1},\ldots,a_{i_k}}(\gamma)$. Then the permutation pattern of $\{1,\ldots,a_1\}$ in the permutation $\tau\gamma$ will be precisely $\tau_1\gamma_{\pi^{-1}(1)}$. Similarly for the remaining permutation patterns $\tau_i\gamma_{\pi^{-1}(i)}$.

To see that the multiset permutation of $\phi_{a_1,\ldots,a_k}(\tau\gamma)$ is in fact $\pi(\gamma_M)$, let $\beta_M = \beta_1\cdots\beta_n$ be the multiset permutation of $\phi_{a_1,\ldots,a_k}(\tau\gamma)$ and $\gamma_M = \alpha_1\cdots\alpha_n$. Then $\alpha_i = j$ implies that $a_{i_1}+\cdots+a_{i_{j-1}} < \gamma(i) \leq a_{i_1}+\cdots a_{i_j}$. This means $\gamma(i)$ is sent to the jth block of τ, so it will be sent to the $\pi(j)$ block in $\tau\gamma$. In other words, when γ_M has a $\alpha_i = j$, $\beta_i = \pi(j)$. Thus, $\pi(\alpha_i) = \beta_i$ for all i and therefore $\beta_1\cdots\beta_n = \pi(\gamma_M)$. □

We look at the following example to get a “pictorial" idea of the proof of this theorem. Because this theorem is actually a generalization of the case where τ is either a direct sum or a skew sum, we will go into more detail.

Example 3.3.3. Consider $\tau = 4\ 7\ 5\ 6\ 10\ 2\ 1\ 3\ 9\ 8$. We see that we can group $\tau = 4756 \mid 10 \mid 213 \mid 98$. Then by definition, we see that $\phi_{3,4,2,1}(\tau) = (213, 1423, 21, 1, 2222411133) = (\tau_1,\tau_2,\tau_3,\tau_4,\tau_M)$. The order permutation of associated with τ is $\pi = 2413$ Now we can take any $\gamma \in S_{10}$, so we just choose $\gamma = 6\ 9\ 1\ 4\ 10\ 8\ 2\ 5\ 7\ 3$. We see that $\phi_{a_2,a_4,a_1,a_3}(\gamma) = \phi_{4,1,3,2}(\gamma) = (1423, 1, 132, 12, 3411431231) = (\gamma_1,\gamma_2,\gamma_3,\gamma_4,\gamma_M)$. We can see that the product $\tau\gamma = 2\ 9\ 4\ 6\ 8\ 3\ 7\ 10\ 1\ 5$. Then we see that $\phi_{3,4,2,1}(\tau\gamma) = (231, 1342, 21, 1, 1322312412)$. We see that in fact $231 = \tau_1\gamma_{\pi^{-1}(1)} = \tau_1\gamma_3$ and similarly for 1342, 21, and 1. Furthermore, we see that 1322312412 is in fact $\pi(\gamma_M) = \pi(3411431231)$. This is shown in Figure 3.3.

In this example, there is only one matching of the various $\tau_i\gamma_j$ which results in defined permutation products, as $\tau_1,\ldots,\tau_4$ are four different permutations from four symmetric groups of distinct sizes and similarly for $\gamma_1,\ldots,\gamma_4$ (and a product of permutations is only defined between two permutations of the same size). Thus, there is only one way to match these two sets together which makes sense. To see this proof is true in general, consider Figure 3.4. Suppose we would like to know what the permutation pattern of $\{1,2,3\}$ in $\tau\gamma$. Because $\pi = 2413$, we know $\pi^{-1}(1) = 3$, which means that τ will send the its third contiguous block to the first contiguous block of $\tau\gamma$ (i.e., it sends $\{6,7,8\}$ to $\{1,2,3\}$ after applying the permutation $\tau_3 = 213$). Thus whatever γ sends to $\{6,7,8\}$, τ will map to $\{1,2,3\}$ after applying 213. We know what elements γ will send to $\{6,7,8\}$ and the order the will appear in. It is exactly the permutation pattern of $\{6,7,8\}$ in γ, which we know from our map $\phi_{4,1,3,2}$ is exactly γ_3. Thus, there permutation pattern of $\{1,2,3\}$ will be exactly τ_1 applied to γ_3. In general, if we want to know what is in the ith block of $\tau\gamma$, we find $\pi^{-1}(i)$ to see which block of τ is sent to the ith block of $\tau\gamma$. Then, we know that the ith block of $\tau\gamma$ will be whatever τ_i applied to whatever is sent to the ith block of τ, namely $\gamma_{\pi^{-1}(i)}$.

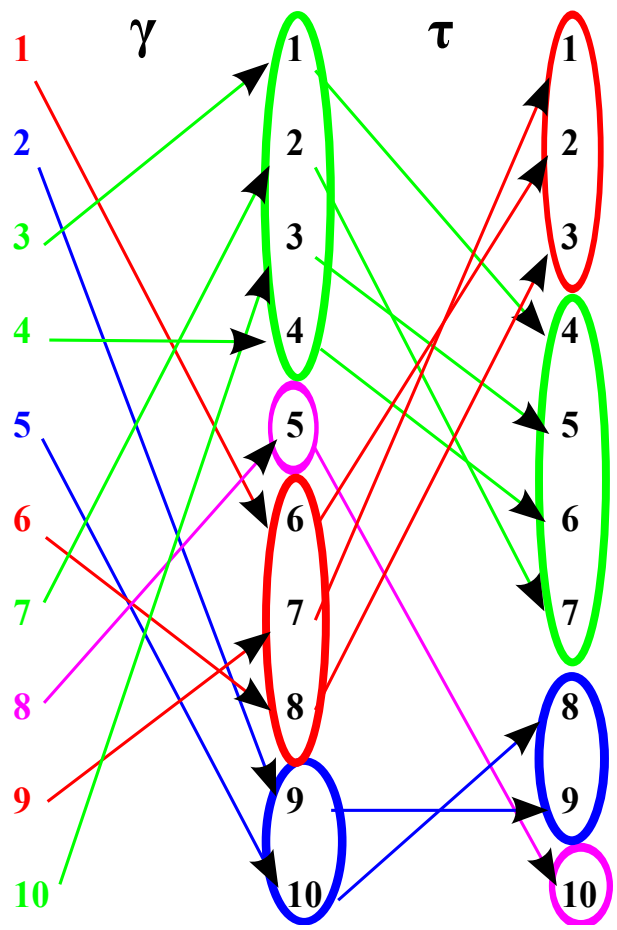

Figure 3.3 The composition of $\tau\gamma$ when τ is a permutation consisting of contiguous blocks, with disjoint subsets highlighted

As we said above, τ being a direct or skew sum actually falls into this category, and thus this theorem is a generalization of the Theorem 3.2.4 and Theorem 3.2.8. However, as this is a generalization, it does not yield a nice, closed form factorization of $f_\tau(t,u)$. Still, we can say the following.

Theorem 3.3.4. *Suppose we have that $\tau \in S_n$ is a contiguous block permutation with order permutation π and $\phi_{a_1,\ldots,a_k}(\tau) = (\tau_1,\ldots,\tau_k,\tau_M)$ where $\tau_1,\ldots\tau_k$ lie in $S_{a_1},\ldots S_{a_k}$ respectively. Then the bi-distance polynomial of τ factors*

$$f_\tau(t,u) = g(t,u)\prod_{j=1}^{k} f_{\tau_j}(t,u)$$

where $f_{\tau_j}(t,u)$ is the bi-distance polynomial of $\tau_j \in S_{a_j}$ and $g(t,u) \in \mathbb{R}[t,u]$ is some polynomial.

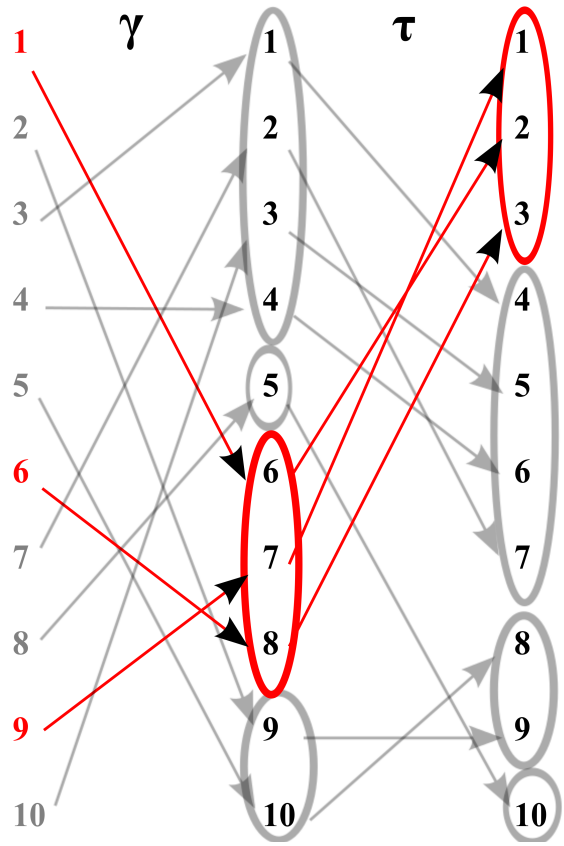

Figure 3.4 The composition of $\tau\gamma$ when τ is a permutation consisting of contiguous blocks, highlighting the permutation pattern of 123 in $\tau\gamma$

Proof. The proof is similar to the cases where τ is a direct or a skew sum. By definition we have

$$f_\tau(t,u) = \sum_{\substack{(\gamma_1,\ldots,\gamma_k,\gamma_M)\in \\ S_{a_{i_1}}\times\cdots\times S_{a_{i_1}}\times\mathfrak{S}_{\{1^{a_1},\ldots,k^{a_k}\}}}} t^{\mathrm{inv}(\gamma_1)+\cdots+\mathrm{inv}(\gamma_k)+\mathrm{inv}(\gamma_M)} u^{\mathrm{inv}(\tau_1\gamma_{\pi^{-1}(1)})+\ldots+\mathrm{inv}(\tau_k\gamma_{\pi^{-1}(k)})+\mathrm{inv}(\pi\gamma_M)}$$

Knowing that we there is some j such that $\pi^{-1}j = i$, we can rearrange this sum in a similar manner as before. We simply need to note that for any $i \in [k]$, we know that γ_i will be matched with $\tau_{\pi(i)}$ after applying $\phi_{a_1,\ldots,a_k}$ to the composition $\tau\gamma$. In other words, if we wanted to know which τ_j will act on γ_i after applying $\phi_{a_1,\ldots,a_k}$ to the composition $\tau\gamma$, it is equivalent to asking which τ_j will act on the ith contiguous block, which will be exactly $\tau_{\pi(i)}$ as we saw before. Therefore

$$
\begin{aligned}
f_\tau(t,u) &= \sum_{\substack{(\gamma_1,\dots,\gamma_k,\gamma_M)\in \\ S_{a_{i_1}}\times\cdots\times S_{a_{i_1}}\times\mathfrak{S}_{\{1^{a_1},\dots,k^{a_k}\}}}} t^{\mathrm{inv}(\gamma_1)+\cdots+\mathrm{inv}(\gamma_k)+\mathrm{inv}(\gamma_M)} u^{\mathrm{inv}(\tau_1\gamma_{\pi^{-1}(1)})+\dots+\mathrm{inv}(\tau_k\gamma_{\pi^{-1}(k)})+\mathrm{inv}(\pi\gamma_M)} \\
&= \Big(\sum_{\gamma_M\in\mathfrak{S}_{\{1^{a_1},\dots,k^{a_k}\}}} t^{\mathrm{inv}(\gamma_M)} u^{\mathrm{inv}(\pi\gamma_M)}\Big)\Big(\prod_{i=1}^{k}\sum_{\gamma_i\in S_{a_i}} t^{\mathrm{inv}(\gamma_i)} u^{\mathrm{inv}(\tau_{\pi(i)}\gamma_i)}\Big) \\
&= \Big(\sum_{\gamma_M\in\mathfrak{S}_{\{1^{a_1},\dots,k^{a_k}\}}} t^{\mathrm{inv}(\gamma_M)} u^{\mathrm{inv}(\pi\gamma_M)}\Big)\Big(\prod_{i=1}^{k} f_{\tau_{\pi(i)}}(t,u)\Big) \\
&= g(t,u)\Big(\prod_{i=1}^{k} f_{\tau_{\pi(i)}}(t,u)\Big)
\end{aligned}
$$

and because we know

$$\prod_{i=1}^{k} f_{\tau_{\pi(i)}}(t,u) = \prod_{i=1}^{k} f_{\tau_{(i)}}(t,u)$$

by the commutative nature of multiplication in the polynomial ring, we have

$$f_\tau(t,u) = g(t,u)\prod_{j=1}^{k} f_{\tau_j}(t,u)$$

□

CHAPTER

4

A THURSTONIAN MODEL FOR PARTIALLY RANKED DATA

4.1 Introduction

Thurstonian models are used to recover a ranking of items from many different incompatible rankings. Introduced by Thurstone in 1927, the Thurstonian model was originally used in studies of human psychology and cognitive science [42]. The model assumes there are n items to be ranked, such as children's drawings or handwriting samples, and someone(s) ranking these items, say a judge. The model is based on the principle that the judge(s) may not always rank things in the same order every time, but rather ranks the items according to some normal distribution based on some underlying "true" psychological construct. The Thurstonian model is still used today in psychological studies; in [40], the authors attempt to reconstruct the true order of certain events, such on the order of the presidents, based on "the wisdom of the crowds" (rankings assigned by a large group of individuals). In this paper, we apply the tenants of a Thurstonian model to biology, specifically to the mutations of the Human Immunodeficiency Virus (HIV) and prostate cancer cells.

When taking samples of a cell or virus that can have mutations at multiple sites, there are two natural groups to divide the mutations for each sample: those that have occurred and those that have

not. We wish to recover a global ordering that the mutations occur in. The global ranking we seek to recover is inherent in all Thurstonian models. The fact that we observe a partial order is different, however, than many applications of Thurstonian models. Most variations of the original Thurstonian model observe totally ranked data. We propose a way to adapt the standard Thurstonian model for use with partially ranked data.

The data we will examine will be 0-1 vectors called mutation vectors. For each observation, we have a mutation vector whose i^{th} entry will be 1 if the mutation has occurred at the time of observation and 0 if it has not yet occurred. We take many such observations from different individuals. The goal is to use this data to gain better understanding of global mutation order. By recovering global orderings, we hope to find the most likely (or unlikely) mutation orders, the mutation most likely to occur first, the total time for all mutations to occur, and any mutations which have prerequisite mutations. In studying these specific traits of the global ordering of the mutations, we could have a better idea of how far the disease has progressed and, potentially, be able to see which steps to take in order to prevent it from progressing further.

We seek to add our Thurstonian model to the many statistical models using partially ranked data are present in the literature. In [2], the authors examine discrete probability distributions that separate a set of events and prove there is both a closed form for the maximum likelihood estimate of the probability for each event occurring as well as a unique maximum likelihood poset for each such probability distribution. The authors of [3] use a Markov model which utilizes 0-1 mutation vectors to find the most likely poset dictating mutation order. The model in [3] assumes that mutation time is exponentially distributed. In [26], the authors use a more general kind of partially ranked data with a form of the Mallows model to make conclusions about the discrete probability distribution over all partial rankings. While this more general model is ideal for a general setting, we propose a new model designed specifically for datasets which are 0-1 vectors and where mutation times are assumed to be distributed normally.

In this chapter, we propose a new model and then analyze it from both a Bayesian and a frequentist perspective. In Section 4.2, we will introduce a new model we will refer to as the partially ranked Thurstonian model or the Thurstonian mutation model which we use throughout the remainder of the chapter. Next, we examine how we can use a Bayesian technique called a Gibbs sampler within this new model. In Section 4.3, we propose a method for parameter estimation of the parameters of the Thurstonian model for partial rankings by means of a Gibbs sampler. In the Section 4.4, we derive a method to compute the maximum likelihood estimate for the parameters of the Thurstonian mutation model. In Section 4.5 we use these techniques to analyze two different datasets, one with mutation vectors for HIV cells and one with mutation vectors of prostate cancer cells. For each of these datasets, we use the Gibbs sampler technique and maximum likelihood estimation–introduced in

Sections 4.3 and 4.4 respectively–to analyze each of these datasets in light of this model. By analyzing the datasets with these techniques, we hope to learn about global trends for mutation order as well as the mutation order which is most likely to occur and the relative timeframes each mutation might occur.

4.2 Partially Ranked Thurstonian Model

Thurstonian models, as mentioned above, traditionally deal with totally ranked data [42]. Because the data we seek to examine is partially ranked, we adapt the traditional Thurstonian model to use partially ranked data to recover a global ordering on all mutations while keeping our latent random variables normally distributed.

We propose the following Thurstonian model for partially ranked data, which we dub the partially ranked Thurstonian model, to describe our system of mutations. Let $X = (X_1, \ldots, X_k)$ be a k-vector of hidden random variables (we think of k as the number of mutations tested for). We will take N samples of our variable X, each of the form $X^{(i)} = (X_1^{(i)}, \ldots, X_k^{(i)})$ where $i = 1, \ldots, N$. Let μ be a k-vector where μ_j is the expected value of X_j and σ a k-vector where σ_j the standard deviation of X_j. The distribution of X_j will be $X_j \sim \mathrm{N}(\mu_j, \sigma_j^2)$, making $X^{(i)}$ distributed i.i.d.

Next, let T be a Gaussian random variable with $T \sim \mathrm{N}(\mu_t, \sigma_t)$. Later, we assume $\mu_t = 0$ and $\sigma_t = 1$, allowing us to circumvent identifiability issues. We will take N observations $T^{(1)}, \ldots, T^{(N)}$. Then $T^{(i)}$ is distributed i.i.d. Furthermore, when considered as a vector, $(X^{(i)}, T^{(i)})$ will be a hidden variable that is independently identically distributed (i.d.d.).

Our observed variable will be $Y^{(i)}$, a k-vector of values determined by $(X^{(i)}, T^{(i)})$ by

$$Y_j^{(i)} = \begin{cases} 0 & \text{if } X_j^{(i)} > T^{(i)} \\ 1 & \text{if } X_j^{(i)} < T^{(i)} \end{cases}.$$

In the case of our own data, $T^{(i)}$ serves as a random cutoff time in sample i. We observe the entries of $Y^{(i)}$ as either having occurred, in which case that entry is a 1, or having not occurred by time $T^{(i)}$, in which case that entry is 0. Then the $Y_j^{(i)}$ will be 1 if and only if $X_j^{(i)} < T^{(i)}$, which is to say $X_j^{(i)}$ occurred before our random cutoff time $T^{(i)}$.

Because we will be doing Bayesian analysis with this model, we will need prior distributions on μ, σ. This assumption will not be included when we find Maximum Likelihood Estimates using the EM Algorithm, but it is necessary for any Bayesian technique. We will say that μ is uniformly

distributed and the prior on all normal distributions on the parameters is $p(\mu,\sigma) \propto 1/\sigma^2$ when we do the Bayesian analysis. This is an improper prior distribution. This will serve as a good conjugate prior (as stated in [40]), making the posterior distribution easier to calculate.

In the partially ranked Thurstonian model, we do not have direct access to the variables $(X^{(i)}, T^{(i)})$. These are hidden variables. The only variables that are observed are the $Y^{(i)}$. The parameters of the Thurstonian mutation model, μ and σ, are not known. The presence of latent variables along with the discrete nature of the observed data coming from continuous random variables makes this a Thurstonian model. The goal is to deduce information about μ,σ and use this information to make conclusions about ranking order.

This partially ranked Thurstonian model differs from a traditional Thurstonian model in a few key ways. All Thurstonian models have latent variables which are normally distributed and seek to recover a “true" ordering on N items. What is observed are judge(s) total rankings of these N items in the form of a permutation. It is assumed that any given judge may not rank the N items in the same way every time he orders them. The fact that the data we observed is not a total ranking but a partial ranking make the partially ranked Thurstonian model different from its traditional counterpart. A second key difference is the partially ranked Thurstonian model considers time as a random variable. The traditional Thurstonian model has no concept of a time variable, as it was assumed that the time the rankings occur should not affect the judge(s) rankings of the N items. From a biological standpoint, it makes sense to consider time as a variable, as not every patient will visit their doctor or get tested at the exact same stage of the disease. In the context of mutations, the “judge" may not make sense. However we do know that mutations do not always occur in some fixed order but do tend to occur in some sort of order. Thus, the idea idea of a judge is more accurately thought of as the virus’ or cell’s propensity to mutate in some order.

A traditional problem with Thurstonian models is, due to the structure of the model (i.e. random

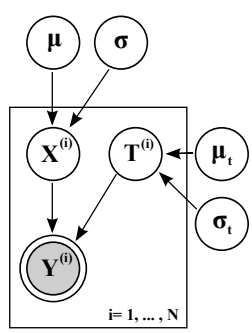

Figure 4.1 Graphical representation of proposed model

variables following normal distribution), there is no closed form for the MLEs of the parameters of the model. This is one reason why other ranking models are preferred to the Thurstonian model. Methods around the rather significant problem of parameter estimation have been proposed [15] [44] for a traditional Thurstonian model. We develop two different methods of recovering global ranking within the partially ranked Thurstonian model as well as efficient methods for parameter estimation.

4.3 Bayesian Methods

The first method we use to analyze data will be a Bayesian technique. All Bayesian analyses involve sampling from the posterior distribution of some variable (in most instances, a parameter). We will use an iterative MCMC algorithm called a Gibbs sampler to sample from the posterior distribution. The authors of [44] showed that the Gibbs sampler is a computationally effective method of parameter estimation for Thurstonian models. Furthermore, it is shown in [15] that partially ranked data can be used to greatly increase computational efficiency within a Thurstonian model, though in this paper the authors used rank dependencies to further simplify many of the probability formulas that do not have a closed form in the Thurstonian model.

The general idea of this algorithm is to sample values of our hidden variables based on those variables' distributions, the current value of our parameter estimation, and the observed variables. A Gibbs sampler is typically used for obtaining a large number of observations approximated from a specific multivariate probability distribution when direct sampling is hard. Because we cannot observe our variables $X^{(i)}, T^{(i)}$ or our parameters μ, σ, we use the Gibbs sampler to approximate values of these from conditional probability distributions. We iterate through this process several times, continually updating our estimations for μ, σ.

For the Thurstonian mutation model, the first step of each iteration will be to go through each of our N observations and take a sample value for $t^{(i)}$. First, we let

$$x_U^{(i)} = \min_{Y_j^{(i)}=0} x_j^{(i)} \qquad \text{and} \qquad x_L^{(i)} = \max_{Y_j^{(i)}=1} x_j^{(i)}$$

be the lowest value of $x_j^{(i)}$ such that $Y_j^{(i)} = 0$ and the highest value of $x_j^{(i)}$ such that $Y_j^{(i)} = 1$, respectively. The values of the $x^{(i)}$ will come from our previous iteration. We want to guarantee that $t^{(i)} < x_j^{(i)}$ if $Y_j^{(i)} = 0$ and $x_j^{(i)} < t^{(i)}$ if $Y_j^{(i)} = 1$; thus $x_U^{(i)}, x_L^{(i)}$ will serve as upper and lower bounds on $t^{(i)}$. By bounding $t^{(i)}$ in this way, we ensure that for every sampled $t^{(i)}, x^{(i)}$, the corresponding $y^{(i)}$ will always be the same as our original observed $Y^{(i)}$. To use this Gibbs sampler, we need to know the conditional

distributions for each of our variables. The distributions are as follows:

$$t^{(i)}|x^{(i)}, Y^{(i)}, \mu, \sigma \sim \mathrm{N}_{\text{truncated}}(0, 1, x_L^{(i)}, x_U^{(i)})$$

where $\mathrm{N}_{\text{truncated}}(0, 1, x_L^{(i)}, x_U^{(i)})$ is the truncated normal distribution with mean 0, *standard deviation* 1, a lower bound of $x_L^{(i)}$ and an upper bound of $x_U^{(i)}$. So, if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Y \sim \mathrm{N}_{\text{truncated}}(\mu, \sigma, a, b)$ is equivalent to $Y = X \mid a \leq X \leq b$. Let

$$x_j^{(i)} \mid \mu_j, \sigma_j, Y_j^{(i)}, t^{(i)} \sim \begin{cases} \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, -\infty, t^{(i)}) & \text{if } Y_j^{(i)} = 1 \\ \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, t^{(i)}, \infty) & \text{if } Y_j^{(i)} = 0 \end{cases}$$

$$\begin{aligned} \sigma_j^2 \mid \mu_j, s_j^2, x_j^{(\cdot)}, t^{(\cdot)}, Y_j^{(\cdot)} &\sim \text{Scale-inv-}\chi^2(N-1, 1/s_i^2) \\ \mu_j \mid \sigma_j, \bar{x}_j, x_j^{(\cdot)}, t^{(\cdot)}, Y_j^{(\cdot)} &\sim \mathrm{N}(\bar{x}_j, (\sigma_j/N)^2) \end{aligned}$$

where $\bar{x}_j$ is the mean of the jth coordinate over all of the sampled $x_j^{(i)}$ and s_j^2 the variance of the jth coordinate over all samples. Here σ is sampled from the scaled inverse chi-square distribution. We know if $X \sim \text{Scale}-\text{inv}-\chi^2(\nu, \tau^2)$, then $\dfrac{X}{\tau^2 \nu} \sim \text{inv}-\chi^2(\nu)$ where ν is the degrees of freedom and τ is the scaling parameter. Then the algorithm for the Gibbs sampler is as follows:

Algorithm 4.3.1. Let $Y^{(1)}, \ldots, Y^{(N)}$ be observed values for the random variable Y. To estimate the values μ, σ, we first take initial estimates of μ, σ such that $p(\mu, \sigma) \propto \frac{1}{\sigma^2}$. We also need initial values of $x^{(i)}, t^{(i)}$ for all $i = 1, \ldots, N$. Then during each iteration we do the following:

1. Sample $t^{(i)} \sim \mathrm{N}_{\text{truncated}}(0, 1, x_L^{(i)}, x_U^{(i)})$ where

$$x_U^{(i)} = \min_{Y_j^{(i)}=0} x_j^{(i)} \qquad \text{and} \qquad x_L^{(i)} = \max_{Y_j^{(i)}=1} x_j^{(i)}$$

   for each $i = 1, \ldots, N$

2. Sample

$$x_j^{(i)} \mid \mu_j, \sigma_j, Y_j^{(i)}, t^{(i)} \sim \begin{cases} \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, -\infty, t^{(i)}) & \text{if } Y_j^{(i)} = 1 \\ \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, t^{(i)}, \infty) & \text{if } Y_j^{(i)} = 0 \end{cases}$$

   for each $j = 1, \ldots, k$ and $i = 1, \ldots, N$

3. Sample $\sigma_j^2 | \mu_j, s_j^2, x_j^{(\cdot)}, t^{(\cdot)}, Y_j^{(\cdot)} \sim \text{Scale-inv-}\chi^2(N-1, 1/s_j^2)$ where s_j^2 is the variance of $x_j^{(i)}$ over

$i = 1, \ldots, N$ for each j.

4. Sample $\mu_i \,|\, \sigma_i, \bar{x_i}, x_j^{(\cdot)}, t^{(\cdot)}, Y_j^{(\cdot)} \sim \mathrm{N}(\bar{x_i}, (\sigma_i/N)^2)$ where $\bar{x_j}$ is the mean of $x_j^{(i)}$ over $i = 1, \ldots, N$ over each j.

5. Repeat

We record estimates for our μ, σ at set intervals. We do not want to record every estimation of μ, σ as it is possible that our algorithm gets "trapped" in a region where estimates for μ and σ are not likely to change much due to the conditions placed on them. Therefore, to ensure our samples are sufficiently random, we ran our procedure with a large burn-in value and multiple chains. From each chain we draw a set number of samples with a large number of iterations between each sample. These guidelines allow us to say the the samples are sufficiently random. After recording each of our sampled μ values, we transform each of them into a permutation and consider the mode of these permutations to be the most likely mutation order.

4.4 Maximum Likelihood Estimation For Thurstonian Modal

Another method of approximating μ and σ is to find the Maximum Likelihood Estimate (MLE) of each. These are the μ, σ which maximize the value of the log-likelihood function $\ell(\mu, \sigma \,|\, Y)$. This is immediately a problem, as it is very hard, even with a known μ, σ to calculate this log-likelihood. Instead, we will find the μ, σ which maximize the log-likelihood function $\ell(\mu, \sigma \,|\, X, T)$, which is much easier. However, maximizing $\ell(\mu, \sigma \,|\, X, T)$ is only easier if we know X, T, which we do not.

In [44], Ennis and Ennis show that by imposing rank dependencies on the items, certain conditional probabilities become much easier to compute. They go on to conclude that the computation efficiency of this method will be increased further by using partially ranked data. We introduce a way to estimate the necessary conditional probabilities rather than compute them directly; these estimates will allow us to make use of the Expectation Maximization algorithm without imposing any rank dependencies.

The Expectation Maximization (EM) algorithm is used when computing parameters directly is difficult, often because there are hidden variables. As this is true of the Thurstonian mutation model, we use the EM algorithm to estimate $\mu_j = \mathbb{E}(X_j^{(\cdot)})$. The general idea is as follows: we will initialize the algorithm with some values for our parameters (μ^0, σ^0). Then

1. We estimate hidden variables $T^{(i)}, X^{(i)}$ at the mth step for each of our N observations based on μ^m, σ^m and $Y^{(i)}$ according to the distribution prescribed by the model which is being examined.

2. Using the estimated values of our $T^{(i)}, X^{(i)}$, we compute the value of our parameters μ^{m+1}, σ^{m+1} which maximize log-likelihood function, $\ell(\mu, \sigma \mid T^{(i)}, X^{(i)})$.

3. Repeat.

While this is intuitively what we want, it is not exactly what is actually done in the EM algorithm. We will not compute the values of $X^{(i)}$, $T^{(i)}$; instead, at the mth step, we will compute the values of the sufficient statistics $\mathbb{E}[X_j^{(i)} \mid Y^{(i)}, \mu, \sigma]$, $\mathbb{E}[(X_j^{(i)})^2 \mid Y^{(i)}, \mu, \sigma]$. Unfortunately, computing the values of these sufficient statistics is difficult, as

$$\mathbb{E}[X_j^{(i)} \mid Y^{(i)}, \mu, \sigma] = \iiint\limits_{\substack{x_r < t \text{ if } Y_\ell^{(i)}=1 \\ x_s > t \text{ if } Y_p^{(i)}=0}} x_j \, F(\mathbf{x}, t \mid \mu, \sigma) \, d\mathbf{x} \, dt$$

$$\mathbb{E}[(X_j^{(i)})^2 \mid Y^{(i)}, \mu, \sigma] = \iiint\limits_{\substack{x_r < t \text{ if } Y_\ell^{(i)}=1 \\ x_s > t \text{ if } Y_p^{(i)}=0}} x_j^2 \, F(\mathbf{x}, t \mid \mu, \sigma) \, d\mathbf{x} \, dt$$

where

$$F(\mathbf{x}, t \mid \mu, \sigma) = f(t \mid \mu, \sigma) \prod_{p=1}^{k} f(x_p \mid \mu, \sigma)$$

and $f(x|\mu, \sigma)$ is the normal probability density function. These integrals are difficult to compute, but without these sufficient statistics, we cannot use the EM algorithm. Direct computation of these estimated values is not feasible, so we will estimate these expected values using importance sampling.

It is possible to estimate expected values by sampling points according to the probability distribution and averaging them. For example, if we want to estimate $\mathbb{E}[X]$ and we know X is distributed according to a specific probability distribution, we draw several random points according to this probability distribution and then take the average of these points. In our model, we know each $X_j^{(i)}$ is normally distributed, so it would seem that we could sample from a normal distribution for each $X_j^{(i)}$ and take the average of these values to get the expected value we are interested in. But drawing a random sample $X^{(i)}$ from a normal distribution might cause our $X^{(i)}$ to fall outside of the region of integration. This means we will not be able to use this exact approach. Importance sampling allows us to estimate this same expected value by drawing samples from a different distribution as long as the two corresponding integrals are the same. Consider the above integral as

$$\iiint\limits_{\substack{x_\ell < t \text{ if } Y_\ell^{(i)}=1 \\ x_p > t \text{ if } Y_p^{(i)}=0}} F(\mathbf{x}, t \mid \mu, \sigma) \, d\mathbf{x} \, dt = \iiint\limits_{\substack{x_\ell < t \text{ if } Y_\ell^{(i)}=1 \\ x_p > t \text{ if } Y_p^{(i)}=0}} \frac{F(\mathbf{x}, t \mid \mu, \sigma)}{G(\mathbf{x}, t \mid \mu, \sigma)} \, G(\mathbf{x}, t \mid \mu, \sigma) d\mathbf{x} \, dt$$

where

$$G(\mathbf{x}, t \mid \mu, \sigma) = f(t \mid \mu, \sigma) \prod_{p=1}^{k} g(x_p \mid \mu, \sigma)$$

$f(t \mid \mu, \sigma)$ is still the normal probability density function, and $g(x \mid \mu, \sigma)$ is the probability density function of the truncated normal distribution bounded above or below by t, according to the corresponding $Y^{(i)}$. Now, using importance sampling, we can approximate our expected values (and therefore our sufficient statistics) by taking many samples of $t \sim \mathrm{N}(\mu_t, \sigma_t)$ and samples of $\mathbf{x}$ from the truncated normal distribution

$$x_j^{(i)} \mid \mu_j, \sigma_j, Y_j^{(i)}, t^{(i)} \sim \begin{cases} \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, -\infty, t^{(i)}) & \text{if } Y_j^{(i)} = 1 \\ \mathrm{N}_{\text{truncated}}(\mu_j, \sigma_j, t^{(i)}, \infty) & \text{if } Y_j^{(i)} = 0 \end{cases}$$

If we take m such samples as directed above, then the expected value will be estimated by

$$\mathbb{E}[X_j^{(i)} \mid Y^{(i)}, \mu, \sigma] \approx \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \frac{F(\mathbf{x}^{(i)}, t^{(i)} \mid \mu, \sigma)}{G(\mathbf{x}^{(i)}, t^{(i)} \mid \mu, \sigma)}$$

which is the average value of $\frac{F}{G}$ evaluated at each of our sampled points.

Knowing, then, that we can approximate these expected values, we can now maximize the log-likelihood $\ell(\mu, \sigma \mid X^{(i)}, T^{(i)})$ and prove that the values that maximize the log-likelihood function are, in fact,

$$\mu_j = \mathbb{E}[X_j \mid Y^{(i)}, \mu, \sigma] = \iiint\limits_{\substack{x_\ell < t \text{ if } Y_\ell^{(i)} = 1 \\ x_p > t \text{ if } Y_p^{(i)} = 0}} x_j \, \frac{F(\mathbf{x}, t \mid \mu, \sigma)}{G(\mathbf{x}, t \mid \mu, \sigma)} \, G(\mathbf{x}, t \mid \mu, \sigma) d\mathbf{x} \, dt$$

and $\sigma_j^2 = \mathbb{E}[X_j^2 \mid Y^{(i)}, \mu, \sigma] - (\mathbb{E}[X_j \mid Y^{(i)}, \mu, \sigma])^2$ where

$$\mathbb{E}[X_j^2 \mid Y^{(i)}, \mu, \sigma] = \iiint\limits_{\substack{x_\ell < t \text{ if } Y_\ell^{(i)} = 1 \\ x_p > t \text{ if } Y_p^{(i)} = 0}} x_j^2 \, \frac{F(\mathbf{x}, t \mid \mu, \sigma)}{G(\mathbf{x}, t \mid \mu, \sigma)} \, G(\mathbf{x}, t \mid \mu, \sigma) d\mathbf{x} \, dt$$

We describe this in the following algorithm.

Algorithm 4.4.1. Let $Y^{(1)}, \ldots, Y^{(N)}$ be observed data for the above model. We initialize vectors $\mu^{(0)}, \sigma^{(0)}$ randomly from a uniform distribution. Set an M which will be the number of samples we draw each time we use importance sampling. Then, at the mth step of the algorithm:

1. Using importance sampling, for $i = 1, \ldots, N$ estimate

$$\mathbb{E}[X_j^{(i)} \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}] = \tfrac{1}{M} \sum_{\ell=1}^{M} x^{(\ell)} \, \frac{F(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)})}{G(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)})}$$
$$\mathbb{E}[(X_j^{(i)})^2 \mid Y^{(i)}, \mu, \sigma] = \tfrac{1}{M} \sum_{\ell=1}^{M} (x^{(\ell)})^2 \, \frac{F(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)})}{G(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)})}$$

   where

$$F(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)}) = f(t \mid \mu^{(m)}, \sigma^{(m)}) \prod_{p=1}^{k} f(x_p^{(\ell)} \mid \mu^{(m)}, \sigma^{(m)})$$

   and $f(x|\mu, \sigma)$ is the probability density function for the normal distribution. Similarly, $G(x^{(\ell)}, t \mid \mu^{(m)}, \sigma^{(m)})$ is the product of the truncated normal probability density functions.

2. Compute

$$\mathbb{E}[X_j^{(\cdot)} \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}] = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbb{E}[X_j^{(i)} \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}]$$
$$\mathbb{E}[(X_j^{(\cdot)})^2 \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}] = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbb{E}[(X_j^{(i)})^2 \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}] \quad .$$

3. Set $\mu_j^{m+1} = \mathbb{E}[X_j^{(\cdot)} \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}]$ and $\sigma_j^{(m+1)} = \mathbb{E}[(X_j^{(\cdot)})^2 \mid Y^{(i)}, \mu^{(m)}, \sigma^{(m)}]$.

4. Repeat.

4.5 Results

We analyze two different datasets using the partially ranked Thurstonian model, using the Gibbs sampler and EM algorithm techniques described above on both datasets. These datasets were analyzed before by Beerenwinkel and Sullivant using the Markov model they proposed in [3]. We compare the results of the two different methods proposed in this paper with each other as well as with the results of Beerenwinkel and Sullivant.

4.5.1 HIV Dataset

The first is a dataset contains 364 observations and is comprised of 7-vectors for HIV mutations. As explained in [3], each mutation is an accumulation of amino acid changes in a segment of the HIV *pol* gene. This is the gene which codes for the viral protein reverse transcriptase (RT), the process HIV uses to trick human cells into producing many copies of the virus. Each of these seven mutations

$X_1 \longleftarrow$ 41L
$X_2 \longleftarrow$ 67N
$X_3 \longleftarrow$ 69D
$X_4 \longleftarrow$ 70R
$X_5 \longleftarrow$ 210W
$X_6 \longleftarrow$ 215Y
$X_7 \longleftarrow$ 219Q

Figure 4.2a HIV mutations and corresponding X_i

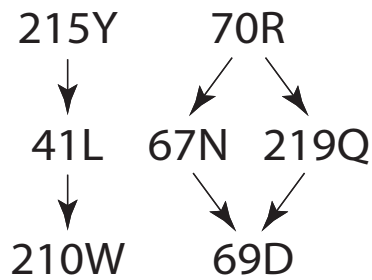

Figure 4.2b Maximum Likelihood Poset for HIV data, from Beerenwinkle and Sullivant [3]

Figure 4.2 The mutations corresponding to variables X_i and the maximum likelihood poset proposed in [3].

is associated to some form of drug resistance, and acquiring all seven mutations renders the virus completely drug resistant. The 364 observations analyzed were extracted from infected patients prescribed treatment with zidovudine, an antiretroviral designed to inhibit RT. These amino acid changes are inferred after DNA sequencing of the *pol* gene.

The seven mutations examined in this data are 41L, 67N, 69D, 70R, 210W, 215Y, and 219Q. These are shorthand, as 41L indicates the presence of the amino acid leucine (L) at position 41 of the RT. Figure 4.2a shows the mutation corresponding to each X_i.

When we applied the Gibbs sampler to the HIV data, we drew a total of 53,756 samples, each from an instance of the Gibbs sampler burn-in value 200 and 20 chains. From each chain we drew 20 samples and preformed 20 iterations between samples. Out of the total 53,756 samples, 33,636 samples had a mutation order corresponding to the permutation (4 5 8 2 7 3 6 1), with the time variable being last. Thus, in 33,636 of the 53,756 samples, mutation 1 occurred fourth, mutation 2 occurred fifth, etc. We notice that this permutation occurs a majority of the time.

The second most likely mutation order corresponded to the permutation (4 6 8 2 7 3 5 1) which occurred 8,140 out of the 53,756 observed sequences. We notice that this particular permutation is a single transposition away from the most common permutation. The third most common permutation is (4 5 8 3 7 2 6 1) which is again a single transposition away from the most observed permutation. This permutation occurred occurred 6,697 times.

As a more succinct way to summarize the results, we can use a paired comparison matrix. This is the matrix whose (i, j) entry is the marginal probability that $(X_i \geq X_j)$ in the posterior distribution under the Bayesian Thurstonian mutation model. We also note that the eighth row and column correspond to the time variable. This data is in strong agreement with the findings in [3]. In their paper, Beerenwinkel and Sullivant use their Markov model to determine the maximum likelihood

Table 4.1 Paired comparison matrix for HIV data

$$
\begin{array}{c}
 \\ 41L \\ 67N \\ 69D \\ 70R \\ 210W \\ 215Y \\ 219Q \\ T
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} 41L & 67N & 69D & 70R & 210W & 215Y & 219Q & T \end{array} \\
\begin{pmatrix}
 & 0.0231 & 0.0 & 0.9999 & 0.0 & 0.9998 & 0.0022 & 0.9999 \\
0.9768 & & 0.00002 & 0.9999 & 0.0008 & 0.9999 & 0.1901 & 0.9999 \\
0.9999 & 0.9999 & & 0.9999 & 0.9667 & 0.9999 & 0.9999 & 0.9999 \\
0.00004 & 0.00002 & 0.00004 & & 0.00006 & 0.1652 & 0.00006 & 0.9987 \\
0.9999 & 0.9991 & 0.0332 & 0.9999 & & 0.9999 & 0.9902 & 0.9999 \\
0.0001 & 0.00006 & 0.0000 & 0.8347 & 0.00007 & & 0.00009 & 0.9999 \\
0.9976 & 0.8097 & 0.0001 & 0.9998 & 0.0097 & 0.9998 & & 0.9998 \\
0.0001 & 0.0001 & 0.0001 & 0.0014 & 0.0001 & 0.0001 & 0.0001 &
\end{pmatrix}
\end{array}
$$

poset dictating mutation order and note this poset is compatible with 87% of the observations. Of the samples taken from the posterior distribution of our Bayesian analysis of the Thurstonian mutation model, nearly 100% of the sampled μ values are compatible with their proposed poset. Only 7 of the $53,756$ sampled μ values were not compatible with the poset. It should be noted that in the structure of the Markov model, stopping time was not considered as a random variable.

To examine the EM algorithm data, we consider the μ, σ pairs with the highest log-likelihood scores. When we examine the maximum likelihood estimates for the parameters μ, σ in the HIV data in the Thurstonian mutation model, the 30 μ, σ pairs with the highest log-likelihood scores vary little from individual sample to sample. The μ, σ pair with the highest log-likelihood score, rounded to 3 decimal places is

$$\mu = (0.703, 2.089, 14.807, 3.417, 1.371, 0.304, 4.988, 0)$$

$$\sigma = (0.490, 2.411, 10.721, 18.234, 0.616, 0.173, 5.468, 1)$$

We notice that the overall behavior of the μ values with the highest log-likelihood score is fairly consistent between the individual μ values. Notice that lost every μ listed has the same corresponding permutation. In fact, the top 30 μ values correspond to just 2 different permutations, which are exactly the top two permutations observed in the Gibbs sampler. This gives us some insight into how likely a particular mutation is to occur.

We see a similar trend in the σ values which correspond to these μ values:

Again, we see these values of σ have fairly consistent behavior. While there is some variation between the observed σ values, the values are fairly consistent. We know that the fourth mutation has a large standard deviation, but because it varies so much, we can not say for certain exactly what the value of the standard deviation for the fourth mutation is. However, the standard deviations for the first, second, fifth, and sixth mutation have a standard deviations that are almost identical from

sample to sample.

This behavior in the μ, σ pairs seems to indicate that these μ, σ pairs are fairly good estimates for the maximum likelihood estimators.

4.5.2 Prostate Cancer Cell Dataset

The second contains 54 observations of 9-vectors for prostate cancer cell mutations. This the same data analyzed in [3], in which the authors describe the data as coming from comparative genome hybridization experiments. This is a technique used to detect large scale genomic alterations, particularly the severing or attachment of chromosome arms, which is common in cancer cells. As an example, the event $4q+$ denotes the gain $(+)$ of additional copies of the large (q) arm of chromosome 4. Similarly, $8p-$ denotes the loss $(-)$ of the small arm (p) of the 8th chromosome. The observations are defined by the presence or absence of nine genetic alterations, $3q+, 4q+, 6q+, 7q+, 8p-, 8q+, 10q-, 13q-$, and $Xq+$.

We can do the same analysis on the prostate data. Unfortunately, while the HIV had a single permutation dominate all the observations, of the $46,585$ observations from the Gibbs sampler for the prostate data, the permutation that occurred most often only occurred 20 times, which is just a little more than 0.04% of the observations. This might suggest that the mutation order for prostate cancer is not fixed. That is, the mutation order for prostate cancer is nearly random.

As before, we can look at the paired comparison matrix of the results. Again, we let the (i, j) entry be the marginal probability that $(X_i \geq X_j)$ in the posterior distribution under the Bayesian Thurstonian mutation model. Again, we let the last row and the last column represent the time variable. These results seem to suggest that, for the most part, the mutation order for prostate cancer cells is not fixed. Although it is almost always the case that the stopping time happens before any mutations occur, it is difficult to say much more with any level of certainty. We see that almost 95% of the time, mutation 1 occurred before mutations 5, 6, and 9. Still, many of the entries in the paired comparison matrix are close to .50, indicating that there an equal chance for either mutation to occur first.

Why then, did the posterior distribution for the prostate cancer dataset come out with so much variability, where the posterior distribution from the HIV dataset had very little? There are perhaps a few reasons. First, there were only 54 observations in this dataset, whereas there were 364 observations in the HIV dataset. Compound that with the fact that this dataset observed 9 mutations, as opposed to the HIV dataset's 7. More mutations and fewer observations are bound to lead to greater uncertainty. The uncertainty could also be a result of the form of the data, as 1/3 of the observations recorded no mutations having occurred. Finally, it could be that there are simply fewer dependencies between the mutations for prostate cancer cells than there are in the mutations of HIV. The authors of [3] do in

Table 4.2 Paired comparison matrix for prostate cancer data

$$
\begin{array}{c}
 \\ 10q- \\ 8p- \\ 13q+ \\ 3q+ \\ 4q+ \\ 6q+ \\ 7q+ \\ 8q+ \\ Xq+ \\ T
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc} 10q- & 8p- & 13q+ & 3q+ & 4q+ & 6q+ & 7q+ & 8q+ & Xq+ & T \end{array} \\
\begin{pmatrix}
* & 0.7831 & 0.8226 & 0.5722 & 0.9437 & 0.9486 & 0.7750 & 0.9047 & 0.9492 & 1 \\
0.2168 & * & 0.5669 & 0.2822 & 0.8062 & 0.8144 & 0.4985 & 0.7223 & 0.8061 & 1 \\
0.1774 & 0.4331 & * & 0.2333 & 0.7472 & 0.7593 & 0.4330 & 0.6564 & 0.7496 & 1 \\
0.4278 & 0.7178 & 0.7666 & * & 0.9134 & 0.9210 & 0.7116 & 0.8654 & 0.9215 & 0.9999 \\
0.0564 & 0.1938 & 0.2528 & 0.0866 & * & 0.5133 & 0.2017 & 0.3949 & 0.4913 & 0.9999 \\
0.0514 & 0.1856 & 0.2407 & 0.0790 & 0.4867 & * & 0.1904 & 0.3849 & 0.4796 & 0.9999 \\
0.2251 & 0.5015 & 0.5670 & 0.2884 & 0.7983 & 0.8096 & * & 0.7165 & 0.8029 & 0.9999 \\
0.0954 & 0.2777 & 0.3437 & 0.1346 & 0.6051 & 0.6150 & 0.2835 & * & 0.5955 & 0.9998 \\
0.0509 & 0.1939 & 0.2504 & 0.0785 & 0.5086 & 0.5204 & 0.1971 & 0.4044 & * & 0.9998 \\
0.0002 & 0.0002 & 0.0002 & 0.0002 & 0.0002 & 0.0002 & 0.0002 & 0.0002 & 0.0002 & *
\end{pmatrix}
\end{array}
$$

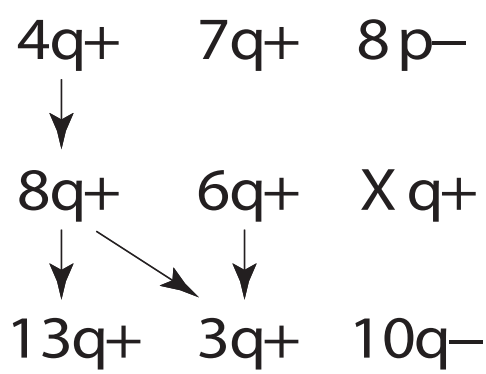

Figure 4.3 Maximum Likelihood Poset for prostate cancer data

fact conclude there are fewer dependencies in prostate cancer cell mutations than in HIV mutations, which can be seen in the Maximum Likelihood Poset the propose in Figure 4.3.

The data we computed is different in many ways than the data presented by Beerenwinkel and Sullivant in [3]. In the HIV data, nearly all the samples taken were compatible with the poset proposed in their findings. In the prostate cancer data, however, 39,951 of the 46,585 samples in the posterior distribution are incompatible with the maximum likelihood poset Beerenwinkel and Sullivant found using the same prostate cancer cell data. This could suggest that when few data points are available, the choice of the model used will greatly affect the conclusions that can be drawn. It is also possible that since Beerenwinkel and Sullivant used maximum likelihood estimation and this is a Bayesian sampling of the posterior distribution, it could be that maximum likelihood estimation is more effective for this particular kind of data.

To see if this is indeed the case, we examine the maximum likelihood estimate pairs μ, σ with the highest log-likelihood values. Examining the 30 μ, σ pairs with the highest log-likelihood score, we see in the top 30 μ have fairly consistent behavior between individual estimates. For each i, the actual value of the estimates of μ_i do not change much from estimate to estimate, but the top 30 estimates for μ correspond to six different mutation orders. This is not entirely surprising, as the estimates for μ_5, μ_6, μ_8 are all grouped tightly around 0.75. This is consistent with our findings in the Gibbs sampler, as in the paired comparison matrix the $(5,6)$ entry is almost exactly 0.5 and the $(5,8)$ and $(6,8)$ entries fall between 0.38 and 0.4. We also see that the corresponding $\sigma_5, \sigma_6, \sigma_8$ are very small. The μ, σ pair with the highest log-likelihood score, rounded to three decimal places, are

$$\mu = (31.020, 11,981, 0.942, 1.347, 0.708, 0.767, 1.289, 0.750, 1.385, 0)$$

$$\sigma = (25.642, 12.082, 0.466, 0.810, 0.350, 0.594, 1.077, 0.082, 1.756, 1)$$

As with the behavior of the means for the HIV data, the means for the prostate data are very similar to one another. However, unlike the HIV data, many of these means are clustered much closer together, such at the mean mutation time for mutations five, six, and eight and mutations four, seven and nine. This might suggest that in among these two subsets, it is equally likely that any of these three mutations occurs first. Also interesting is the fact that most of these mutations have a standard deviation of less than 1. We also notice that there are six different mutation orders in these top 30 μ values, as opposed to the two different mutation orders in the μ values for the HIV data.

BIBLIOGRAPHY

[1] Archambeau, Cédric and François Caron. Plackett-Luce regression: a new Bayesian model for polychotomous data. *International conference on Uncertainty in Artificial Intelligence* (2012), Catalina Island USA. arXiv:1210.4844

[2] Beerenwinkel, N., N. Eriksson and B. Sturmfels. Conjunctive Bayesian networks. *Bernoulli* 13 (2007)., 893Ð909.

[3] Beerenwinkel, Niko and Seth Sullivant. Markov models for accumulating mutations. *Biometrika* (2009), Vol. 96: 645-661.

[4] Bi, Jian and Carla Kuesten. Estimating and Testing Parameters of the Thurstonian Model for Torgerson's Method of Triads. *Journal of Sensory Studies* (2015), Vol. 30: 33Ð45. doi: 10.1111/joss.12134

[5] Björner, A. and F. Brenti *Combinatorics of Coxeter Groups*, Springer Science, 2005.

[6] Bonin, Joseph, Louis Shapiro and Rodica Simion. Some q-Analogs of the Schröder Numbers Arising from Combinatorial Statistics on Lattice Paths. *Journal of Statistical Planning Inference* 34, 35-55, 1993.

[7] Caron, François, Yee Whye Teh and and Thomas Brendan Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programs. *The Annals of Applied Statistics* 2014, Vol. 8 No. 2, 1145-1181. DOI: 10.1214/14-AOAS717

[8] Ceberio, Josu, Alexander Mendiburu and Jose A. Lozano. Introducing the Mallows model on estimation of distribution algorithms. *Neural Information Processing*. Springer Berlin Heidelberg 2011, 461-470. doi: 10.1007/978-3-642-24958-7_54.

[9] Christensen, Rune Haubo Bojesen, Hye-Seong Lee and Per Bruun Brockhoff. Estimation of Thurstonian model for the 2-AC protocol. *Food Quality and Preference* (April 2012), Vol. 24: 119-128. doi:10.1016/j.foodqual.2011.10.005

[10] Cox, D., J. Little and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Third edition. Springer, 2007.

[11] Dempster, A.P, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* (1977), Series B 39 (1), pp. 1Ð38.

[12] Diaconis, P. *Group Representations in Probability and Statistics*. Hayward, CA. Institute of Mathematical Statistics, 1988.

[13] Dorfman, D. and E. Alf. Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals–Rating-Method Data. *Journal of Mathematical Psychology*, Volume 6, 487-496, 1969.

[14] Drton, M., B. Sturmfels, S. Sullivant. *Lectures on Algebraic Statistics*. Birkhäuser Verlag, 2009.

[15] Ennis, Daniel M. and John M. Ennis. A Thurstonian ranking model with rank-induced dependencies. *Journal of Classification* (2013) Vol. 30, 124 -147.

[16] Ennis, Daniel M., and Benoît Rousseau. A Thurstonian model for the degree of difference protocol. *Food Quality and Preference* (April 2015), Vol. 41: 159-162. doi:10.1016/j.foodqual.2014.11.011

[17] Ennis, John M., Daniel M. Ennis, Dorene Yip, and Michael O'Mahony. Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical & Statistical Psychology* (1998), Vol. 51, 205-215.

[18] Francis, Brian, Regina Dittrich, Reinhold Hatzinger, and Les Humphreys. A mixture model for longitudinal partially ranked data. *Communications in Statistics–theory and Methods* 43, 2014. 722-734.

[19] Geman, Stuart and Donald Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal Of Applied Statistics*(1993) Vol. 20, Issue 5-6, pp. 25 - 62.

[20] Gianola, Daniel and Henner Simianer. A Thurstonian model for quantitative genetic analysis of ranks: a Bayesian approach. *Genetics* (2006), Vol. 174. DOI: 10.1534/genetics.106.060673

[21] Guiver, John, and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. Inproceedings, ICML 17 June 2009.

[22] Hassett, B. *Introduction to Algebraic Geometry*. Cambridge University Press, 2007.

[23] Kassel, Christian and Vladimir Turaev. *Braid Groups*. Springer New York, 2008. doi: 10.1007/978-0-387-68548-9.

[24] Kim, Hee-Jin, Seon Young Jeon, Kwang-Ok Kim and Michael O'mahony. THURSTONIAN MODELS AND VARIANCE I: EXPERIMENTAL CONFIRMATION OF COGNITIVE STRATEGIES FOR DIFFERENCE TESTS AND EFFECTS OF PERCEPTUAL VARIANCE. *Journal of Sensory Studies* (2006), Vol. 21: 465Ð484. doi: 10.1111/j.1745-459X.2006.00074.x

[25] Kim,Hee-Jin, Seon Young Jeon, Kwang-Ok Kim and Michael O'Mahony. THURSTONIAN MODELS AND VARIANCE II: EXPERIMENTAL CONFIRMATION OF COGNITIVE STRATEGIES FOR DIFFERENCE TESTS AND EFFECTS OF PERCEPTUAL VARIANCE. *Journal of Sensory Studies* (2006), Vol. 21: 465Ð484. doi: 10.1111/j.1745-459X.2006.00079.x

[26] Lebanon, Guy and Yi Mao. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research 9*, 2008, 2401-2419.

[27] Lee, Michael D., Mark Steyvers, Mindy de Young, and Brent J. Miller. A model-based approach to measuring expertise in ranking tasks. CogSci 2011 proceedings.

[28] Luce, R. D. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., 1959.

[29] Mallows, C. L. . Non-null ranking models. *Biometrika*, 44:114-130, 1957.

[30] Mollica, Cristina and Luca Tardella. Bayesian mixture of Plackett-Luce models for partially ranked data, (2015), arXiv:1501.03519

[31] Mollica, Cristina and Luca Tardella. (2014), Epitope profiling via mixture modeling of ranked data, Int. J. Numer. Meth. Fluids (2014), Vol. 33, pages 3738Ð3758. DOI: 10.1002/sim.6224

[32] Ott, R. Lyman and Michael Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Sixth Edition. Brooks/Cole, Cenage Learning 2010.

[33] Plackett, Robin L. The analysis of permutations. *Applied Statistics*, 24(2):193Ð202, 1975.

[34] De Poi, Pietro. On higher secant varieties of rational normal scroll. Matematiche (Catania), 51 (1996), no. 1, 3Ð21.

[35] Roberts, G. O. and A. F. M. Smith. Simple conditions for conditions of the Gibbs sampler and Metropolis-Hastings algorithm. *Stochastic Processes and their Applications*, Vol. 49 Issue2, February 1994, 207-218.

[36] Sagan, Bruce E. *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*. Second Edition. Springer-Verlag, 2001.

[37] Sastry, N.S. Narasimha. *Groups of Exceptional Type, Coxeter Groups and Related Geometries*. Springer India, 2014. dpi: 10.1007/978-81-322-1814-2.

[38] Sidman, Jessica and Seth Sullivant. Prolongations and computational algebra. *Canadian Journal of Mathematics* **61** no. 4 (2009) 930-949 **math.AC/0611696**

[39] Stanley, Richard P. *Ennumerative Combinatorics, Volume I*. Second Edition. Cambridge University Press, 2012.

[40] Steyvers, Mark, Michael Lee, Brent Miller, and Pernille Hemmer. “Wisdom of Crowds in the Recollection of Order Information." *Advances in Neural Information Processing Systems 22* (2009), 1785-1793.

[41] Sturmfels, Bernd and Volkmar Welker. Commutative Algebra of Statistical Ranking. *Journal of Algebra* 361 (2012), 264-286.

[42] Thurstone, Louis Leon. The law of comparative judgement. *Psychological Review*, Vol 34(4), Jul, 1927. pp. 273-286.

[43] Wasserman, Larry . *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

[44] Yao, Grace and Ulf Böckenholt. Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology* (1999), Vol. 52, 79Ð 92.