**Title**
Algebraic Matroids in Applications



**Author**
Rosen, Zvi

**Publication Date**
2015

Peer reviewed|Thesis/dissertation

# Algebraic Matroids in Applications

by

Zvi H Rosen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bernd Sturmfels, Chair
Professor Lior Pachter
Professor Yun S. Song

Spring 2015

**Algebraic Matroids in Applications**

## Abstract

Algebraic Matroids in Applications

by

Zvi H Rosen

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Bernd Sturmfels, Chair

Algebraic matroids are combinatorial objects defined by the set of coordinates of an algebraic variety. These objects are of interest whenever coordinates hold significance: for instance, when the variety describes solution sets for a real world problem, or is defined using some combinatorial rule. In this thesis, we discuss algebraic matroids, and explore tools for their computation. We then delve into two applications that involve algebraic matroids: probability matrices and tensors from statistics, and chemical reaction networks from biology.

In memory of my grandparents

| | |
|---|---|
| Isaac and Ann Davis | יצחק אהרן וחנה רייזל דייוויס ז"ל |
| Isaac and Sala Rosen | יצחק יוסף ושרה רוזן ז"ל |

whose courage and perseverance through adversity
will inspire their families for generations.

# Contents

# List of Figures

# List of Tables

## Acknowledgments

אוֹדה ה׳ בכל לבי אספרה כל נפלאותיך: (תהלים ט:ב)
*I thank the Lord with my whole heart ...* (Psalms 9:2)

Working towards a doctorate can come with unexpected challenges. The process involves learning a great deal about one's area of research, as well as learning a lot about yourself. I have been blessed to have many supporters who encouraged me to pursue a career in mathematics. Amitai Bin-Nun encouraged me to pursue research when I was an undergraduate, helping me apply to summer programs; he continues providing insightful advice when I call. Jessica Sidman gave me my first experience in math research, getting me hooked on the experience; she also continues to be a great advisor. Many professors from my undergraduate training, especially Dennis DeTurck and Andreea Nicoara, taught me beautiful mathematics and helped me apply to graduate school.

My mathematical life in Berkeley included terrific teaching experiences with Per Persson and Slobodan Simic. My graduate research experiences were fun and enriching thanks to all of my coauthors, especially Elizabeth Gross, Heather Harrington, and Kaie Kubjas, with whom I wrote the work forming much of this thesis. I also want to extend my gratitude to Carina Curto, Alex Fink, Heather Harrington, and Shmuel Onn for their help in the process of applying to postdoctoral grants and jobs. I owe a great debt to Bernd Sturmfels for taking me on as a student. It has been a privilege to benefit from his mathematical insight and to participate in the community of great thinkers that surrounds him. I also want to thank him for going through multiple drafts of all my writing including this dissertation, and providing detailed feedback to improve it. He has demonstrated tremendous patience and support, and I am incredibly grateful.

The professors who served on the committee for my qualifying examination and my dissertation were Mark Haiman, Lior Pachter, Satish Rao, Yun Song, and Dan-Virgil Voiculescu. Their willingness to participate is highly appreciated. The Berkeley math department has an incredible staff who make the lives of the graduate student population much easier. In particular, I want to extend my appreciation to Barb Waller, Marsha Snow, and Judie Filomeo. I also spent a semester at the Max Planck Institute for Mathematics in Bonn, Germany, and Anke Völzmann was incredibly helpful in locating resources.

Moving to Berkeley, CA from my home on the East Coast was a daunting enterprise in itself and I thank the members of Congregation Beth Israel, particularly Rabbi Yonatan and Frayda Cohen and Avraham and Ruchama Burrell whose hospitality and kindness to me has been nothing short of extraordinary. I also want to thank Maharat Victoria Sutton and Adam Brelow, Zeev and Tamara Neumeier, Naaman and Meechal Kam, and Maayan and Elishav Rabinovich for being so welcoming and generous. My group of friends, particularly Ali Austerlitz, Arabella Bangura, Benjamin Epstein, Boaz Haberman, Amalya Lehmann, Matty Lichtenstein, Shivaram Lingamneni and Jonathan Thirman made Berkeley feel like home, and gave me tremendous support and friendship. I also want to thank Eitan Adler, Yonatan Cantor, Jonathan Eskreis-Winkler, Elie Friedman, Miki Friedmann, Yoni Halpern,

Ariella Meisel, Noam Pratzer, David Pruwer, and Mordechai Treiger whose support from afar kept me going.

I could never have made it this far without the love and support of my family: my mother and father especially, for always making themselves available in times of need, and for serving as living examples for me. My siblings and my brothers-in-law have also been unfailingly kind and welcoming whenever I have needed their help. It is a wonderful gift to have family that always picks up the phone.

# Chapter 1

# Introduction

Both chapters 1 and 2 are based on material from *Computing Algebraic Matroids* [48]. The definitions and introductory examples are brought to aid exposition, but similar examples can be found in standard texts such as [45] and [56].

Algebraic matroids have a surprisingly long history. They were studied as early as the 1930's by van der Waerden, in his textbook Moderne Algebra [55, Chapter VIII], and MacLane in one of his earliest papers on matroids [35]. The topic lay dormant until the 70's and 80's, when a number of papers about representability of matroids as algebraic matroids were published: notably by Ingleton and Main [23], Dress and Lovasz [11], the thesis of Piff [46], and extensively by Lindström ([32, 33, 30, 31], among others). In recent years, the algebraic matroids of *toric* varieties found application (e.g. in [43]); however, they have been primarily confined to that setting.

Renewed interest in algebraic matroids comes from the field of matrix completion, starting with [26], where the set of entries of a low-rank matrix are the ground set of the algebraic matroid associated to the determinantal variety. In applied algebra in general, coordinates typically carry real-world significance, and the matroid has inherent interest as the dependence structure among those quantities. Even for varieties arising in pure mathematics, distinguished coordinates may have combinatorial meaning, in which case the matroid also provides insight.

## 1.1 Summary of Main Results

The purpose of this dissertation is to provide the relevant tools for computation of algebraic matroids and to actually use them in practice. The remainder of Chapter 1 introduces matroids to the reader and defines technical language that will be used in the thesis. We prove some basic results where the proof is particularly instructive or relevant. In Chapter 2, we summarize two techniques for computing algebraic matroids associated to prime ideals. An original definition in this chapter is the *non-matroidal locus* of an affine variety. This is a subvariety which exhibits degenerate behavior under coordinate projection. In Propositions

2.2.2 and 2.2.4, we also give an explicit formula; as a corollary, this demonstrates that it has positive codimension. Therefore, generic computation of the Jacobian returns the correct matroid.

The sample computations, which fully describe some large matroids and their algebraic decorations are also new results. In particular, we compute the matroid associated to the Plackett-Luce model for statistical rankings of size 4, the mixture model matroid of rank 3 for $4 \times 4$ matrices, the matroid for the steady state variety of the MAP kinase network, the Grassmannian $Gr(3, 6)$, and two realizations of the non-Pappus matroid. These motivate many questions and justify the matroid as an object of interest.

Chapter 3 explores a problem from algebraic statistics in depth: Completely describe the algebraic matroid associated to the variety of rank-1 matrices whose entries sum to one. In Theorem 3.1.14, we prove necessary and sufficient conditions for a subset of entries of a matrix to be completable to a rank-1 matrix with nonnegative entries summing to one. We also turn this into an algorithm for outputting such a matrix – Algorithm 3.3.2 for arbitrary partial matrices and Algorithm 3.3.3 for partial matrices with only two connected components. This gives a rejection criterion for the hypothesis that two discrete random variables are independent given a partial joint probability mass function.

In Chapter 4, we embark on a full algebraic study of a particular chemical reaction network: the shuttle model for the Wnt signaling pathway. The model has not been previously studied, as it was only recently introduced in [37]; therefore, all results about its properties are original. Techniques used in the analysis include symbolic computation, numerical algebraic geometry, polyhedral analysis. We prove that the system has nine solutions generically in Theorem 4.0.3, and that it can achieve as many as three real positive steady states in Theorem 4.3.1. In Proposition 4.5.2, we show that nonzero values for rate parameters and conserved quantities imply nonzero concentrations for all species at steady-state. Original techniques introduced here use the algebraic matroid to find combinatorially nice coordinate parametrizations of the steady-state variety, in Proposition 4.4.3 and to perform parameter inference with only a subset of species data.

Finally, we conclude the dissertation in Chapter 5 with open problems and research directions. Some of these are extensions of topics already discussed in the thesis, and others peek into untapped areas. Algebraic matroids are natural and intriguing objects, and with this dissertation and ongoing research, we aim to contribute to a more widespread discussion of their applications.

## 1.2 Examples

Matroids were introduced as a way to generalize phenomena termed "independence" from across mathematics. Before defining a matroid, it is helpful to describe some of the foundational examples that motivated their invention. In each case, we will describe the associated structures, whose definitions we will formalize in the next section.

**Example 1.2.1** (Linear Matroid)**.** *Let $\{v_1, \ldots, v_4\}$ be a collection of vectors as pictured in Figure 1.1. We let linear independence be the guiding principle here, defining a collection of associated structures. For brevity, we use $E = \{v_1, \ldots, v_4\}$ and $2^E =$ the power set of $E$.*

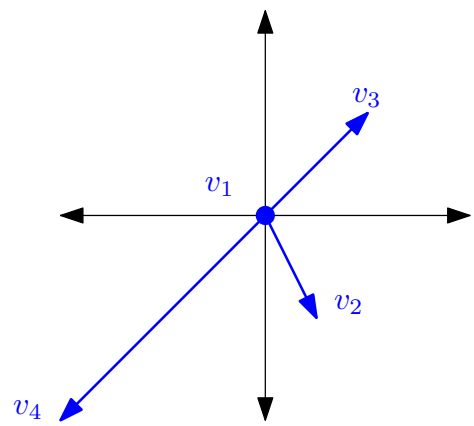



Figure 1.1: Example of linear matroid.

| | | |
|---|---|---|
| *Independent Sets* | $\mathcal{I}$ | *Subsets of $E$ that are linearly independent.*<br>*Here:* $\{\emptyset, \{2\}, \{3\}, \{4\}, \{2,3\}, \{2,4\}\}$. |
| *Bases* | $\mathcal{B}$ | *Vector space bases of the span of $E$; equivalently, maximal linearly independent subsets.*<br>*Here:* $\{\{2,3\}, \{2,4\}\}$ |
| *Dependent sets* | $\mathcal{D}$ | *Subsets of $E$ that have some nonzero linear dependency.*<br>*Here:* $\{\{1\}, \{1,2\}, \{1,3\}, \{1,4\}, \{3,4\}, \{2,3,4\}, \{1,3,4\}, \{1,2,4\}, \{1,2,3\}, \{1,2,3,4\}\}$. |
| *Circuits* | $\mathcal{C}$ | *Minimal subsets of $E$ that are dependent.*<br>*Here:* $\{\{1\}, \{3,4\}\}$ |
| *Hyperplanes* | $\mathcal{H}$ | *Subsets of $E$ that do not span the full space, but after adding any vector, the resulting set does span the full space.*<br>*Here:* $\{\{1,2\}, \{1,3,4\}\}$. |
| *Closure function* | $c$ | *Function from $2^E \to 2^E$ which takes a set of vectors and returns all vectors in their span.* |
| *Rank function* | $\rho$ | *Function $2^E \to \mathbb{N}_0$ which takes a set of vectors and returns the dimension of their span.* |

**Example 1.2.2** (Graphic Matroid)**.** *Let $G = (V, E)$ be an undirected graph, not necessarily simple: in particular, there may be loops or parallel edges. For instance, let $G$ be as pictured in Figure 1.2. The guiding principle in this example is acyclicity. For brevity, we use $E = \{e_1, \ldots, e_5\}$ and $2^E =$ the power set of $E$.*

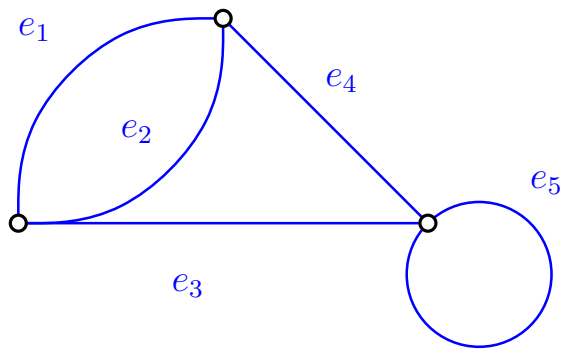



Figure 1.2: Example of graphic matroid.

| | | |
|---|---|---|
| *Independent Sets* | $\mathcal{I}$ | *Forests of $E$.*<br>*Here:* $\{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$. |
| *Bases* | $\mathcal{B}$ | *Spanning trees of $G$, i.e. maximal acyclic edge sets.*<br>*Here:* $\{\{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$ |
| *Dependent sets* | $\mathcal{D}$ | *Subsets of $E$ containing cycles.*<br>*Here:* $\{\{5\}, \{1,2\}, \{1,5\}, \{2,5\}, \{3,5\}, \{4,5\}, \ldots$<br>*Twenty-two subsets of $E$ not in $\mathcal{I}$.* |
| *Circuits* | $\mathcal{C}$ | *Simple cycles of $G$ - i.e. cycles with no repeated vertex.*<br>*Here:* $\{\{5\}, \{1,2\}, \{1,3,4\}, \{2,3,4\}\}$ |
| *Hyperplanes* | $\mathcal{H}$ | *The set of all edges resulting from inducing a subgraph on both parts of a bipartition.*<br>*Here:* $\{\{1,2,5\}, \{3,5\}, \{4,5\}\}$. |
| *Closure function* | $c$ | *Function from $2^E \to 2^E$ which includes an edge if the two endpoints are connected by another path in the input.* |
| *Rank function* | $\rho$ | *Function $2^E \to \mathbb{N}_0$ which returns the size of the largest contained independent set.* |

**Example 1.2.3** (Algebraic Matroid)**.** *Let $K/k$ be a field extension. For this example, we let $k = \mathbb{Q}$ and $K = \mathbb{Q}(x, y)$. Consider the ground set $E$ given by the following elements of $K$:*

$$\alpha_1 = \frac{1}{y-1}, \quad \alpha_2 = x^3 + 5, \quad \alpha_3 = xy^2, \quad \alpha_4 = y - 1$$

*The idea determining the structures in this example is algebraic independence. A set of elements $\{\alpha_1, \ldots, \alpha_s\}$ in the field extension $K/k$ is considered* algebraically independent*, if there exists no nonzero polynomial $P \in k[X_1, \ldots, X_s]$ such that $P(\alpha_1, \ldots, \alpha_s) = 0$.*

| | | |
|---|---|---|
| *Independent Sets* | $\mathcal{I}$ | *Algebraically independent subsets of $E$.*<br>*Here:* $\{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}\}$. |
| *Bases* | $\mathcal{B}$ | *Transcendence bases of $K$.*<br>*Here:* $\{\{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}\}$ |
| *Dependent sets* | $\mathcal{D}$ | *Algebraically dependent sets (i.e. those containing a subset which satisfy a polynomial relation).*<br>*Here:* $\{\{1,4\}, \{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}\}$. |
| *Circuits* | $\mathcal{C}$ | *Minimal algebraically dependent subsets.*<br>*Here:* $\{\{1,4\}, \{1,2,3\}, \{2,3,4\}\}$ |
| *Hyperplanes* | $\mathcal{H}$ | *Subsets of $E$ s.t. adding any of the remaining elements induces a basis.*<br>*Here:* $\{\{1,4\}, \{2\}, \{3\}\}$. |
| *Closure function* | $c$ | *Function from $2^E \to 2^E$ which takes a subset $S \subset E$ and adds any element algebraic over $k(S)$.* |
| *Rank function* | $\rho$ | *Function $2^E \to \mathbb{N}_0$ which takes $S \subset E$ and returns the transcendence degree of $k(S)$.* |

In each of these cases, any one of the listed structures is enough to specify all of the others. They are each also subject to a set of axioms that they must satisfy. In the next section, we will discuss the commonality among these concepts, introducing definitions and concepts that will be used later in the thesis.

## 1.3 Definitions, Axioms, and Notation

Having seen various motivating examples for the definition of a matroid, we will state it formally here. Matroids have many characterizations; we present the axiom systems for independent sets, bases, and circuits. The computation problem tackled in Chapter 2 returns as output the list of bases and circuits of an algebraic matroid. For this reason, having these axioms may be useful.

The following sets of axioms paraphrase [56]:

**Definition 1.3.1** (Independent Set Axioms)**.** *A matroid is an ordered pair* $(E, \mathcal{I})$ *where* $E$ *is a finite set, and* $\mathcal{I} \subset 2^E$*, such that* $\mathcal{I}$ *satisfies:*

1. $\emptyset \in \mathcal{I}$.
2. $I' \subset I$ *and* $I \in \mathcal{I}$ *implies* $I' \in \mathcal{I}$.
3. *(Augmentation Axiom)* $I_1, I_2 \in \mathcal{I}$ *and* $|I_1| < |I_2|$ *implies there exists* $x \in E$ *such that* $I_1 \cup x \in \mathcal{I}$.

**Definition 1.3.2** (Basis Axioms)**.** *A matroid is an ordered pair* $(E, \mathcal{B})$ *where* $E$ *is a finite set, and* $\mathcal{B} \subset 2^E$*, such that* $\mathcal{B}$ *satisfies:*

1. $\mathcal{B}$ *is not empty.*
2. *(Basis Exchange Axiom) If* $B_1, B_2 \in \mathcal{B}$ *are distinct, then there exist* $x \in B_1 \setminus B_2$ *and* $y \in B_2 \setminus B_1$ *such that* $B_1 \setminus \{x\} \cup \{y\} \in \mathcal{B}$.

**Definition 1.3.3** (Circuit Axioms)**.** *A matroid is an ordered pair* $(E, \mathcal{C})$ *where* $E$ *is a finite set, and* $\mathcal{C} \subset 2^E$*, such that* $\mathcal{C}$ *satisfies:*

1. $\mathcal{C}$ *is not empty.*
2. $C' \subseteq C$ *and* $C, C' \in \mathcal{C}$ *implies* $C' = C$.
3. *(Circuit Elimination Axiom) If* $C_1, C_2 \in \mathcal{C}$ *are distinct such that* $x \in C_1 \cap C_2$*, then there exist* $C_3 \subseteq C_1 \cup C_2 \setminus \{x\}$ *such that* $C_3 \in \mathcal{C}$.

As mentioned earlier, every set of axioms can be derived from any other set of axioms. Since this is the case, proving that a given independence structure is a matroid can be accomplished by showing it satisfies any set of axioms. With these axioms established as the definition of a matroid, we can prove that algebraic matroids, the topic of this thesis, are in fact matroids:

**Theorem 1.3.4.** *Let* $E$ *be a finite set of elements in a field extension* $K/k$*, and let* $\mathcal{B}$ *be all subsets that are transcendence bases of* $k(E)$*. Then* $(E, \mathcal{B})$ *is a matroid.*

*Proof.* Since $E$ generates the field extension, it must contain a transcendence basis. As for the basis exchange axiom, let $B_1, B_2$ be transcendence bases of $k(E)$. Then, for any $y \in B_2$, $y$ must be algebraic over $k(B_1)$. This implies it satisfies an algebraic relation $P(y) = 0$ where $P$ has coefficients in $k(B_1)$.

Since $y$ is not algebraic over $k$, at least one other $x_i$ must appear in the coefficients for some $i$. That means that $x_i$ is algebraic over $B_1 \setminus \{x_i\} \cup \{y\}$; since the algebraic closures contains a transcendence basis for the full extension, it must be a transcendence basis as well. $\square$

In Example 1.2.3, an algebraic matroid was presented by giving a collection of elements in a purely transcendental extension. An alternative approach is to first define elements, and then quotient out by the ideal of relations, as in the following definition.

**Definition 1.3.5** (Matroid of a Prime Ideal)**.** *Let $k$ be a field, and $R = k[x_1, \ldots, x_n]$, a polynomial ring.*

*Let $P \subseteq k[x_1, \ldots, x_n]$ be a prime ideal. The ring $S = k[x_1, \ldots, x_n]/P$ is an integral domain, so the function field $K = Frac(S)$ is well-defined.*

*Let $E = \{\overline{x_1}, \ldots, \overline{x_n}\} \subset K$ be the image of $\{x_1, \ldots, x_n\}$ under the composition of the quotient and the injection $\varphi : k \to S \to K$. Independence is defined as usual in an algebraic matroid: algebraic independence over the ground field $k$. $\mathcal{M}(P)$ denotes the matroid obtained from a prime ideal in this manner.*

In fact, every algebraic matroid $\mathcal{M}$ can be obtained as $\mathcal{M}(P)$ for some prime ideal $P$. Start from an algebraic matroid $\mathcal{M}$ of size $n$ with ground set contained in $K/k$. Set the ground set $E$ to be the image of the variables in a ring map $\phi_E : k[x_1, \ldots, x_n] \to K$. The image of the induced map of varieties is an irreducible variety in $k^n$. The associated prime ideal $P$, obtainable by implicitization, satisfies $\mathcal{M} = \mathcal{M}(P)$. Despite this property, it is often convenient to study the matroid of a variety purely in terms of the variety's parametrization. Given a map $\phi : k[x_1, \ldots, x_n] \to K$, the notation $\mathcal{M}(\phi)$ will then refer to the algebraic matroid with ground set $\{\phi(x_i) : i = 1, \ldots, n\}$.

## Decorated Bases and Circuits

We can infuse more of the algebraic structure of the ideal into the matroid via "matroid decorations." This approach of enhancing a matroid with more information has been taken in various forms: oriented matroids [5], arithmetic matroids [9], valuated matroids [12] and matroids over rings [15], to name a few. Circuits of algebraic matroids have a natural decoration, based on the following fact ([25, Lemma 5.6]):

| | | |
|---|---|---|
| $C = \{x_{i_1}, \ldots, x_{i_k}\}$ is a circuit of $\mathcal{M}(P)$ | $\iff$ | The ideal $I \cap k[C]$ is principal with generator $\theta_C$ s.t. $\text{support}(\theta_C) = C$. |

The generator $\theta_C$ of the principal ideal, called a *circuit polynomial* is unique up to unit. This invariant was used by Dress and Lovasz in [11] as well as Lindström in [30] in the process of proving structural facts about algebraic matroids. More recently in [25], the circuit polynomials are studied as objects of interest in their own right. If the polynomial itself is too unwieldy, we may prefer to record only some aspect of the polynomial: (1) The *Newton Polytope* associated to the polynomial. (2) *Top-Degree*: This aspect is explored in [25]. It is the vector in $\mathbb{N}^n$ given by $(\deg_{x_i} \theta_C)_{i \in [n]}$. Equivalently, it is the outer vertex of the tightest bounding box for the Newton polytope. When we try to construct points in a variety based on subsets of the coordinates, the top-degree allows us to determine the cardinality of the solution set. (3) *Degree*: A natural concise invariant.

**Remark 1.3.6.** *A question whose answer is still not fully understood:* What constraints are there for a set of integers attached to the circuits of a matroid to be the degrees of circuit polynomials for some algebraic matroid?

*If $\mathcal{M}$ is a matroid with corank 1, then there is a unique circuit: the full set of variables. Assuming there is more than one variable, we can find an irreducible circuit polynomial of any degree involving the circuit variables. On the other hand, suppose that a matroid has loops. Over an algebraically closed field, the degree of a loop's polynomial is forced to 1, since a polynomial in one variable breaks into linear factors and the polynomial must be irreducible. Over $\mathbb{Q}$, on the other hand, any degree is possible for a loop, since there are irreducible polynomials in every degree. (See Section 2.3 for another example where the field places constraints on the matroid decorations.)*

The bases of an algebraic matroid have the following nice property (cf. [25, Definition 2.6]):

$\{x_{i_1}, \ldots, x_{i_k}\}$ is a base of $\mathcal{M}(P)$ $\iff$ the projection from the variety onto the $i_1, \ldots, i_k$-th coordinates is dominant with generically finite fibers.

The cardinality of the generic fiber (or, equivalently, the degree of the projection) is the decoration on the bases and will be referred to as the *base degree*.

## Matroid Duality

The final matroid property we will discuss in this chapter will be matroid duality. The nature of duality for algebraic matroids is the most well-known open problem about this object [56, Exercise 11.2.3]. Duality of linear matroids will be useful for us in Chapter 2. The construction is as follows:

**Definition 1.3.7.** *Let $\mathcal{M} = (E, \mathcal{B})$ be a matroid. The dual matroid $\mathcal{M}^* = (E, \mathcal{B}^*)$ is defined by*

$$\mathcal{B}^* = \{E \setminus B \ \mid \ \forall B \in \mathcal{B}\}.$$

The basis exchange axiom is inherited by $\mathcal{B}^*$ so the construction is well-defined. The language of duality is appropriate since $\mathcal{M}^{**} = \mathcal{M}$.

There are a number of nice bijections between structures of $\mathcal{M}$ and $\mathcal{M}^*$. Structures associated to $\mathcal{M}^*$ are given the prefix co-. Since they have not been mentioned before, we note that a *spanning set* is a set whose closure is $E$, a loop is a circuit consisting of just one element, and a *bridge* is an element in $\cap_{B \in \mathcal{B}} B$ (Graph theory makes the last two notations intuitive).

| | | |
|---|---|---|
| Base | Cobase | $B \subset E \longleftrightarrow$ Complement of $B$, |
| Independent Set | Cospanning Set | $I \subset E \longleftrightarrow$ Complement of $I$, |
| Hyperplane | Cocircuit | $H \subset E \longleftrightarrow$ Complement of $H$, |
| Bridge | Coloop | $x \longleftrightarrow x$ (identity). |

Obviously, switching the "co"s preserves the bijection.

*Proof.* Since independent sets are those sets contained in some base, the complement contains a cobase, which means it is a cospanning set. The hyperplane is a subset such that adding any element returns a spanning set; on the dual side, removing any element from its complement gives a coindependent set. However, the hyperplane is not a spanning set, so its complement is codependent; therefore, it must be a circuit. Finally, a bridge is an element whose complement is a hyperplane, so the same element is a cocircuit. □

The map from a linear matroid $\mathcal{M}$ to its dual is simple.

**Theorem 1.3.8.** *Let $A$ be a matrix whose columns comprise the ground set of the linear matroid $\mathcal{M}$. Let $K$ be a set of vectors spanning* $\ker(A)$*, and let $A'$ be a matrix whose rows are $K$. The columns of $A'$ is the ground set of a linear matroid identical to $\mathcal{M}^*$.*

*Proof.* The ground set clearly has the right cardinality, and the rank-nullity theorem indicates that the rank will be $|E| - \rho(\mathcal{M})$, the correct rank for the dual.

Suppose $B \subset E$ is a base of the matroid. Then for any $x \in E \setminus B$, there is a vector in the kernel supported on $B \cup x$. Since the row span of $A'$ contains vectors with support $B \cup x$ for each $x \in E \setminus B$, the matrix restricted to $E \setminus B$ must have full rank. So, it must be a cobase of the matroid.

If a subset of the appropriate size is not a base, there is some vector with support $B$ in the row span of $A'$. Row reduce so that one of the rows has support $B$, and only $|E| - \rho(\mathcal{M})$ rows are nonzero. (Row operations do not affect the matroid). Then, the matrix restricted to $E \setminus B$ does not have full rank. □

# Chapter 2

# Computation

In this chapter, we will explore the various options for computing algebraic matroids with base degrees and circuit polynomials. After discussing two primary strategies for the computation, we will demonstrate through several examples of various sizes from different areas of mathematics. This material comes from *Computing Algebraic Matroids* [48].

The two strategies we will outline are the *symbolic algorithm* and the *linear algorithm*. In both approaches, an ideal in a polynomial ring is taken, and a list of bases with base degrees and a list of circuits with circuit polynomial (or alternative decoration like degree) are returned. The linear algorithm allows us to turn the algebraic matroid into a linear matroid by specializing the Jacobian at a generic point of the variety. We define the subvariety where this specialization fails to obtain the correct matroid. We then use the computational tools to describe matroids for: the Plackett-Luce model for statistical rankings of size 4, the mixture model matroid of rank 3 for $4 \times 4$ matrices, the matroid for the steady state variety of the MAP kinase network, the Grassmannian $Gr(3, 6)$, and two realizations of the non-Pappus matroid.

Within each regime, we employ techniques on two different levels: *oracles*, which extract matroid features from the ideal, and *matroid algorithms*, which turn one type of matroid data (e.g. rank) into another type (e.g. circuits). Matroid algorithms are well-studied, and will not be the focus of the chapter, though our software does rely upon them. In most cases, we use naïve matroid algorithms, though we have also implemented sophisticated methods like ReverseSearch [1] and the circuit enumeration algorithm of Boros *et al.* [6] to list bases and circuits respectively. These occasionally perform better than the naïve algorithms; however, in the majority of cases, they do not significantly accelerate the computation. Since the bottleneck of algebraic computation, in the form of Gröbner bases or homotopy continuation, is very time-consuming, only relatively small examples are feasible. At this scale, the difference in complexity for the combinatorial algorithms is not noticeable.

## 2.1 Symbolic Algorithm

In the symbolic realm, elimination is at the core of the computations. It is used in the *rank oracle*, which will be justified by the following proposition:

**Proposition 2.1.1** (Rank Oracle)**.** *Let $k$ be a field, $P \subseteq k[x_1, \ldots, x_n]$ be a prime ideal, $E = \{\overline{x_1}, \ldots, \overline{x_n}\} \subset \mathrm{Frac}(k[x_1, \ldots, x_n]/P)$ be the ground set, and $\mathcal{M}(P)$ be the matroid of this prime ideal as in Definition 1.3.5, and $S \subset E$. Then,*

$$\rho(S) = |S| - \mathrm{height}(P \cap k[S]),$$

*where* height *denotes the height of the ideal.*

*Proof.* The rank of $S$ in the algebraic matroid is given by the transcendence degree of $k(S)$. This is the dimension of the coordinate projection of $\mathcal{V}(P)$ to the space with coordinates in $S$. The ideal $P \cap k[S]$ is the defining ideal of the closure of the image of $\mathcal{V}(P)$ in $k^S$ under the projection. This ideal lives in the smaller polynomial ring $k[S]$; the height of $P \cap k[S]$ as an ideal of this ring is equal to its codimension in the $S$-subspace. Subtracting from $|S|$ gives the dimension. □

Based on this, elimination of $E \setminus S$ followed by computation of height will serve as a rank oracle. Matroid algorithms use the rank oracle to enumerate bases and circuits. For circuits, we may also use the characterization of circuits in Section 1.3 to define a *circuit oracle*: test a set $S$ by first computing the elimination ideal $P \cap k[S]$, and then checking that the generator has full support on the variables of $S$.

The decoration of the circuits is a natural byproduct of the symbolic algorithm. For the base degree, fix a base $B = \{x_{i_1}, \ldots, x_{i_d}\}$. Under the projection map from $k^n \to k^d$ onto the coordinates in the base, the preimage of a generic point $(\lambda_1, \ldots, \lambda_d)$, generated randomly, is a subspace with defining ideal $I_B = \langle x_{i_s} - \lambda_s \mid s = 1, \ldots, d \rangle$. The fiber of the projection then has defining ideal $P + I_B$, and the degree of that ideal is the base degree.

An implementation of the symbolic approach, called `matroids.m2`, written for the commutative algebra platform Macaulay2 [18], can be found at: `http://math.berkeley.edu/~zhrosen/matroids.html`. Important commands are listed in Table 2.1.

The code has two sources of complexity - the complexity of elimination of variables via Gröbner bases, and the combinatorial complexity of listing and testing all potential bases, resp. circuits. For this reason, the code has difficulty with large ground sets, large-rank matroids, and ideals with high degree generators. In trials, the code works quickly for matroids with $|E| \leq 18, \rho(\mathcal{M}) \leq 6$, and generators in degree $\leq 4$. For larger or higher-rank matroids, one should use a more tailored approach, as in Example 2.3.2.

## 2.2 Linear Algebra

A classical result in the study of algebraic matroids states: algebraic matroids defined over a field $k$ of characteristic zero can also be realized as a linear matroid over a field $k(T)$ where

| | |
|---|---|
| `bases` | *Input:* Polynomial Ring, Ideal.<br>*Output:* List of bases of the matroid $\mathcal{M}(I)$. |
| `decoratedBases` | *Input:* Polynomial Ring, Ideal.<br>*Output:* List of bases of the matroid $\mathcal{M}(I)$ with base degree. |
| `circuits` | *Input:* Polynomial Ring, Ideal.<br>*Output:* List of circuits of the matroid $\mathcal{M}(I)$. |
| `pCircuits` | *Input:* Polynomial Ring, Ideal.<br>*Output:* List of ordered pairs: circuits of the matroid $\mathcal{M}(I)$ together with circuit polynomials.<br>Degree-decorated circuits can be computed with:<br>`apply(pCircuits, c -> (c_0, degree(c_1)))`. |
| `topDegree` | *Input:* Polynomial Ring, Polynomial.<br>*Output:* Top-degree vector of the polynomial w.r.t. the variables of the ring.<br>Top-degree decorated circuits can be computed with:<br>`apply(pCircuits(Ring, Ideal), c -> (c_0, topDegree(Ring,c_1)))`. |

Table 2.1: Commands to compute algebraic matroids using `matroids.m2`

$T$ is a finite set of transcendentals ([45, Proposition 6.7.10]). In particular, when $P$ is defined by generators $\langle f_1, \ldots, f_m \rangle \subseteq k[x_1, \ldots, x_n]$, we define the Jacobian matrix $\mathcal{J}(P)$ as:

$$\left(\frac{\partial f_i}{\partial x_j}\right) \; : \; 1 \leq i \leq m, \; 1 \leq j \leq n. \tag{2.1}$$

This matrix, when considered as a matroid with columns as the ground set and linear independence over $\mathrm{Frac}(k[\mathbf{x}]/P)$ defining the independent sets $\mathcal{I}$, represents the **dual matroid** to $\mathcal{M}(P)$. Though the derivatives are computed symbolically in the polynomial ring $k[x_1, \ldots, x_n]$, we then consider linear algebra over the function field of the variety.

When the variety is defined by a parametrization $\phi(t_1, \ldots, t_d) = (g_1(\mathbf{t}), \ldots, g_n(\mathbf{t}))$, we write $\mathcal{J}(\phi)$ for the Jacobian matrix of the following form:

$$\left(\frac{\partial g_j}{\partial t_i}\right) \; : \; 1 \leq i \leq d, \; 1 \leq j \leq n. \tag{2.2}$$

Note that the indices in top and bottom are flipped. This matrix, again setting the columns as $E$ and using linear independence over $\mathrm{Frac}(k[\mathbf{x}]/P)$ to define $\mathcal{I}$, represents $\mathcal{M}(P)$ (not its dual). Since symbolic computation is more costly, certain values of $\overline{x_1}, \ldots, \overline{x_n}$, the ambient coordinates, can be substituted for the variables.

**Definition 2.2.1** (NM-Locus)**.** *Let the* non-matroidal locus $\mathcal{NM}(I)$ *denote the locus of points in* $\mathcal{V}(I)$*, at which the specialization of the Jacobian matrix does* not *represent the dual*

*of the algebraic matroid. Similarly, $\mathcal{NM}(\phi)$ is the set of points in the parameter space where the specialization of the Jacobian matrix does not represent the algebraic matroid.*

This pair of definitions specifies the values that should be avoided when specializing the linear matroid. To help describe the non-matroidal locus, we set the following notation: $I_d(M)$ will refer to the ideal generated by $d \times d$ minors of a matrix $M$. Further, $M\{S\}$ denotes the submatrix of $M$ obtained by restricting to the columns with indices in $S$.

**Proposition 2.2.2.** *Let $V = \mathcal{V}(P)$ be a variety of dimension $d$ in an ambient space of dimension $n$, with Jacobian $\mathcal{J}(P)$ representing the dual of $\mathcal{M}(P)$. Then $\mathcal{NM}(P)$ is defined by the ideal:*

$$I = P + \bigcap_{B \in \mathcal{B}(\mathcal{M})} I_{n-d}(\mathcal{J}(P)\{B\}),$$

*or, equivalently, the intersection of $I_{n-d}(\mathcal{J}(P)\{S\})$ over all $S$ such that $I_{n-d}(\mathcal{J}(P)\{S\}) \not\subseteq P$. In the special case where $\mathcal{J}(P)$ has $n-d$ rows, this is a principal ideal generated by the lcm of all nonzero (mod $P$) maximal minors.*

*Proof.* A matroid is fully described by its list of bases. Given any cobase of $\mathcal{M}(V)$, the corresponding $m \times (n-d)$ matrix has rank $n-d$, so some $(n-d) \times (n-d)$ minor is nonvanishing. The last fact follows from the properties of intersections of principal ideals. □

**Example 2.2.3.** *We compute the non-matroidal locus of a torus in $\mathbb{R}^3$. Let $P = \langle (x^2+y^2+z^2+3)^2 - 16(x^2+y^2) \rangle \subseteq \mathbb{R}[x,y,z]$, the defining ideal for the torus with minor radius 1 and major radius 2. The Jacobian $\mathcal{J}(P)$ is a $1 \times 3$ matrix:*

$$\begin{array}{ccc} x & y & z \\ [4x^3+4xy^2+4xz^2-20x & 4x^2y+4y^3+4yz^2-20y & 4x^2z+4y^2z+4z^3+12z] \end{array}$$

*Since the dual matroid has rank 1, the non-matroidal locus $\mathcal{NM}(P)$ is a principal ideal generated by the lcm of the entries.*

$$\mathcal{NM}(P) = P + \langle -x^5yz - 2x^3y^3z - xy^5z - 2x^3yz^3 - 2xy^3z^3 - xyz^5 + 2x^3yz + 2xy^3z + 2xyz^3 + 15xyz \rangle.$$

*This defines the non-matroidal locus as a subvariety of the torus, and we compute the associated primes:*

$$\langle x, y^2+z^2+4y+3 \rangle, \langle x, y^2+z^2-4y+3 \rangle, \langle y, x^2+z^2+4x+3 \rangle, \langle y, x^2+z^2-4x+3 \rangle,$$

$$\langle z, x^2+y^2-9 \rangle, \langle z, x^2+y^2-1 \rangle, \langle \mathbf{z^2+3, x^2+y^2} \rangle, \langle z+1, x^2+y^2-4 \rangle, \langle z-1, x^2+y^2-4 \rangle.$$

*The boldface ideal has empty real variety, but the other 8 ideals define 8 circles around the torus, four for each generator of the fundamental group. Specializing at any point on those circles gives the wrong matroid for $\mathcal{M}(P)$.*

The corresponding statement for parametrized varieties is given in Proposition 2.2.4.

**Proposition 2.2.4.** *Let $V = \mathcal{V}(\phi)$ be a variety of dimension $d$ with Jacobian $\mathcal{J}(\phi)$ representing $\mathcal{M}(\phi)$. Then $\mathcal{NM}(\phi)$ is defined by the ideal:*

$$I = \bigcap_{B \in \mathcal{B}(\mathcal{M})} I_d(\mathcal{J}(\phi)\{B\}),$$

*or, equivalently, the intersection of all nonzero $I_d(\mathcal{J}(\phi)\{S\})$. This is a principal ideal generated by the lcm of all nonzero maximal minors.*

For the linear algorithm, the goal is to specialize the Jacobian at a point so that we can perform linear algebra to compute the matroid. The propositions imply that selecting a non-root of the ideal $I$ guarantees the desired matroid. From the formulation, it is clear that the non-matroidal locus has positive codimension in the ambient variety; therefore, selection of a generic point is sufficient to guarantee that the NM-locus is avoided.

For parametrized varieties, the Jacobian at a generic point is obtained simply by plugging in random numbers for each parameter. For varieties defined by ideals, a generic point can be computed using numerical algebraic geometry software; we use `Bertini` [3].

**Remark 2.2.5.** *When we select points in the variety numerically, we often need to use very high accuracy. A set of columns may have minors with polynomial values that evaluate to zero when passing to the quotient; however, when we specialize to a point with low accuracy, we may find that the minors corresponding to these columns are $\gg 0$. The required accuracy depends on the degree of the ideal generators or polynomial parametrizations.*

Software that computes linear matroids is then used to transform the matrix into a list of circuits and bases; we use numerical linear algebra in `Sage` [50], as well as its `Matroid` implementation. These lists are translated into a set of $\{0, 1\}$ vectors that are sent back to Bertini. We then use Bertini's `TrackType:5` routine, which carries out coordinate projections to the subspaces specified by the $\{0, 1\}$ vectors. Bertini performs each projection, obtaining the base degree for the list of basis vectors, and the degree or top-degree of the circuit polynomial for the list of circuit vectors. Bertini returns these values using numerical algebraic geometry techniques (see [4] for more details). In this mode of computation, Gröbner basis complexity is avoided; however, combinatorial complexity is still a fixture. The original calculation of the witness set can also be expensive for a high-dimensional variety and ambient space. Examples of code for `Sage` and `Bertini` are included in the website `http://math.berkeley.edu/~zhrosen/matroids.html`, mentioned earlier.

## 2.3 Sample Computations for Applications

In this section, examples from different areas of mathematics will demonstrate that decorated algebraic matroids are natural and provide fundamental insight into the independence structure of a system of distinguished coordinates. As mentioned earlier, this is an approach which has already been explored in matrix completion [26], and which can be applied much more broadly. An application to statistics will be explored in Chapter 3, and matroids as part of an algebraic approach to chemical reaction networks is outlined in Chapter 4. This section is meant as a proof of concept, and includes examples that show the capabilities of our computational techniques.

### Algebraic Statistics

The first area we look into is the field of algebraic statistics. Statistical models have distinguished coordinates describing the probability of an event; the relationship among those coordinates is therefore an obvious and natural question. The decorated algebraic matroid is the way to succinctly describe the independence structure among those probabilities. An in-depth study of a statistical matroid problem appeared in [29], and will be presented as Chapter 3. In this section, we discuss two specific models from [28, 52].

**Example 2.3.1** ($PL_4$ matroid)**.**

$$\boxed{\mathcal{M}(PL_4): \quad |E| = 24, \quad \rho = 4, \quad |\mathcal{B}| = 10560, \quad |\mathcal{C}| = 41346.}$$

*Consider a random variable $X$ which takes as values the permutations of the letters $1, \ldots, n$, which correspond to rankings of a set of preferences. Probability functions $p_\pi : \Theta \to [0,1]$ assign probabilities to each ranking as a function of some parameters $\theta_1, \ldots, \theta_k$.*

*The geometry of the variety defined by the image of the $p_\pi$'s in $[0,1]^{n!}$ is the object of interest. We forget the cube and consider the variety in $\mathbb{C}^{n!}$ for simplicity. In [52, Section 7], the Plackett-Luce model is defined by:*

$$p_\pi \mapsto \prod_{i=1}^{n-1} \frac{1}{\sum_{j=1}^{i} \theta_{\pi(j)}}.$$

*Since algebraic dependence is not changed by reciprocating elements, we instead consider $p_\pi \mapsto \prod_{i=1}^{n-1}(\sum_{j=1}^{i} \theta_{\pi(j)})$ for easier computation. The variety defining the Plackett-Luce model for $n = 4$ is 4-dimensional with degree 27; the corresponding ideal is minimally generated by 9 polynomials of degree 1 and 36 degree-2 polynomials. Since $\mathcal{M}(PL_4)$ has rank 4, the matroid may be represented by an affine representation in 3-space, depicted by its Schlegel diagram in Figure 2.1, made with* `Polymake` *[17].*

*At first glance, this arrangement looks like the vertices of a permutohedron; however, some sets expected to be contained in facets are in fact full-dimensional. The polytope has four hexagonal facets given by fixing the last element of the ranking and acting with $S_3$ on*

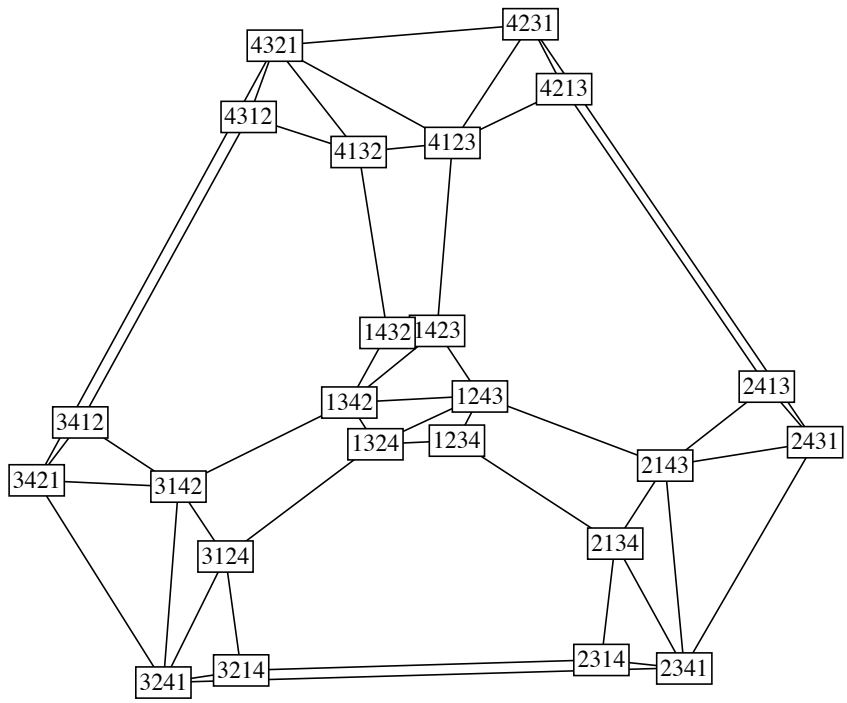
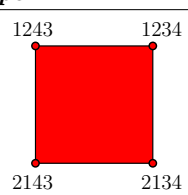
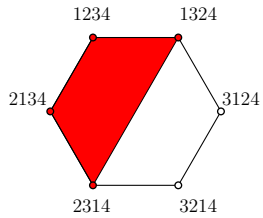
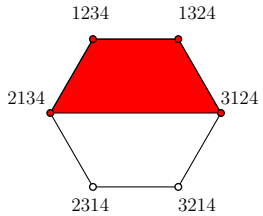
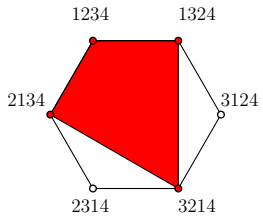
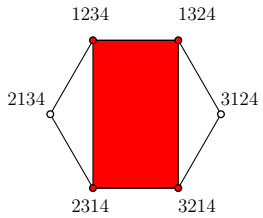



Figure 2.1: Schlegel diagram for the affine representation of $\mathcal{M}(PL_4)$

*the others. The four facets of the permutohedron corresponding to fixing the first element are triangulated. This may be due to the fact that in the parametrization the last element of the ranking does not make an appearance. (e.g.* $p_{1234} \mapsto x_1(x_1+x_2)(x_1+x_2+x_3)$*). The full matroid is too large for computation; instead, we use combinatorial tools in* `Sage` *[50] to find representatives of each base and circuit modulo the natural* $S_4$*-action on the set of variables. We will refer to distinct bases and circuits modulo the group action as* base classes *and* circuit classes*, respectively.*

*1.* Decorated Circuits*: There are five circuit classes of size* 4 *with orbit size* $6, 12, 12, 12, 24$*:*

| ***Type:*** | ***Polynomial:*** | ***Orbit Size:*** |
|---|---|---|
|   | $p_{1243}p_{2134} - p_{1234}p_{2143}$ | *6* |

| | | |
|---|---|---|
|  | $p_{1234}^2 p_{2134} - p_{1234}p_{1324}p_{2134} - p_{1234}p_{2134}^2 - p_{1324}p_{2134}^2 + p_{1234}^2 p_{2314} + p_{1234}p_{2134}p_{2314}$ | *12* |
|  | $p_{1234}^2 p_{1324} - p_{1234}p_{1324}^2 - p_{1324}^2 p_{2134} + p_{1234}^2 p_{3124}$ | *12* |
|  | $p_{1234}^4 - 2p_{1234}^3 p_{1324} + p_{1234}^2 p_{1324}^2 - p_{1234}^3 p_{2134} - p_{1234}^2 p_{1324} p_{2134} + 2p_{1234}p_{1324}^2 p_{2134} + p_{1234}p_{1324}p_{2134}^2 + p_{1324}^2 p_{2134}^2 - p_{1234}^3 p_{3214} - p_{1234}^2 p_{2134} p_{3214}$ | *24* |
|  | $p_{1234}^2 p_{2314} - 2p_{1234}p_{1324}p_{2314} + p_{1324}^2 p_{2314} - p_{1324}p_{2314}^2 + p_{1234}^2 p_{3214} - 2p_{1234}p_{1324}p_{3214} + p_{1324}^2 p_{3214} + p_{1234}p_{2314}p_{3214} + p_{1324}p_{2314}p_{3214} - p_{1234}p_{3214}^2$ | *12* |

*There are 1720 circuit classes of size* 5*, each of which has orbit size* 24*, yielding an additional* 41,280 *circuits. Important to note here: the* `Macaulay2` *computation ran for 10 days without producing circuit polynomials.* `Bertini` *was able to produce a witness set in approximately 8 hours and compute 1720 projections in approximately 6 hours (working in parallel). The degrees of the circuit polynomials are recorded in Figure 2.2.*

2. Decorated Bases*: The 464 classes of size* 4 *that are not circuits are bases. The computation of base decorations produces the distribution of base degrees in Figure 2.3.*

   *The highest base degree is* 24*, which is also the cardinality of the matroid, and the size of the symmetry group. The base of degree 24 is* $\{p_{1234}, p_{2341}, p_{3412}, p_{4123}\}$*. In other words, pick a ranking and apply the 4-cycle to it. The degree of the variety, which tells us the degree of a fiber under generic projection, is* 27*, indicating that all of the coordinate projections are "special."*

*Knowing about the decorated bases and circuits of this matroid allows us to understand its coordinate projections, and gives valuable information about reconstructing partial data.*

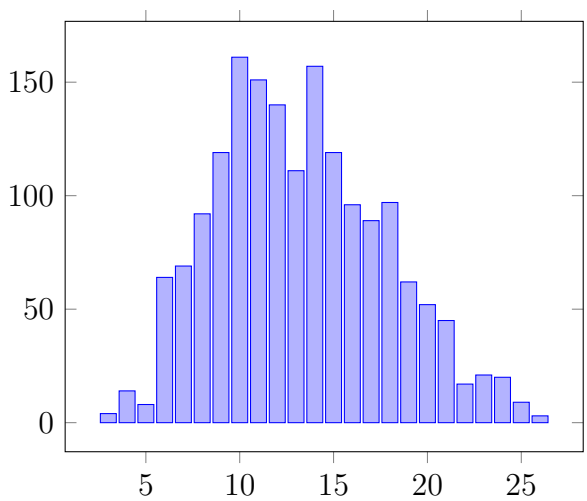



Figure 2.2: Circuit degree frequency for $\mathcal{M}(PL_4)$.

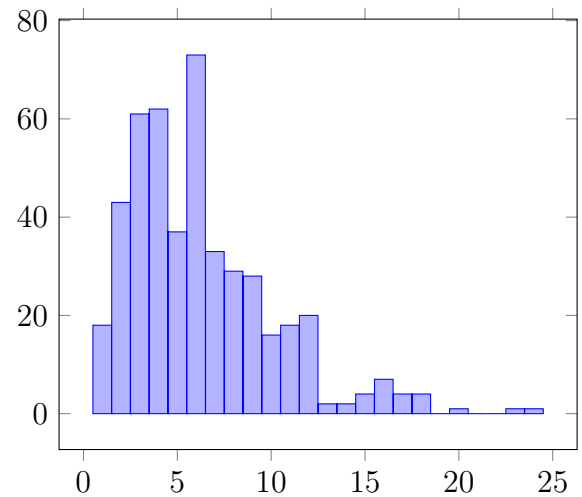



Figure 2.3: Base degree frequency for $\mathcal{M}(PL_4)$.

Another application of matroids to algebraic statistics is in the study of mixture models. The $r$-th mixture model of a pair of discrete random variables $X$ and $Y$, with $m$ and $n$ states respectively, models the situation where $X \perp\!\!\!\perp Y$ conditional on a "hidden" variable $Z$ which occupies $r$ states. In [28], the algebraic boundary of the mixture model for $m = n = 4$ and $r = 3$ is computed; it has 288 components, one of which is analyzed in the example. Studying this example gives insight into the combinatorics of *all* of the components of the variety, in addition to the independence structure of this particular component.

**Example 2.3.2** (Mixture Model Matroid)**.**

$$\boxed{\mathcal{M}(I_{mix}): \quad |E| = 16, \quad \rho = 14, \quad |\mathcal{B}| = 112, \quad |\mathcal{C}| = 11.}$$

*We examine one of the components of the algebraic boundary of the mixture model of rank* 3 *for* $4 \times 4$ *matrices, as defined in [28, Example 5.2]. Let* $I_{mix}$ *denote the defining ideal of this component;* $I_{mix}$ *is generated by the* $4 \times 4$ *minors of the following matrix:*

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & 0 \\ p_{21} & p_{22} & p_{23} & p_{24} & 0 \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{33}(p_{11}p_{22} - p_{12}p_{21}) \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{41}(p_{12}p_{23} - p_{13}p_{22}) + p_{43}(p_{11}p_{22} - p_{12}p_{21}) \end{pmatrix}$$

1. Decorated Bases*: The base enumeration in this case is very quick. There are* 120 *subsets of size* 14*; checking all of them took* 0.5 *seconds to run in* `Macaulay2` *before returning a list of* 112 *bases.*

   *The cobases are the pairs of variables for which the corresponding edge* ***is not*** *one of the 8 edges in Figure 2.4.*

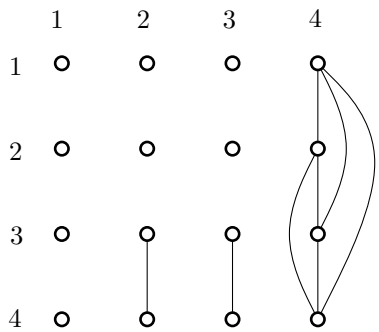



Figure 2.4: Non-cobases of $\mathcal{M}(I_{mix})$.

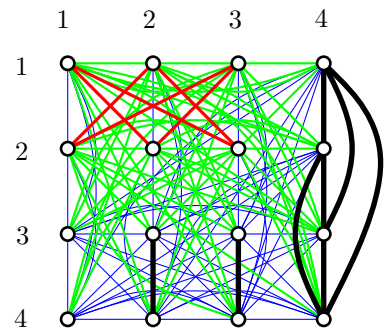



Figure 2.5: Base degrees for $\mathcal{M}(I_{mix})$.

*Computation of base degree yields the following numbers:*

| ***Base Degree*** | 1 | 2 | 3 |
|---|---|---|---|
| ***# of Bases*** | 52 | 54 | 6 |

*The blue edges in Figure 2.5 indicate the complements of degree-*1 *bases, the green edges are the degree-*2 *bases, and the six red edges are the degree-*3 *bases.*

2. Decorated Circuits*: The matroid has* 11 *circuits, which can be read from Figure 2.4 as the complements of each connected component of the graph; this coincidence is because the cohyperplanes have rank one. The computation checking all sets of size* $\leq 11$ *took 3090 seconds. When we started instead from the fundamental circuits associated to some base, and checked mutual eliminations (the Boros et al algorithm), took only 2.7 seconds to produce all circuits, and another 293 seconds to verify that the list was complete. This is one example where, due to high rank, the alternative circuit enumeration algorithm is preferred. The circuit polynomials are too big to record here: the number of terms in each polynomial are 24, 27, 27, 19, 150, 136, 24, 136, 150, 150, and 150, respectively. Instead, we record relevant statistics:*

| ***Circuit Complement*** | ***Circuit Top-Degree*** | ***Circuit Degree*** |
|---|---|---|
| $p_{32}, p_{42}$ | $(2,1,2,1,2,1,2,1,1,0,1,1,1,0,1,1)$ | 6 |
| $p_{41}$ | $(1,2,2,1,1,2,2,1,1,1,1,1,0,1,1,1)$ | 6 |
| $p_{31}$ | $(1,2,2,1,1,2,2,1,0,1,1,1,1,1,1,1)$ | 6 |
| $p_{34}, p_{44}, p_{14}, p_{24}$ | $(2,2,2,0,2,2,2,0,1,1,1,0,1,1,1,0)$ | 6 |
| $p_{22}$ | $(3,1,3,2,2,0,2,2,2,1,2,2,2,1,2,2)$ | 9 |
| $p_{21}$ | $(1,3,3,2,0,2,2,2,1,2,2,2,1,2,2,2)$ | 9 |
| $p_{33}, p_{43}$ | $(2,2,1,1,2,2,1,1,1,1,0,1,1,1,0,1)$ | 6 |
| $p_{11}$ | $(0,2,2,2,1,3,3,2,1,2,2,2,1,2,2,2)$ | 9 |
| $p_{12}$ | $(2,0,2,2,3,1,3,2,2,1,2,2,2,1,2,2)$ | 9 |
| $p_{13}$ | $(2,2,0,2,3,3,1,2,2,2,1,2,2,2,1,2)$ | 9 |
| $p_{23}$ | $(3,3,1,2,2,2,0,2,2,2,1,2,2,2,1,2)$ | 9 |

*These circuit statistics tell us how many completions are possible for every projection.*

*The combinatorial characterization of this* component *also carries information for the* global *structure. Consider the action of $S_4 \times S_4 \times \mathbb{Z}_2$ on the labeled graph of Figure 2.4. The orbit of the graph has cardinality 144. Indeed, in [28, Example 5.2], it is shown that the* 288 *components are paired by taking the transpose of the factorization, and both components in a given pair have the same matroid.*

## Chemical Reaction Networks

In the algebraic study of chemical reaction networks (CRNs), steady-state concentrations of chemical species lie in some algebraic variety. The matroid associated to this variety may be used to design an experiment where measurements of each coordinate are obtained only through some specified costs. Then, bases of the matroid would be appropriate to measure if the goal is to find all concentrations; if we aim to test the validity of the model, a circuit may be a good choice for model rejection. A thorough description of how algebraic matroids inform chemical reaction network analysis may be found in Chapter 4, based on the article *Algebraic Systems Biology: A Case Study for the Wnt Pathway* [20], joint work with Elizabeth Gross, Heather Harrington, and Bernd Sturmfels.

**Example 2.3.3** (MAPK Network)**.**

$$\boxed{\mathcal{M}(I_{MAPK}): \quad |E| = 12, \quad \rho = 3, \quad |\mathcal{B}| = 190, \quad |\mathcal{C}| = 303.}$$

*This ideal comes from [38], which analyzes the polynomials defining the steady-state of a certain CRN. Each variable corresponds to the concentration of some chemical species:*

$$R = \mathbb{R}[KS_{00}, KS_{01}, KS_{10}, FS_{01}, FS_{10}, FS_{11}, K, F, S_{00}, S_{01}, S_{10}, S_{11}].$$

$$\begin{aligned} I_{MAPK} = \langle & a_{00} \cdot K \cdot S_{00} + b_{00} \cdot KS_{00} + \gamma_{0100} \cdot FS_{01} + \gamma_{1000} \cdot FS_{10} + \gamma_{1100} \cdot FS_{11}, -a_{01} \cdot K \cdot S_{01} + b_{01} \cdot KS_{01} + \\ & c_{0001} \cdot KS_{00} - \alpha_{01} \cdot F \cdot S_{01} + \beta_{01} \cdot FS_{01} + \gamma_{1101} \cdot FS_{11}, -a_{10} \cdot K \cdot S_{10} + b_{10} \cdot KS_{10} + c_{0010} \cdot KS_{00} - \alpha_{10} \cdot F \cdot S_{10} + \beta_{10} \cdot \\ & FS_{10} + \gamma_{1110} \cdot FS_{11}, -\alpha_{11} \cdot F \cdot S_{11} + \beta_{11} \cdot FS_{11} + c_{0111} \cdot KS_{01} + c_{1011} \cdot KS_{10} + c_{0011} \cdot KS_{00}, a_{00} \cdot K \cdot S_{00} - (b_{00} + c_{0001} + \\ & c_{0010} + c_{0011}) \cdot KS_{00}, a_{01} \cdot K \cdot S_{01} - (b_{01} + c_{0111}) \cdot KS_{01}, a_{10} \cdot K \cdot S_{10} - (b_{10} + c_{1011}) \cdot KS_{10}, \alpha_{01} \cdot F \cdot S_{01} - (\beta_{01} + \gamma_{0100}) \cdot \\ & FS_{01}, \alpha_{10} \cdot F \cdot S_{10} - (\beta_{10} + \gamma_{1000}) \cdot FS_{10}, \alpha_{11} \cdot F \cdot S_{11} - (\beta_{11} + \gamma_{1101} + \gamma_{1110} + \gamma_{1100}) \cdot FS_{11}, -a_{00} \cdot K \cdot S_{00} + (b_{00} + \\ & c_{0001} + c_{0010} + c_{0011}) \cdot KS_{00} - a_{01} \cdot K \cdot S_{01} + (b_{01} + c_{0111}) \cdot KS_{01} - a_{10} \cdot K \cdot S_{10} + (b_{10} + c_{1011}) \cdot KS_{10}, -\alpha_{01} \cdot F \cdot S_{01} + \\ & (\beta_{01} + \gamma_{0100}) \cdot FS_{01} - \alpha_{10} \cdot F \cdot S_{10} + (\beta_{10} + \gamma_{1000}) \cdot FS_{10} - \alpha_{11} \cdot F \cdot S_{11} + (\beta_{11} + \gamma_{1101} + \gamma_{1110} + \gamma_{1100}) \cdot FS_{11} \rangle. \end{aligned}$$

*The constants $a, b, c, \alpha, \beta,$ and $\gamma$ are taken to be random real numbers, i.e. a set of algebraically independent transcendentals over $\mathbb{Q}$. If the rate constants are originally taken to be part of the matroid, this specialization amounts to matroid contraction. The ideal $I_{MAPK}$ is radical with two associated primes: (1) a variety of degree 10 and dimension 3, and (2) a coordinate subspace with ideal $\langle F, K, FS_{11}, FS_{10}, FS_{01}, KS_{1011}, KS_{01}, KS_{00} \rangle$. In the chemical reaction, the latter component to the steady-state achieved by the disappearance of these reactants. We are interested in the rank* 3 *matroid associated to the former component. A quick symbolic calculation determines that the matroid has affine representation as in Figure 2.6.*

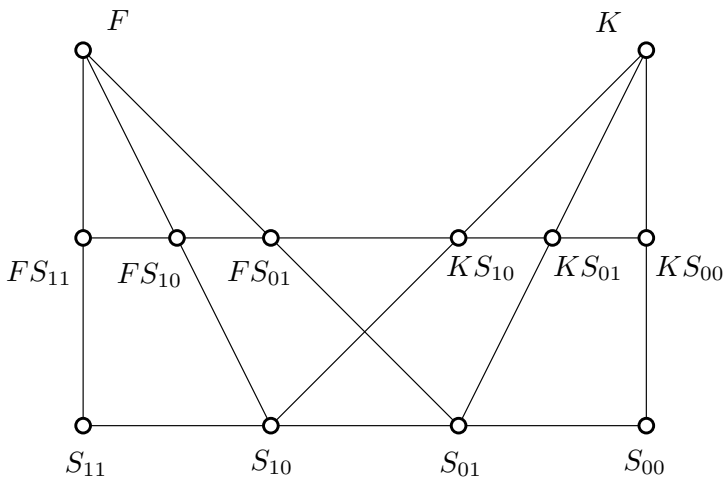



Figure 2.6: Affine representation of the MAP kinase matroid.

1. Decorated Bases*: Any non-collinear set of* 3 *elements from the diagram are a basis of the matroid; there are 190 in total. Of these,* 52 *have base degree* 1*,* 124 *have degree* 2*, and* 14 *have degree* 3*.*

2. Decorated Circuits*: There are circuits of size* 3 *and* 4*. The size* 3 *circuits are the* 30 *collinear sets of* 3*: these have degree* 2 *except for* $\{S_{00}, S_{01}, S_{11}\}$ *and* $\{S_{00}, S_{10}, S_{11}\}$*, which have degree* 3*.*

   *There are* 273 *circuits of size* 4*; the degrees of the circuit polynomials have the following frequencies:*

| ***Circuit Degree*** | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| ***# of Circuits*** | 13 | 76 | 125 | 49 | 10 |

*Possessing this data aids in experimental design, as mentioned above; however, it also distills the combinatorial essence of a chemical reaction network. This demonstrates the power of algebraic matroids in summarizing structure.*

## The Grassmannian

In algebraic geometry and representation theory, some important objects have a distinguished set of coordinates. For the Grassmannian $Gr(r, n)$, the Plücker coordinates are the variables of choice. When $r = 2$, the Grassmannian is defined by skew-symmetric $n \times n$ matrices; this is thematically similar to the content of [25]. We examine the next case of interest: $r = 3$.

**Example 2.3.4.**

$$\boxed{\mathcal{M}(Gr(3,6)): \quad |E| = 20, \quad \rho = 10, \quad |\mathcal{B}| = 184,590, \quad |\mathcal{C}| = 51,005.}$$

*$Gr(3,6)$ is the variety of 3-dimensional subspaces of $\mathbb{C}^6$, with coordinates given by the Plücker coordinates $p_{ijk}$, with $1 \leq i,j,k \leq 6$, distinct. The ideal of the Grassmannian is generated by 35 Plücker relations of degree 2.*

1. Decorated Bases*: The bases are sets of size 10. Computation is aided here by using* `Sage` *to give only one representative of each class up to the $S_6$ action on the Grassmannian. The rest is carried out in 7 seconds by* `Macaulay2`*. There are 197 base classes of degree 1, 42 of degree 2, two of degree 3, and one of degree 7.*

   *The degree-7 base appears to be an outlier, so further examination seems appropriate. The appearing variables correspond to the triangles in the beautifully symmetric complex in Figure 2.7. This image is familiar as a minimal triangulation of $\mathbb{RP}^2$; we plan to explore the connection between this image and high-degree projections of the Grassmannian in future work.*

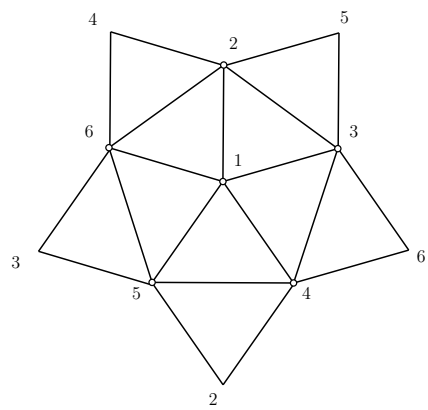



Figure 2.7: Base of $\mathcal{M}(Gr(3,6))$ with base degree 7.

2. Decorated Circuits*: For circuit computation,* `Sage` *once again proved vital in cutting down the number of required tests by a factor of approximately 6!. The testing on the circuit class representatives took 55 seconds in Macaulay2 before returning the list of circuits. There are 97 total circuit classes, with degree of circuit polynomials distributed as in Figure 2.8. Taking into account the full orbits of each circuit, we have 51,005 total circuits. The only circuit class of degree 12 is obtained from our special base by adding one triangle, e.g. the variable $p_{456}$ as in Figure 2.9.*

## Matroid Representations

There is a small collection of algebraic matroids over finite fields that are not representable as linear matroids. Note that base degree is more delicate when working over a finite field

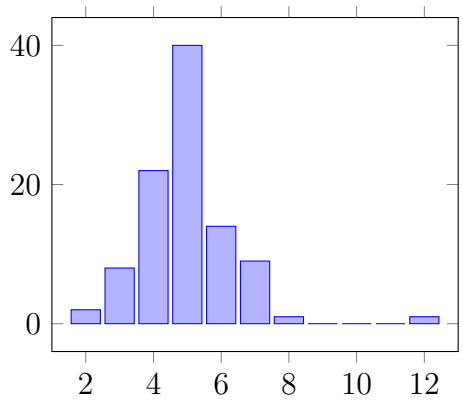



Figure 2.8: Circuit degree frequency for $\mathcal{M}(Gr(3,6))$.

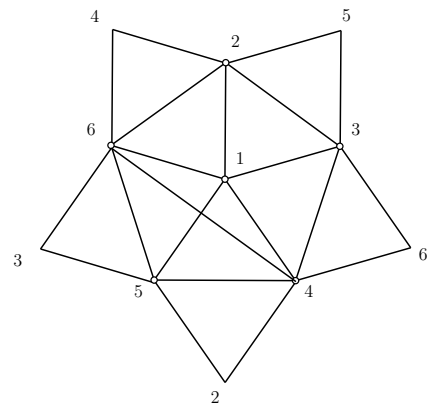



Figure 2.9: Circuit of $\mathcal{M}(Gr(3,6))$ with degree 12.

as the "generic fiber" is not defined; instead, one would consider the degree of the algebraic field extension $k(E)/k(B)$. In any case, computation of the corresponding ideal with circuit polynomials can give insight into the structure of the matroid. One such matroid is explored in the example:

**Example 2.3.5** (Non-Pappus Matroid)**.** $\boxed{\mathcal{M}(I): \quad |E| = 9, \quad \rho = 3, \quad |\mathcal{B}| = 76, \quad |\mathcal{C}| = 86.}$ *The* non-Pappus Matroid *is algebraic over every finite field while not being linearly representable over any field. Since linear representability is equivalent to algebraic representability for fields of characteristic* 0*, this is as extreme as a matroid can be.*

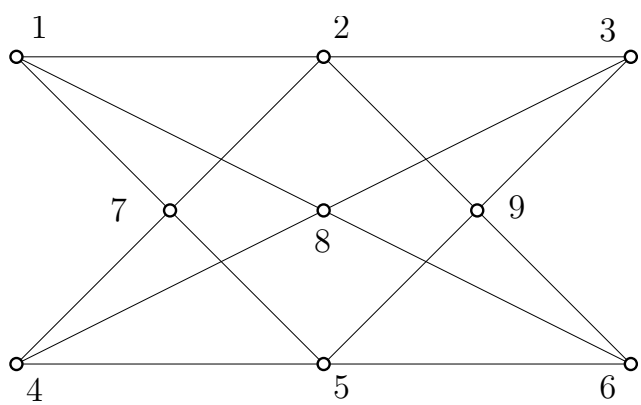



Figure 2.10: Non-Pappus matroid.

*Algebraic Matroid over* $\mathbb{F}_4$ $(\lambda \neq \lambda^2)$:

$$\begin{aligned}
\varphi(1) &= x^2 + y, \\
\varphi(2) &= x, \\
\varphi(3) &= x + y, \\
\varphi(4) &= y + z, \\
\varphi(5) &= y + \lambda z, \\
\varphi(6) &= z, \\
\varphi(7) &= (\lambda - 1)x^2 + \lambda y + \lambda z, \\
\varphi(8) &= x^2 + y + z - z^2, \\
\varphi(9) &= \lambda z - x.
\end{aligned}$$

*Algebraic Matroid over* $\mathbb{F}_2$:

$$\begin{aligned}
\varphi(1) &= x, \\
\varphi(2) &= x + y, \\
\varphi(3) &= y, \\
\varphi(4) &= x + y + \tfrac{xz}{x+y}, \\
\varphi(5) &= z, \\
\varphi(6) &= x + y + \tfrac{yz}{x+y}, \\
\varphi(7) &= xz, \\
\varphi(8) &= xy + \tfrac{xyz}{x+y}, \\
\varphi(9) &= yz.
\end{aligned}$$

*The algebraic representation on the right was used by Lindström in [32] to prove that the non-Pappus matroid is algebraic. The algebraic representation on the left is a valid algebraic representation over* $\mathbb{F}_{p^2}$ *for any p prime, used in [33] to prove an infinite algebraic characteristic set. We can be a bit more precise in assessing these matroid representations, by computing the implicit ideal of each.*

*The representation over* $\mathbb{F}_2$ *has defining ideal generated as:*

$$\langle t_4 + t_5 + t_6, t_1 + t_2 + t_3, t_5t_8 + t_3t_9 + t_5t_9 + t_6t_9, t_3t_7 + t_2t_9 + t_3t_9,$$
$$t_2t_7 + t_6t_7 + t_2t_9 + t_5t_9 + t_6t_9, t_3t_5 + t_9, t_2t_5 + t_7 + t_9, t_3^2 + t_3t_6 + t_8 + t_9, t_2^2 + t_2t_6 + t_9\rangle.$$

*Compiling the degrees of the circuit polynomials, we have:*

| ***Degree*** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| ***# of Circuits*** | 2 | 33 | 24 | 21 | 4 | 0 | 2 |

*The representation over* $\mathbb{F}_4$ *has defining ideal generated as:*

$$\langle t_4 + t_5 + (\lambda + 1)t_6, t_3 + t_5 + t_9, t_2 + \lambda t_6 + t_9, t_1 + \lambda t_5 + \lambda t_7,$$
$$t_9^2 + \lambda t_5 + t_6 + (\lambda + 1)t_7 + (\lambda + 1)t_8, t_6^2 + \lambda t_5 + t_6 + \lambda t_7 + t_8\rangle.$$

*The degrees of the circuit polynomials appear with the following frequency:*

| ***Degree*** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| ***# of Circuits*** | 12 | 59 | 0 | 15 |

*Further examination of the* decorated *algebraic matroid may give insight into the various possible representations of this and similar nonlinear matroids.*

With these examples, we have demonstrated that computation of algebraic matroids is feasible, even in quite large examples. The statistics and biology applications show that we can completely describe matroids whose interest comes from real-world completion problems; we will see more matroids of this type in Chapters 3 and 4. The results for the Grassmannian and the non-Pappus matroid present examples where our computations raise interesting questions in pure mathematics.

# Chapter 3

# Statistics: Joint Probability Matroid

This chapter comes mostly from the paper *Matrix Completion for the Independence Model* [29], joint with Kaie Kubjas. The problem comes from statistics and asks a question with a very matroidal flavor. Our treatment does not rely on matroid theory, but in Section 3.4, we make the underlying matroids explicit. The matroid perspective organizes our answer to the statistics question, helps explain why generalization will be difficult, and leads us to more questions about the geometry of algebraic matroids. The question coming from statistics is as follows:

**Problem 3.0.1.** *Given some entries of a matrix, is it possible to add the missing entries so that the matrix has rank* 1*, is nonnegative, and its entries sum up to one?*

The answer to this question takes many different forms. For example, as we shall prove later, the partial probability matrix

$$\begin{pmatrix} 0.16 & & & \\ & 0.09 & & \\ & & 0.04 & \\ & & & 0.01 \end{pmatrix}$$

has a unique completion:

$$\begin{pmatrix} 0.16 & 0.12 & 0.08 & 0.04 \\ 0.12 & 0.09 & 0.06 & 0.03 \\ 0.08 & 0.06 & 0.04 & 0.02 \\ 0.04 & 0.03 & 0.02 & 0.01 \end{pmatrix}.$$

On the other hand, perturbing any entry of the original matrix by $\epsilon > 0$ makes the matrix have no eligible completions, and perturbing any entry by $\epsilon < 0$ introduces an infinite number of completions.

The motivation for studying Problem 3.0.1 comes from statistics. Let $X$ and $Y$ be two discrete random variables with $m$ and $n$ states respectively. Define the probability matrix:

$$P = (p_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}, \quad \text{where } \; p_{ij} = Pr(X = i, Y = j).$$

For any probability matrix $P$, we have $p_{ij} \geq 0$ for all $i, j$ and $\sum p_{ij} = 1$. We say that the random variables $X$ and $Y$ are independent if

$$Pr(X = i, Y = j) = Pr(X = i) \cdot Pr(Y = j)$$

for all $i, j$. This can be translated into the statement

$$P = \begin{pmatrix} Pr(X = 1) \\ Pr(X = 2) \\ \vdots \\ Pr(X = m) \end{pmatrix} \begin{pmatrix} Pr(Y = 1) & Pr(Y = 2) & \cdots & Pr(Y = n) \end{pmatrix}.$$

Hence, the probability matrix $P$ of two independent random variables has rank 1, is nonnegative, and its entries sum to one.

Suppose that the probabilities $Pr(X = i, Y = j)$ are measurable only for certain pairs $(i, j)$. A situation in which this might arise in applications is a pair of compounds in a laboratory that only react when in certain states. A complete answer to Problem 3.0.1 will allow us to reject a hypothesis of independence of the events $X$ and $Y$, based only on this collection of probabilities.

For each type of partial matrix, we derive an inequality which is satisfied if and only if the partial matrix can be completed to a rank-1 probability matrix. This main result is derived in Theorem 3.1.14. In Theorem 3.2.3, we generalize this characterization to diagonal partial tensors which can be completed to rank-1 probability tensors, i.e. rank-1 tensors whose entries are nonnegative and sum to one. Rank-1 probability tensors correspond to joint probabilities of independent random variables $X_1, X_2, \ldots, X_n$. An entry $p_{i_1 i_2, \ldots i_n}$ of a probability tensor expresses the joint probability $Pr(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$.

We design Algorithm 3.3.2 for checking completability of partial matrices to rank-1 probability matrices. Moreover, we will explain how to construct desired completions. In case there is more than one desired completion, we will show how to use Lagrange multipliers to find a completion that minimizes a distance measure to a fixed probability distribution.

Problem 3.0.1 is a variation on the well-studied problem of low-rank matrix completion. Király *et al* [26] introduced algebraic matroid techniques for matrix completion problems. In Section 3.4, we will study the algebraic matroids that arise in the context of completions of probability matrices. The problem has an immediate generalization to probability matrices of higher rank: We can ask if a partial matrix can be completed to a probability matrix of (nonnegative) rank $r$. Nonnegative rank $r$ probability matrices express joint probabilities of random variables $X$ and $Y$ that are independent given a hidden random variable $Z$ with $r$ states. However, to study nonnegative rank $r$ probability completions, one has to consider rank-$r$ probability completions first. We will explore an example for matrices of rank 2.

## Outline

In Section 3.1, we derive, for each type of partial matrix, an inequality which is fulfilled if and only if the partial matrix is completable to a rank-1 probability distribution. Our

discussion starts with diagonal probability masks, before moving on to general probability masks. Theorem 3.1.2 and Theorem 3.1.14 characterize when a diagonal, respectively a general, partial matrix is completable.

Partial matrices which can be completed to rank-1 probability matrices form a semialgebraic set, see Proposition 3.2.1. In Section 3.2, we will study the semialgebraic description of completable partial matrices and tensors. In Theorem 3.2.3, we will derive a characterization of diagonal partial tensors which can be completed to rank-1 probability tensors, and in Proposition 3.2.4, we study the polynomial inequalities defining this semialgebraic set.

In Section 3.3, we use results in Section 3.1 to define an algorithm which checks completability. We then present an algorithm to recover the $\leq 2$ possible solutions when the solution set is finite. Afterwards, we show how to construct a completion if there are infinitely many completions. We will use Lagrange multipliers to construct a rank-1 probability completion which maximizes or minimizes a certain function, e.g. the distance from the uniform distribution or the probability of a particular state.

In Section 3.4, we examine the algebraic matroids arising from this problem, following the approach of Király *et al* [26] for low-rank matrix completion. Finally, in Section 3.5, we study generalization of our results to higher rank probability matrices and probability tensors.

Implementations of algorithms can be found on

`math.berkeley.edu/~zhrosen/probCompletion.html`

The following notation will be used throughout the chapter:

| Notation | Definition |
|---|---|
| $\Delta^n$ : | Standard $n$-simplex $\{\mathbf{x} \in \mathbb{R}^{n+1} : \sum x_i = 1$ and $x_i \geq 0$ for all $i\}$ |
| $\Delta_0^n$ : | $n$-simplex as a corner of the $n$-cube $\{\mathbf{x} \in \mathbb{R}^n : \sum x_i \leq 1$ and $x_i \geq 0$ for all $i\} = \mathrm{conv}(\{0\} \cup \Delta^{n-1})$ |
| $\Pi_{m\times n}$ : | Ideal of algebraic relations among entries of a rank-1 matrix with entries summing to 1. |
| $\mathcal{V}(\Pi_{m\times n})$ : | Variety of rank-1 matrices with entries summing to 1. |
| $\pi_S$ : | Probability mask. Projection of $\mathbb{R}^{m\times n}$ onto the coordinates indexed by $S$. Summarized by 0-1 matrix with 1's for coordinates in the image, and 0's for coordinates in the kernel. |
| $\pi_S(M)$ : | Partial matrix. The image of a matrix $M$ under the probability mask $\pi_S$. Summarized by matrix with values in the coordinates indexed by $S$ and blanks elsewhere. |

The completion problem can be restated as a question about geometry.

**Remark 3.0.2.** *The set of rank-*1 *matrices in* $\mathbb{R}^{m\times n}$ *with entries summing to* 1 *is an algebraic variety* $\mathcal{V}(\Pi_{m\times n})$*. The projection map* $\pi_S : \mathcal{V}(\Pi_{m\times n}) \to \mathbb{R}^{|S|}$ *takes a rank-*1 *matrix with entries summing to 1 and returns some subset of its entries. To move backwards from that subset to the full rank-*1 *matrix summing to* 1*, i.e. matrix completion, amounts to computing*

*the fiber of the projection, in particular* $\pi_S^{-1}(M) \cap (\mathcal{V}(\Pi_{m\times n}) \cap \Delta^{mn-1})$. *The intersection with the simplex restricts our attention to matrices with nonnegative entries.*

## 3.1 Completability of Partial Probability Matrix

### Diagonal Partial Matrices

The simplest case to analyze is the set of partial matrices with entries visible along the diagonal. For a $1 \times 1$ matrix, this is trivially completable, indeed completed, if and only if the observed entry is 1. For a $2 \times 2$ matrix, there is more to consider.

**Example 3.1.1.** *Let $M$ be the partial probability matrix given by:*

$$M = \begin{pmatrix} a & \\ & b \end{pmatrix}.$$

*In order for the matrix to be completed, both the rank 1 requirement and the summing to 1 must be addressed. First, for rank 1, the off-diagonal entries are set to $x$ and $ab/x$, then the quantity $a + ab/x + x + b$ is set equal to 1. The equivalent quadratic equation is $x^2 + (a + b - 1)x + ab = 0$. In order for a real solution for $x$ to exist, the discriminant must be greater than or equal to 0, i.e.*

$$(a + b - 1)^2 - 4ab \geq 0.$$

*This inequality, along with the requirement that $a + b \leq 1$ and both $a$ and $b$ are greater than or equal to 0, is necessary and sufficient to guarantee that $x$ gives a completion in $\Delta^3$, see Figure 3.1.*

For $n > 2$, we take advantage of the factorization of rank-1 matrices as products of vectors to obtain the following more general result:

**Theorem 3.1.2.** *Let $M$ be an $n \times n$ partial probability matrix, where $n \geq 2$, with nonnegative observed entries along the diagonal:*

$$M = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix}.$$

*Then $M$ is completable if and only if $\sum_{i=1}^n \sqrt{a_i} \leq 1$, or equivalently, $||(a_1, \ldots, a_n)||_{1/2} \leq 1$. In the special case $\sum_{i=1}^n \sqrt{a_i} = 1$, the partial matrix $M$ has a unique completion.*

*Proof.* Recall that a rank-1 $n \times n$ probability matrix can be factored as $\mathbf{u}^T\mathbf{v}$ for $\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^n$ where the sum of the entries in each vector is 1. For this problem, consider all possible

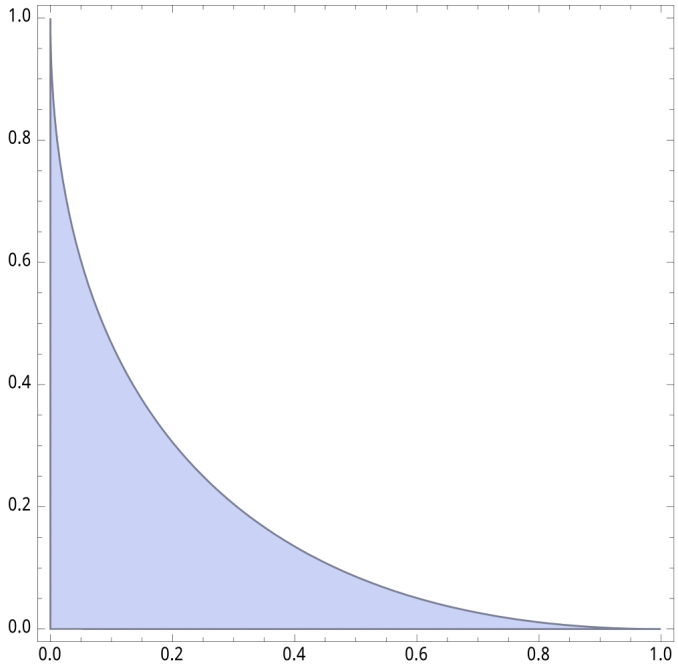



Figure 3.1: Completable probability masks of $2 \times 2$ matrices with diagonal entries observed.

values of $\mathbf{u}$ in $\Delta^{n-1}$, but do not restrict the values of $\mathbf{v}$ to the simplex. Instead, let $\mathbf{v}$ be formulated in terms of $\mathbf{u}$ and the entries of the matrix. Explicitly, we have

$$\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \begin{pmatrix} a_1/u_1 & \cdots & a_n/u_n \end{pmatrix} = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix}.$$

A probability completion will arise when $\sum v_i = 1$. Here we assume $a_i > 0$ and thus $u_i > 0$ for all $i$. We will consider the case when $a_i = 0$ for some $i$ separately.

Let $f(\mathbf{u}) = \sum_{i=1}^n v_i = \sum_{i=1}^n a_i/u_i$ denote this quantity and compute the minimum of $f$ on the simplex. For this computation, consider $u_1, \ldots, u_{n-1}$ as independent variables and $u_n = 1 - \sum u_i$. We thus consider the following function on the simplex $\Delta_0^{n-1}$.

$$f = \left( \sum_{i=1}^{n-1} \frac{a_i}{u_i} \right) + \frac{a_n}{1 - \sum_{i=1}^{n-1} u_i}$$

$$\Rightarrow \frac{\partial f}{\partial u_i} = -\frac{a_i}{u_i^2} + \frac{a_n}{(1 - \sum_{i=1}^{n-1} u_i)^2}$$

Setting $\partial f / \partial u_i = 0$ for all $i$ implies $a_i/u_i^2 = k$ is constant for all $i$. Since $\mathbf{u}$ is in the simplex, we have $k$ equal to $(\sum \sqrt{a_i})^2$. The value of $f$ (i.e. the sum of $v_i$) at this point is $(\sum \sqrt{a_i})^2$. If

this is $\leq 1$, continuity of $f$ implies that a completion exists somewhere between our minimum and the boundary, because within an $\epsilon$ of the boundary of $\Delta^{n-1}$, we have

$$\sum v_i = \frac{a_1}{u_1} + \cdots + \frac{a_n}{u_n} \gg 1.$$

If this minimum value is $> 1$, the function will not achieve 1 anywhere in the simplex, so no completion is possible.

Now assume $a_i = 0$ for $i \in I$. If $|I| \geq n-1$, then the statement of the theorem is clearly satisfied. Assume $|I| \leq n-2$. If $\sum_{i\in[n]\setminus I} \sqrt{a_i} \leq 1$, then the probability mask with the rows and columns in $I$ removed has a rank 1 probability completion, and it can be extended to a rank 1 completion of the original matrix by replacing the entries in the removed rows and columns with zeros. On the other hand, a completion of the original matrix gives a completion of the reduced matrix with sum of entries $\leq 1$. By continuity, the reduced matrix also has a completion with sum of entries equal to 1. Hence $\sum_{i\in[n]\setminus I}^{n} \sqrt{a_i} \leq 1$.

Finally, when equality is attained, the solution must be unique, since $(\sum \sqrt{a_i})^{-1}(\sqrt{a_1}, \ldots, \sqrt{a_n})$ is unique as a minimum in the simplex. □

**Corollary 3.1.3.** *Let $\sum_{i=1}^{n} \sqrt{a_i} < 1$. For $n = 2$, the probability mask $M$ has exactly two completions. If $n > 2$, then the semialgebraic set of completions of $M$ is $(n-2)$-dimensional.*

*Proof.* If $M$ contains no zeros, then every path from $(\sum \sqrt{a_i})^{-1}(\sqrt{a_1}, \ldots, \sqrt{a_n})$ to the boundary of the simplex will contain exactly one completion. For $\mathbf{u} \in \Delta^1$, there are only two distinct paths to the boundary. For higher-dimensional simplices, this will induce a codimension 1 set inside the $(n-1)$-dimensional simplex, which gives an $(n-2)$-dimensional set.

If $M$ has zeros, let $I = \{i : a_i = 0\}$ denote the indices of zero observed entries. The set of completions of $M$ is

$$\bigcup_{R,C\subseteq I: R\cup C = I} \{\text{completions of } M \text{ with rows in } R \text{ and columns in } C \text{ zero}\}.$$

By the previous discussion, the dimension of the semialgebraic set of completions with the sum of entries $\leq 1$ of the $(n-|I|) \times (n-|I|)$ submatrix of $M$ corresponding to nonzero observed entries is $(n-|I|-1)$. For each such completion, we can freely fix all but one of the row sums $u_i$ with $i \in I\setminus R$ and columns sums $v_j$ with $j \in I\setminus C$: Without loss of generality assume $|C| \leq |R|$. Row sums $u_i$ with $i \in I\setminus R$ determine all row sums. The nonzero observed entries together with row sums determine column sums in columns $[n]\setminus I$. All but one of the rest of the columns sums can be chosen freely such that they sum to one. Hence the dimension of the set of completions corresponding to $R$ and $C$ is $(n-|I|-1) + (|I|-|R| + |I|-|C|-1) \leq (n-|I|-1) + (|I|-1) = n-2$. The equality is obtained for any $R$ and $C$ with $|R| + |C| = |I|$. □

**Remark 3.1.4.** *The analysis of Theorem 3.1.2 works to derive the constraint for the* $2 \times 2$ *diagonal probability mask as in Example 3.1.1:*

$$\sqrt{a} + \sqrt{b} \leq 1 \Leftrightarrow a + b + 2\sqrt{ab} \leq 1 \Leftrightarrow 2\sqrt{ab} \leq 1 - a - b$$
$$\Leftrightarrow 4ab \leq (1 - a - b)^2 \text{ and } 0 \leq 1 - a - b.$$

**Example 3.1.5.** *The matrix* $\mathrm{diag}(1/4, 1/4, 1/4)$ *does not have a completion to a rank-*1 *probability matrix, since* $3\sqrt{1/4} = 3/2$. *The set of* $3 \times 3$ *diagonal partial matrices that are completable to rank-*1 *probability matrices is shown in Figure 3.2.*

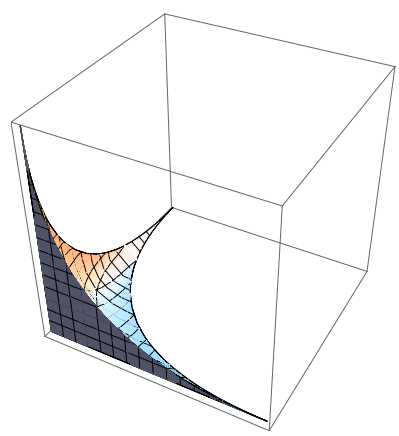


Figure 3.2: Completable probability masks of $3 \times 3$ matrices with diagonal entries observed.

## Block Diagonal Matrices

In this section, we will use our result about diagonal partial matrices to prove completability conditions for different types of partial matrices with increasing generality:

$$\text{Diagonal} \implies \text{Block Diagonal} \implies \text{Acyclic} \implies \text{Feasible}.$$

To discuss non-diagonal masks, we introduce graph notation used in various places in the matrix completion literature.

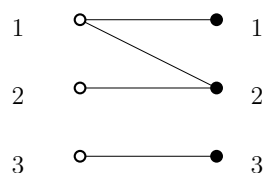
**Definition 3.1.6.** *Let* $M = \mathbf{u}^T\mathbf{v}$ *be an* $m \times n$ *matrix of probabilities whose entries sum to* 1, *and let* $\mathbf{u}$ *and* $\mathbf{v}$ *be vectors in* $\Delta^{m-1}$ *and* $\Delta^{n-1}$ *respectively. A bipartite graph can be associated to* $M$ *in the following way:*

| ***Graph*** | ***Matrix*** |
|---|---|
| *White vertex* $r_i$ | $i$*-th row* |
| *Black vertex* $c_j$ | $j$*-th column* |
| *Edge* $(r_i c_j)$ | $(i, j)$*-th entry* |
| *Weight* $\omega(r_i)$ | *Sum of* $i$*-th row, or* $u_i$ |
| *Weight* $\omega(c_j)$ | *Sum of* $j$*-th column, or* $v_j$ |
| *Edge weight* $\omega(r_i c_j)$ | *Value of entry* $m_{ij}$ |

*The bipartite graph associated to a probability mask is the graph obtained by deleting the edges corresponding to unobserved entries, and omitting vertex weights.*

**Example 3.1.7.** *On the left is a partial probability matrix with entries, and on the right is the corresponding bipartite graph.*

$$\begin{pmatrix} x_{11} & x_{12} & \\ x_{21} & & \\ & & x_{33} \end{pmatrix}$$




Note that in this formulation, the question of completability is equivalent to the existence of a vertex labeling so that the black vertex weights and white vertex weights each sum to 1, and the edge weights satisfy $\omega(r_i c_j) = \omega(r_i)\omega(c_j)$. The diagonal case describes those masks whose graphs are the union of disjoint edges. We now consider more general bipartite graphs.

**Lemma 3.1.8.** *Let $G$ be a bipartite graph with a connected component $K_{p,q}$, with edge weights $a_{1,1}, \dots a_{p,q}$, so that the corresponding submatrix is rank 1. Let $H$ be the graph $(G \backslash K_{p,q}) \cup K_{1,1}$, with the edge weight on $K_{1,1}$ given by*

$$\omega(a_{1,1}) = \sum_{i=1}^{p} \sum_{j=1}^{q} a_{i,j}.$$

*Then completability of $G$ is equivalent to completability of $H$.*

*Proof.* ($\Rightarrow$) Begin with a vertex weighting on $G$. Replace the white vertices weighted $u_1, \dots, u_p$ with a single white vertex weighted $\sum_{i=1}^{p} u_i$ and the black vertices weighted $v_1, \dots, v_q$ by a single black vertex labeled $\sum_{i=1}^{q} v_i$. Since $K_{p,q}$ was disconnected from the other vertices, no other observed entries will be changed by this replacement.

($\Leftarrow$) Begin with a vertex weighting on $H$. The fact that $a_{1,1}, \dots, a_{p,q}$ form a rank-1 $p \times q$ matrix implies that there is a rank-1 factorization $\mathbf{u}'^T \mathbf{v}'$. Scale the vector $\mathbf{u}'$ with the constant $u_1 / \sum u'_i$; and scale $\mathbf{v}'$ by the inverse. The resulting submatrix is the same but now $\sum u'_i = u_1$ and $\sum v'_i = v_1$. The new vertex weights give a completion of $G$. □

This lemma establishes that Theorem 3.1.2 applies to block diagonal matrices as well. To extend to general acyclic matrices, we need to make a definition to account for exceptional cases involving zeros:

**Definition 3.1.9.** *Let $M$ be a partial matrix, with corresponding graph $G$. We say that $M$ is* prunable *if there is a set of vertices $W \subset V(G)$ such that every edge labeled zero is adjacent to some $w \in W$, but no edge with nonzero label is adjacent to any $w \in W$.*

Pruning *refers to removing $W$ and all incident edges from the graph, or equivalently, considering only the induced subgraph on $V \setminus W$. A* careful pruning *takes $W$ so that the remaining graph has the largest possible number of components.*

**Proposition 3.1.10.** *A partial matrix is completable to a rank-1 probability matrix only if it is prunable.*

*Proof.* Let $W$ be the set of vertices whose incident edges are all labeled zero. Suppose some $e \in E(G)$ labeled zero is not adjacent to any $w \in W$. Then, if $e = (r_i, c_j)$, both $r_i$ and $c_j$ are connected to some other vertices $c_k$ and $r_l$ with nonzero edges. The $2 \times 2$ minor of $M$ defined by rows $i, l$ and columns $j, k$ then has two nonzero entries along one diagonal and zero in the other. Therefore, the matrix $M$ cannot be rank one. □

The next lemma allows us to confine our conversation to matrices with nonzero entries:

**Lemma 3.1.11.** *Let $G$ be a bipartite graph with a vertex $v$ such that all edges incident to $v$ have weight* 0*. In particular, the vertex $v$ might be an isolated vertex. Completability of $G$ is equivalent to completability of $G \setminus v$, except when $G \setminus v$ is connected.*

*Proof.* ($\Rightarrow$) Assume that the probability mask $M$ corresponding to $G$ is completable. Remove the row or column corresponding to $v$. Contract rows and columns corresponding to each completable block to one row and column respectively. Add the corresponding entries in a completion of $M$. After permuting rows and columns, we get a diagonal probability mask that has a nonnegative completion with the sum of entries $\leq 1$ (by modifying the completion of $M$ in the same way as we have modified $M$). By the proof of Theorem 3.1.2, the diagonal probability mask is completable as long as $n > 1$. By Lemma 3.1.8, also the probability mask corresponding to $G \setminus v$ is completable. In the case where $G \setminus v$ is connected, this reasoning does not hold, since the sum must be $= 1$.

($\Leftarrow$) Begin with a probability matrix $M$ which is a completion of $G \setminus v$. Simply add in a row or column of zeros corresponding to $v$ to obtain the desired completion of $G$. □

**Example 3.1.12.** *To illustrate why the exception in Lemma 3.1.11 is necessary, consider the following partial matrix:*

$$\begin{pmatrix} 0.15 & & \\ 0.05 & 0.1 & \\ & 0 & 0 \end{pmatrix}$$

*If we prune off the bottom row, the remaining matrix is completable; however, if we prune off the last column, the matrix has only one rank-*1 *completion and its entries do not sum to* 1*.*

We say that a block of a matrix is 1-closable (as in [26]), if the corresponding graph is a spanning tree. If all labels in a spanning tree are nonzero, then the rest of the entries in the block can be completed using fundamental cycles (see Section 3.3 for details). So, Lemma 3.1.11 will allow us to prune a partial matrix to the form where it can be completed to a block diagonal partial matrix using cycles. The result from the discussion so far is that for matrices that can be pruned to acyclic graphs with nonzero edge labels and $n > 1$ component, the question is reduced to the problem solved in Section 3.1.

In order to allow cycles in the graph, we recall that a universal Gröbner basis for the ideal of relations among entries in a rank-1 matrix is indexed by the set of cycles of the bipartite

graph (this follows from Theorems 4.11 and 8.11 in [51]); see Sections 3.3 and 3.4 for more detail. If the cycles in the graph satisfy these relations, they admit a rank-1 completion. We summarize all of our conditions in the following definition:

**Definition 3.1.13.** *A matrix $M$ is said to be* feasible*, if $M$ is prunable, and cycles in the graph of $M$ satisfy the binomial relations from the universal Gröbner basis.*

**Theorem 3.1.14.** *Let $M$ be a feasible partial probability matrix such that after careful pruning, its graph $G$ has $s$ connected components. Let $b_i$ be the sum of the weights in the $i$-th connected component of $G$* ***after rank-1 closure****. If $s = 1$, then $M$ is completable if and only if $b_1 = 1$. For $s > 1$, the partial matrix $M$ is completable to a probability matrix if and only if:*

$$\sum_{i=1}^{s} \sqrt{b_i} \leq 1.$$

*Proof.* By Lemma 3.1.11, if the pruned matrix has more than one component, completability of the pruned matrix is equivalent to completability of the original. In the exceptional case where every choice of vertices leaves one component, completability of the pruned matrix is also equivalent to the original. This is because every row and column removed was forced to be all zeros so cannot contribute to the total probability.

On the remaining nonzero matrix, blocks are completed using cycles, and then contracted from a block diagonal matrix to a diagonal matrix; this does not change completability by Lemma 3.1.8. Finally, we use Theorem 3.1.2. □

**Example 3.1.15.** *The probability mask*

$$\begin{pmatrix} x_{11} & x_{12} & \\ x_{21} & & \\ & & x_{33} \end{pmatrix} \tag{3.1}$$

*with all observed entries nonnegative has a completion if and only if*

$$\sqrt{x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11}} + \sqrt{x_{33}} \leq 1.$$

*This is equivalent to the conditions*

$$\begin{aligned} (x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11} + x_{33} - 1)^2 - 4(x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11})x_{33} &\geq 0, \\ \textit{and} \qquad x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11} + x_{33} &\leq 1. \end{aligned}$$

*By clearing the denominators, we get polynomial inequalities in the observed entries whose solutions are all completable probability masks of form (3.1).*

## 3.2 Semialgebraic Description

**Proposition 3.2.1.** *Partial matrices with a specified set of observed entries, which can be completed to rank-1 probability matrices, form a semialgebraic set.*

*Proof.* The independence model is a semialgebraic set defined by $2 \times 2$-minors, nonnegativity constraints and entries summing to one. The statement of the proposition follows by the Tarski-Seidenberg theorem, which states that projections of semialgebraic sets are semialgebraic. This is also known as quantifier elimination. □

The goal of this section is to find a semialgebraic description of this semialgebraic set. The difference from characterizations in Theorems 3.1.2 and 3.1.14 is that we aim to derive a description without square roots. For $2 \times 2$ partial matrices with diagonal entries, a semialgebraic description is given in Example 3.1.1 and its derivation from the inequality containing square roots is explained in Remark 3.1.4.

**Example 3.2.2.** *Let us consider $3 \times 3$ probability masks with diagonal entries $a, b, c$. Denote the elementary symmetric polynomials by $S_1 = a + b + c$, $S_2 = ab + bc + ca$ and $S_3 = abc$. By consecutive squaring and reordering terms, we get:*

$$\sqrt{a} + \sqrt{b} + \sqrt{c} = 1 \tag{3.2}$$

$$\Rightarrow 2(\sqrt{ab} + \sqrt{bc} + \sqrt{ca}) = 1 - S_1 \tag{3.3}$$

$$\Rightarrow 8(a\sqrt{bc} + b\sqrt{ac} + c\sqrt{ab}) = (1 - S_1)^2 - 4S_2 \tag{3.4}$$

$$\Rightarrow 128S_3(\sqrt{ab} + \sqrt{bc} + \sqrt{ca}) = ((1 - S_1)^2 - 4S_2)^2 - 64S_1S_3 \tag{3.5}$$

*Substituting Equation (3.3) into Equation (3.5) gives:*

$$64S_3(1 - S_1) = ((1 - S_1)^2 - 4S_2)^2 - 64S_1S_3 \tag{3.6}$$

$$\Leftrightarrow ((1 - S_1)^2 - 4S_2)^2 - 64S_3 = 0 \tag{3.7}$$

*The degree four polynomial equation (3.7) with 35 terms in the region*

$$a, b, c \geq 0, \tag{3.8}$$

$$1 - S_1 \geq 0, \tag{3.9}$$

$$(1 - S_1)^2 - 4S_2 \geq 0 \tag{3.10}$$

*gives the equation (3.2). The left hand sides of inequalities (3.9) and (3.10) are given by right hand sides of equations (3.3) and (3.4). The region defined by*

$$\sqrt{a} + \sqrt{b} + \sqrt{c} \leq 1$$

*is the same as*

$$((1 - S_1)^2 - 4S_2)^2 - 64S_3 \geq 0 \tag{3.11}$$

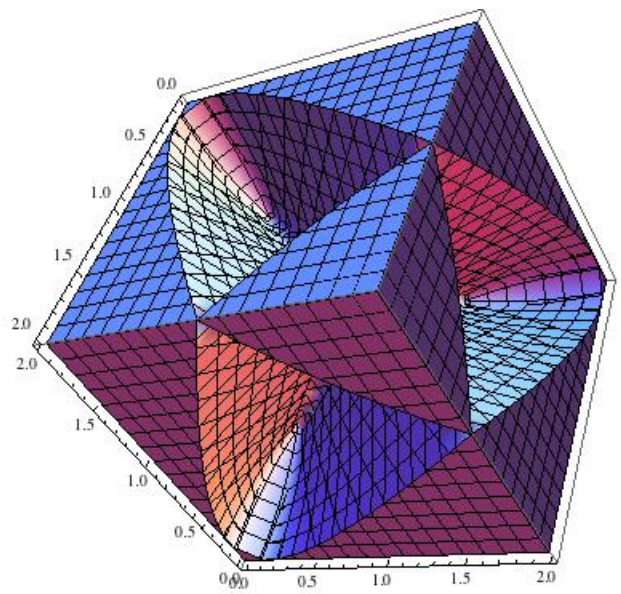



Figure 3.3: Region defined by (3.11) inside the cube $[0,2]^3$.

*together with inequalities (3.8), (3.9) and (3.10). This is a semialgebraic description of* $3 \times 3$ *diagonal partial matrices which can be completed to rank-*1 *probability matrices. Figure 3.3 shows the large semialgebraic set defined by inequality (3.11); only after slicing with inequalities (3.8), (3.9) and (3.10), do we obtain the set of completable matrices.*

Before we continue with studying semialgebraic descriptions of general partial matrices, we make a detour and characterize diagonal partial tensors that can be completed to rank-1 probability tensors. This is also a semialgebraic set and its semialgebraic description can be studied in a similar way to the partial matrix case.

## Tensors

The reasoning for diagonal matrices translates nicely into higher-dimensional tensors as follows:

**Theorem 3.2.3.** *Suppose we are given a partial probability tensor* $T \in \Delta^{n^d-1} \subset (\mathbb{R}^n)^{\otimes d}$ *with nonnegative observed entries* $a_i$ *along the diagonal, i. e. we have* $t_{ii\ldots i} = a_i$ *for* $1 \leq i \leq n$*, and all other entries unobserved. Then* $T$ *is completable if and only if*

$$\sum_{i=1}^{n} a_i^{1/d} \leq 1.$$

*Proof.* The proof is analogous to the proof of Theorem 3.1.2, with a few adjustments to deal with the multiple parametrizing vectors. The tensor $T$ can be factored as $\mathbf{u}^1 \otimes \ldots \otimes \mathbf{u}^d$, where

each $\mathbf{u}^i \in \Delta^{n-1} \subset \mathbb{R}^n$. The relations $a_i = u_i^1 \cdots u_i^d$ imply that the coordinates of $\mathbf{u}^d$ can be expressed as functions on the product of simplices $(\Delta^{n-1})^{d-1}$. Define $f : (\Delta^{n-1})^{d-1} \to \mathbb{R}$ by:

$$f(\mathbf{u}^1, \ldots, \mathbf{u}^{d-1}) = \sum_{i=1}^{n} u_i^d = \sum_{i=1}^{n} \frac{a_i}{u_i^1 \cdots u_i^{d-1}}.$$

Since every variable appears in the denominator of some term, the function $f$ approaches infinity at the boundary of the product of simplices. A candidate vector $\mathbf{u}^d$ will be available if and only if the minimum value of $f$ on the product of simplices is below one. To find the minimum, compute partial derivatives; as in the proof of Theorem 3.1.2, we let $u_n^j = 1 - \sum_{k=1}^{n-1} u_k^j$ for each $j = 1, \ldots, d-1$.

$$\frac{\partial f}{\partial u_i^j} = -\frac{a_i}{u_i^j \prod_{k=1}^{d-1} u_i^k} + \frac{a_n}{u_n^j \prod_{k=1}^{d-1} u_n^k}.$$

Setting the partial derivatives to zero gives us the following:

$$\frac{a_i}{u_i^j \prod_{k=1}^{d-1} u_i^k} = \frac{a_d}{u_n^j \prod_{k=1}^{d-1} u_d^k} \qquad \forall i \in [n-1], \forall j \in [d-1].$$

Let $c_j$ be the value on both sides of this equation. Picking two values of $j$, w.l.o.g., 1 and 2, this designation means:

$$\frac{a_i}{u_i^1 \prod_{k=1}^{d-1} u_i^k} \Big/ \frac{a_i}{u_i^2 \prod_{k=1}^{d-1} u_i^k} = \frac{u_i^2}{u_i^1} = \frac{c_1}{c_2}.$$

Applying to all indices, we have $u_i^j = \frac{c_1}{c_j} u_i^1$ for all $i, j$. Since $\sum_i u_i^j = 1 = \frac{c_1}{c_j} \sum_i u_i^1 = c_1/c_j$ for all $j$, implying that every $c_j = c_1$, and

$$\frac{a_i}{(u_i^j)^d} = c_1 \quad \forall i \in [n], \forall j \in [d-1]$$

$$\Rightarrow u_i^j = (a_i/c_1)^{1/d} = \kappa(a_i)^{1/d}.$$

for some constant $\kappa$. Since the sum $\sum_i u_i^j = 1$, the value of $\kappa = (\sum_{i=1}^n a_i^{1/d})^{-1}$. Plugging in these values of $u_i^j$, we obtain

$$f = \sum_{i=1}^{n} \frac{a_i}{(\kappa(a_i)^{1/d})^{d-1}} = \sum_{i=1}^{n} \frac{a_i^{1/d}}{\kappa^{d-1}} = \left( \sum_{i=1}^{n} a_i^{1/d} \right)^d.$$

Since $f$ is at its minimum here, the value must be less than one for a solution to exist, proving the theorem. □

We will characterize the semialgebraic set of diagonal partial tensors which can be completed to rank-1 probability tensors. This is the positive part of the unit ball in the $L^{\frac{1}{d}}$ space.

**Proposition 3.2.4.** *There exists a unique irreducible polynomial $f$ of degree $d^{n-1}$ with constant term 1 that vanishes on the boundary of diagonal partial tensors which can be completed to rank-1 probability tensors. The semialgebraic description takes the form $f \geq 0$, coordinates $\geq 0$ plus additional inequalities that separate our set from other chambers in the region defined by $f \geq 0$.*

The proof of Proposition 3.2.4 was suggested to us by Bernd Sturmfels. For analogous proof idea, see [42, Lemma 2.1].

*Proof.* Denote the diagonal entries of the partial tensor by $x_1, \ldots, x_n$. We will show that the defining polynomial of the $\frac{1}{d}$-unit ball can be written as

$$p_{d,n} = \prod_{\substack{y_i \text{ s.t. } y_i^d = x_i^{1/d} \\ \text{for each } i}} \left( (1 - y_1 - ... - y_{n-1})^d - x_n \right). \tag{3.12}$$

We want to eliminate $y_1, \ldots, y_n$ from the ideal

$$I = \left\langle y_1^d - x_1, \ldots, y_n^d - x_n, \sum_{i=1}^{n} y_i - 1 \right\rangle \subset \mathbb{Q}[x_1, \ldots, x_n, y_1, \ldots, y_n].$$

First replace $y_n$ by $1 - y_1 - ... - y_{n-1}$ in the equation $y_n^d - x_n$. We consider the field of rational functions $K = \mathbb{Q}(x_1, \ldots, x_n)$. Solving the first $n-1$ equations $y_i^d - x_i$ is equivalent to adjoining the $d$-th roots of $x_i$ for $i \in \{1, \ldots, n-1\}$ to the base field. This gives a Galois extension $L$ of degree $d^{n-1}$ over $K$. The Galois group of the extension $L/K$ is $(\mathbb{Z}/d\mathbb{Z})^{n-1}$. The product (3.12) is the orbit of $(1 - y_1 - ... - y_{n-1})^d - x_n$ under the action of the Galois group, and thus lies in the base field $K$. Every factor in the product (3.12) is integral over $\mathbb{Q}[x_1, \ldots, x_n]$, hence the product (3.12) is a degree $d^{n-1}$ polynomial in $x_1, \ldots, x_n$. No subproduct is left invariant under the Galois group, so (3.12) is irreducible. □

In Example 3.2.2, the polynomial $f$ is the left hand side of Equation (3.7). Inequalities (3.8), (3.9) and (3.10) separate the set $\sqrt{a} + \sqrt{b} + \sqrt{c} \leq 1$ from other chambers of $f \geq 0$.

The semialgebraic description for an arbitrary probability mask can be constructed in five steps using the semialgebraic description for diagonal masks:

1. Take all elements of the universal Gröbner basis of $\Pi_{m \times n}$ (see Section 3.3) that contain only observed entries. With these equations we check that observed entries do not contradict the rank-1 condition.

2. For each $R \subseteq [m], C \subseteq [n]$ consider the semialgebraic set

$$\{x_{ij} \text{ is observed} : x_{ij} = 0 \text{ if } i \in R \text{ or } j \in C \text{ and } x_{ij} > 0 \text{ otherwise}\}. \tag{3.13}$$

For example, the partial matrix

$$\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}.$$

does not belong to the semialgebraic set (3.13) for any $R$ and $C$

3. For fixed $R$ and $C$, express all completable entries as rational functions in observed entries, see Equation (3.15). This gives a block diagonal matrix.

4. Construct the diagonal mask corresponding to the block diagonal mask, where each diagonal entry is equal to the sum of entries in the corresponding block. Clear denominators. Intersect the semialgebraic set for this diagonal mask with the semialgebraic set in Step 2.

5. Take the union of semialgebraic sets in Step 5 for all $R$ and $C$ and intersect with the variety in Step 1.

## 3.3 Completion Algorithms

### Algorithm for Checking Completability

When the bipartite graph corresponding to a probability mask is a tree with all nonzero entries, there is a unique completion. The missing entries can be computed by the following relations among entries in a cycle:

$$\prod_{(i,j)\in E_1} x_{ij} - \prod_{(i,j)\in E_2} x_{ij} = 0, \tag{3.14}$$

where $(E_1, E_2)$ is a partition of the edges in a cycle so that no two edges in either $E_i$ are adjacent. In particular, since the graph $G$ is connected, if there is a missing edge $(r_k, c_l)$ there is a path from $r_k$ to $c_l$ in $G$; ordering those edges in the path order from $r_k$ to $c_l$, let $S_1$ be every other edge in the list starting with $r_k$, and let $S_2$ be the complement in the path. Then:

$$\omega(r_k c_l) = \prod_{(ij)\in S_1} x_{ij} \,/\, \prod_{(ij)\in S_2} x_{ij}. \tag{3.15}$$

This formula is implemented in [26, Algorithm 8].

**Example 3.3.1.** *Given the base indicated at left, we can uniquely complete as at right using polynomials (3.14):*

$$\begin{pmatrix} a & b & \\ c & & d \\ & e & \end{pmatrix} \qquad \to \qquad \begin{pmatrix} a & b & \frac{ad}{c} \\ c & \frac{bc}{a} & d \\ \frac{ae}{b} & e & \frac{ade}{bc} \end{pmatrix}$$

The entries are assumed to be nonzero; if an observed entry is zero, we omit an appropriate row and/or column and complete the rest. If a probability matrix only has one connected component, then completability amounts to checking that the unique rank-1 completion has entries summing to 1.

Combining Theorem 3.1.14 with the algorithm just described, we may now present an algorithm for checking completability of an arbitrary partial probability matrix.

**Algorithm 3.3.2** (Completability of Arbitrary Partial Matrices)**.** *Let $M$ be a partial matrix with nonnegative entries whose sum is less than or equal to 1. The following algorithm answers "Is $M$ completable?":*

1. *Translate $M$ into the corresponding bipartite graph $G$ including edge weights.*
2. *If all edge weights are nonzero, proceed to Step 3; otherwise:*
   a) *If $G$ is not prunable, return "NO."*
   b) *If $G$ is prunable, execute a careful pruning, as described in Definition 3.1.9.*
3. *In the remaining graph, suppose there are $k$ connected components $C_1, \ldots, C_k$. If $k = 1$:*
   a) *Check that Equation (3.14) holds for every cycle in the graph. If a cycle fails, return "NO."*
   b) *Uniquely complete to $K_{m,n}$. If the edge weights after completion add up to 1, return "YES." Else, return "NO."*
4. *If $k > 1$:*
   a) *Check that Equation (3.14) holds for every cycle in the graph. If a cycle fails, return "NO."*
   b) *Add in edges to make each component a complete bipartite graph, with edge weights from Equation (3.15). Let $S = \sum_{i=1}^{k} \sqrt{b_i}$ where $b_i$ is the sum of the entries in $C_i$ after the last step. If $S > 1$, return "NO." Else, return "YES."*

## Algorithm for Completing Partial Matrices with Two Components

In the last section, an algorithm was presented that determines completability for an arbitrary mask. In the special case where Step 3 returns $k = 2$ connected components, there is a finite set of completions, with cardinality between 0 and 2 (see Section 3.4 for an explanation). The following algorithm returns this set of completions:

**Algorithm 3.3.3.** *Begin with a graph $G$, and carry out Algorithm 3.3.2 until Step 3. Let $H$ be the graph returned from Step 3 with two complete bipartite components $C_1$ and $C_2$.*

1. *Choose any edge connecting $C_1$ and $C_2$; set it equal to $X$. Fill in the remaining entries (in terms of the known entries and $X$) using Equation (3.15).*

2. *Solve the quadratic equation $\sum p_{ij} = 1$ for $X$, where $p_{ij}$'s are the set of entries obtained in Step 1. Substitute the two values for $X$ into the completed matrix from Step 3 to obtain two (usually distinct) completions.*

3. *Reintroduce any vertices removed in Step 2 of Algorithm 3.3.2 as rows/columns with all zero entries.*

Note that since the entries involving $X$ all have the same sign, and their sum is fixed, the solutions for $X$ (if real) must be both positive or both negative.

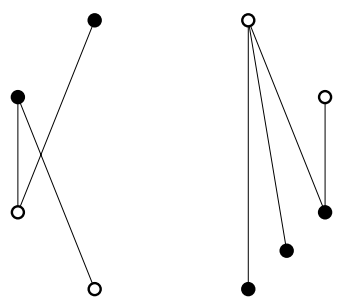
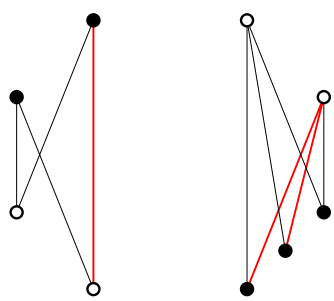
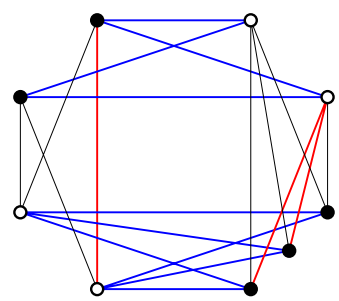
| | | |
|---|---|---|
| **START:** Spanning forest with two connected components.  | 1) Complete each component to $K_{m,n}$ uniquely using binomials.  | 2) Connect components using two solutions to $\sum p_{ij} = 1$.  |

Table 3.2: Algorithm for completing probability matrix projections.

**Example 3.3.4.** *Consider the projection of a $3 \times 3$ matrix into $\Delta_0^4$ indicated by the partial matrix below:*

$$\begin{pmatrix} .06 & .09 & \\ .08 & & \\ & & .15 \end{pmatrix}$$

***Step 1:*** *Add in the missing edges so that both connected components are complete bipartite.*

$$\begin{pmatrix} .06 & .08 & \\ .09 & .12 & \\ & & .15 \end{pmatrix}$$

***Step 2:*** *Add in $X$ to connect the components, and fill in the remaining edges.*

$$\begin{pmatrix} .06 & .08 & X \\ .09 & .12 & 1.5X \\ .009/X & .012/X & .15 \end{pmatrix}$$

***Step 3:*** *Set $\sum p_{ij} = 1$ and solve for $X$:*

$$(.06 + .08 + .09 + .12 + .15) + X + 1.5X + .009/X + .012/X = 1$$

$$\Rightarrow .5 + 2.5X + .021/X = 1 \Rightarrow 2.5X^2 - .5X + .021 = 0$$

*The two solutions for $X$ yield the following two completions:*

$$\begin{pmatrix} .06 & .08 & .06 \\ .09 & .12 & .09 \\ .15 & .2 & .15 \end{pmatrix} \qquad \begin{pmatrix} .06 & .08 & .14 \\ .09 & .12 & .21 \\ 9/140 & 3/35 & .15 \end{pmatrix}$$

If a set of entries is indeed a projection of a probability matrix, this algorithm will recover it. Though the generic fiber has two points, there will be a unique completion if and only if the discriminant of the quadratic polynomial in $X$ is zero.

**Example 3.3.5.** *Applying the algorithm to a random point $x \in \Delta_0^{m+n-2}$ does not necessarily produce a matrix in the probability simplex $\Delta^{mn-1}$. Indeed, even in the smallest cases, this is seen to be false; the matrix below left is obviously in $\Delta_0^2$, but its fiber in $\mathcal{V}(\Pi_{m\times n})$ consists of the matrix at right and its complex conjugate:*

$$\begin{pmatrix} \frac{1}{3} & \\ & \frac{1}{3} \end{pmatrix} \to \begin{pmatrix} \frac{1}{3} & \frac{1}{6} + \frac{i}{2\sqrt{3}} \\ \frac{1}{6} - \frac{i}{2\sqrt{3}} & \frac{1}{3} \end{pmatrix}$$

## Completing Partial Matrices with More Than Two Components

When the graph has $n > 2$ components, and the quantity $\sum_{i=1}^n \sqrt{b_i}$ is less than 1, there is an $(n-2)$-dimensional set of completions. For example, consider a diagonal partial $3\times 3$ matrix with each observed entry equal to the same constant $c$. In Figure 3.4, each curve represents values of $\mathbf{u}$ that parametrize a completion of the partial matrix with $c$ on the diagonal, for various values of $c$. Here $\mathbf{u}$ is projected onto the first two coordinates.

For practical applications, some completions are more useful than others. We may want to minimize a distance measure $d$ from a fixed probability distribution. We will explain how to use Lagrange multipliers to solve this optimization problem if $d$ is the Euclidean distance from the uniform distribution.

Let $S = \sum_{i=1}^n \sqrt{a_i}$. Let us parametrize a vector $\mathbf{u} \in \Delta^{n-1}$ by

$$\mathbf{u}(\mathbf{t}) = \left( \tfrac{\sqrt{a_1}}{S} + t_1, \tfrac{\sqrt{a_2}}{S} + t_2, \ldots, \tfrac{\sqrt{a_{n-1}}}{S} + t_{n-1}, \tfrac{\sqrt{a_n}}{S} + t_n \right),$$

where $t_n = -t_1 - t_2 - \ldots - t_{n-1}$. Then

$$f(\mathbf{u}(\mathbf{t})) = \sum_{i=1}^n v_i(\mathbf{t}) = \sum_{i=1}^n \frac{a_i}{u_i(\mathbf{t})} = \sum_{i=1}^n \frac{a_i}{\frac{\sqrt{a_i}}{S} + t_i} = \sum_{i=1}^n \frac{a_i S}{\sqrt{a_i} + t_i S}.$$

**Proposition 3.3.6.** *The semialgebraic set of completions of a diagonal probability mask $diag(a_1, a_2, \ldots, a_n)$ is given by $f(\mathbf{u}(\mathbf{t})) = 1$ and $\mathbf{u}(\mathbf{t}) \geq 0$ (after clearing denominators).*

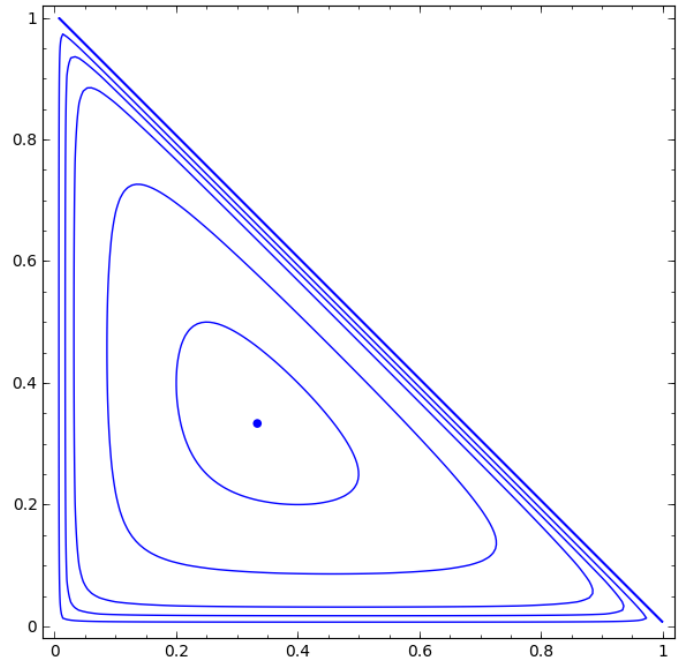



Figure 3.4: Solution curves for $c = 1/9, 1/10, 1/16, 1/36, 1/64$, and $1/150$.

By the method of Lagrange multipliers, an element $\left(t_1, t_2, \cdots, t_{n-1}\right)$ in this semialgebraic set is a critical point for a distance function $d$ if and only if the gradient of $d$ is a constant multiple of the vector of partial derivatives $\frac{\partial f}{\partial t_i}$. To compute all the critical points of the function $d$ on the variety given by $f(\mathbf{t}) = 1$, we need to solve the system of rational equations given by $f(\mathbf{t}) = 1$ and all the $2 \times 2$ minors of the matrix

$$L = \begin{pmatrix} \frac{\partial f}{\partial t_1} & \frac{\partial f}{\partial t_2} & \cdots & \frac{\partial f}{\partial t_{n-1}} \\ \frac{\partial d}{\partial t_1} & \frac{\partial d}{\partial t_2} & \cdots & \frac{\partial d}{\partial t_{n-1}} \end{pmatrix}.$$

Finally we need to check for all real solutions which satisfy $\mathbf{u}(\mathbf{t}) \geq 0$ which one minimizes the distance $d$.

**Example 3.3.7.** *Let us return to the matrix $A = \mathrm{diag}(1/4, 1/25, 1/36)$, and find a completion that minimizes the Euclidean distance from the uniform distribution:*

$$d = \left[\sum_{i,j} \left(p_{ij} - \frac{1}{9}\right)^2\right]^{1/2} = \left[\sum_{i,j} \left(u_i v_j - \frac{1}{n^2}\right)^2\right]^{1/2}.$$

*Our method is not restricted to Euclidean distance and can be used for any distance measure.*

*We construct the Lagrange matrix*

$$L = \begin{pmatrix} \frac{\partial f}{\partial t_1} & \frac{\partial f}{\partial t_2} \\ \frac{\partial d}{\partial t_1} & \frac{\partial d}{\partial t_2} \end{pmatrix}$$

*and find the critical points of $d$ on the variety $f = 1$ by solving the system of rational equations $\{f = 1, det(L) = 0\}$. We use* `maple` *to construct $L$ and to solve the system of equations. This system has 18 solutions, out of which ten are real and four are feasible, i.e. they satisfy $\mathbf{u} \geq 0$. The minimum is achieved at*

$$M = \begin{pmatrix} 0.250 & 0.049 & 0.215 \\ 0.204 & 0.040 & 0.176 \\ 0.032 & 0.006 & 0.028 \end{pmatrix} \text{ and } M^T.$$

*The Euclidean distance from the uniform distribution is 0.683.*

Any path from the local minimum to the boundary of the simplex will strike at least one solution. If any completion is acceptable, we can designate a simple path and find its points of intersection with the semialgebraic set of completions.

**Proposition 3.3.8.** *Let $A = \mathrm{diag}(a_1, \ldots, a_n)$, such that $n > 2$ and $S = \sum \sqrt{a_i} < 1$. Then, a completion of the matrix is given by $\mathbf{u}^T\mathbf{v}$ with:*

$$\mathbf{u} = \left( \frac{\sqrt{a_1}}{S} + t, \frac{\sqrt{a_2}}{S} - t, \frac{\sqrt{a_3}}{S}, \ldots, \frac{\sqrt{a_n}}{S} \right),$$

*where $t$ is one of the solutions to the following quadratic equation:*

$$\begin{aligned} \left( \frac{\sqrt{a_1} + \sqrt{a_2}}{S} \right) t^2 + & \left( a_2 - a_1 - \left( \frac{\sqrt{a_1} + \sqrt{a_2}}{S} \right)^2 \right) t \\ & + \left( \frac{a_1\sqrt{a_2} + a_2\sqrt{a_1}}{S} - \frac{\sqrt{a_1 a_2}(\sqrt{a_1} + \sqrt{a_2})}{S^3} \right) = 0, \end{aligned} \tag{3.16}$$

*both of which lie in the interval $[-\sqrt{a_1}/S, \sqrt{a_2}/S]$. The coordinates of $\mathbf{v}$ are obtained by setting $v_i = a_i u_i^{-1}$.*

*Proof.* The trajectory traced for values of $t \in [-\sqrt{a_1}/S, \sqrt{a_2}/S]$ is a line segment on the simplex. Setting the sum of the coordinates of $(a_i/u_i)_{i=1,\ldots,n}$ equal to 1 and clearing denominators gives the quadratic equation above. Since it passes through the local minimum, continuity implies existence of two solutions in the desired interval. □

**Example 3.3.9.** *Consider the matrix $A = \mathrm{diag}(1/4, 1/25, 1/36)$. To obtain a completion, we set $\mathbf{u} = (\frac{15}{26} + t, \frac{3}{13} - t, \frac{5}{26})$ and solve the quadratic equation (3.16), which turns into*

$$\frac{77}{90} t^2 + \frac{28}{325} t - \frac{28}{845} = 0$$

*giving solutions $t = -\frac{6}{715}\sqrt{586} - \frac{36}{715}$, or $\frac{6}{715}\sqrt{586} - \frac{36}{715}$. Using the latter value, we obtain $\mathbf{u} \approx (\ .730,\ \ .078,\ \ .192\ ), \mathbf{v} \approx (\ .343,\ \ .513,\ \ .144\ )$, and the matrix completion:*

$$\mathbf{u}^T\mathbf{v} \approx \begin{pmatrix} 0.250 & 0.374 & 0.105 \\ 0.027 & 0.040 & 0.011 \\ 0.066 & 0.098 & 0.028 \end{pmatrix}.$$

## 3.4 Algebraic Matroids

In this section, we make the connection to matroid theory explicit. A closer focus on algebraic matroids associated to determinantal ideals may be found in [25]; earlier, its application to low-rank matrix completion was explored in [26]. Later, we will discuss how the complicated matroid structures for the variety of rank-$r$ matrices and the variety of higher-order rank-1 tensors make their analysis less accessible.

### Rank-One Determinantal Matroid

The determinantal matroid is the algebraic matroid associated to the determinantal ideal. In particular, the determinantal ideal $I_{r,m\times n}$ is generated by $(r+1)\times(r+1)$-minors of an $m\times n$ matrix of variables $x_{ij}$, and its vanishing set is the set of $m\times n$ matrices of rank $\leq r$. The matroid encoding the algebraic relationships among the $x_{ij}$'s is denoted $\mathbf{D}(m\times n, r)$. In the case of the determinantal matroid $\mathbf{D}(m\times n, 1)$, much is clearly understood. In particular, the matroid $\mathbf{D}(m\times n, 1)$ is the graphic matroid on $K_{m,n}$. Some consequences are in the following proposition:

**Proposition 3.4.1.** *The following facts hold about* $\mathbf{D}(m\times n, 1)$*:*

1. *Rank of* $\mathbf{D}(m\times n, 1) = m+n-1$.
2. $\mathcal{B} = \{$*spanning trees of* $K_{m,n}\}$.
3. $\mathcal{C} = \{$*simple cycles of* $K_{m,n}\}$.

Circuit polynomials are the essentially unique relations among the elements of each circuit of the matroid. The circuit polynomials of $\mathbf{D}(m\times n, 1)$ are each of the form described in Equation (3.14).

### Probability Matroid

The rank-1 probability matroid is the algebraic matroid associated to the ideal of algebraic relations among the entries $p_{ij} = P(X = x_i, Y = y_j)$ for $X, Y$ independent discrete random variables. Let this ideal be denoted $\Pi_{m\times n}$; it is generated as

$$\Pi_{m\times n} = \left\langle 2\times 2 \text{ minors}, \sum p_{ij} - 1 \right\rangle;$$

the associated variety is the section of $\mathcal{V}(I_{1,m\times n})$ with the hyperplane $\sum p_{ij} - 1$.

**Proposition 3.4.2.** *The dimension of* $\mathcal{V}(\Pi_{m\times n})$ *is* $m+n-2$.

*Proof.* There is a parametrization of $\mathcal{V}(\Pi_{m\times n})$ analogous to the standard parametrization of the determinantal variety.

$$(a_1, \ldots, a_{m-1}, b_1, \ldots, b_{n-1}) \longmapsto (a_i b_j : 1 \leq i \leq m, 1 \leq j \leq n),$$

where we set $a_m = 1 - \sum_{i=1}^{m-1} a_i$ and $b_n = 1 - \sum_{i=1}^{n-1} b_i$. This map is actually invertible, via:

$$(p_{ij}) \longmapsto \left( \sum_j p_{1j}, \ldots, \sum_j p_{m-1,j}, \sum_i p_{i1}, \ldots, \sum_i p_{i,n-1} \right)$$

The existence of these maps implies that the dimension of the two varieties is equal. □

The corollary for the matroid is that $\rho(\mathcal{M}(\Pi_{m\times n})) = m + n - 2$; however, the matroid can be much more tightly characterized:

**Theorem 3.4.3.** *The matroid $\mathcal{M}(\Pi_{m\times n})$ has bases given by $\mathcal{B} = \{$spanning forests with 2 connected components$\}$, and circuits given by $\mathcal{C} = \{$spanning trees and non-spanning simple cycles$\}$.*

*Proof.* The bases of a matroid characterize it completely; we claim that spanning forests on two components are the bases of $\mathcal{M}(\Pi_{m\times n})$. From [25], a base of an algebraic matroid is a subset of the coordinates for which the projection map is finite surjective. Then, the algorithm in Section 3.3 returns the finite fiber over any point, since we are not restricted to the simplex in the more general algebraic context. This gives the set of bases, and the set of circuits follows from the two types of relations in the ideal. □

**Remark 3.4.4.** *The matroid-theoretic way of formulating Theorem 3.4.3 is that $\mathcal{M}(\Pi_{m\times n})$ is the truncation to rank $m + n - 2$ of the determinantal matroid $\mathbf{D}(m \times n, 1)$, see [56, Definition 4.1].*

*A question worth exploring with the rank one determinantal variety, and for varieties in general, is as follows:* What hyperplane sections induce a truncation of the algebraic matroid? *Intuition dictates that almost all hyperplanes would lead to a truncation; characterizing the subvariety of the projective space of hyperplanes that would induce a different matroid seems very interesting but nontrivial.*

The base degree, as defined in Section 1.3, is the cardinality of a generic fiber in a projection onto the base's coordinates. If a base $B$ has an isolated vertex, the base degree is 1; if both components have positive number of edges, the base degree is 2.

The circuit polynomials associated to cycles are the binomials inherited from the determinantal ideal. As for the spanning trees, the circuit polynomials are obtained either by elimination in $\Pi_{m\times n}$, **or** we complete the matrix using the binomials and then set the sum equal to 1. After clearing denominators, we obtain a polynomial with the circuit polynomial as a factor; indeed, it is actually the circuit polynomial.

**Proposition 3.4.5.** *Let $F$ be the set of edges in a spanning tree $C$ not incident to a leaf. Let $P_C$ be the set of paths in $C$ from white leaves to black leaves.*
*For a path $p \in P_C$, take $p_1$ to be the set of alternate edges of $p$ containing the leaves, let $p_2$ denote the set of alternate edges of $p$ not containing the leaves, and let $m(p)$ denote the product of $p_1$ divided by the product of $p_2$.*

*The circuit polynomial for $C$ is the degree $|F|+1$ polynomial:*

$$\prod_{(k,l)\in F} x_{kl} \left( \sum_{(i,j)\in C} x_{ij} + \sum_{p\in P} m(p) - 1 \right)$$

*Proof.* We first need to show that the rational function inside of the parentheses evaluates to zero on a rank-1 matrix summing to one. Then, we need to check that after clearing denominators by multiplying the monomial on the left, we have an irreducible polynomial.

Each term in the sum of entries is a Laurent monomial of total degree 1. The variables in the first sum are the revealed entries. The $m(p)$ terms are the Laurent monomials we using the fundamental cycle to assign a value to every missing edge from $K_{m,n}$. So the parenthetical expression is in the ideal.

There are $mn$ terms of degree $|F|+1$ and one term of degree $|F|$. Let $fg$ be a factorization of $\theta_C$. Then exactly one of $f$ and $g$ must have terms in two degrees; otherwise, the product would have terms in at least three degrees. Let $f = f_1 + f_2$ where $f_1$ is the part of highest degree. The product $f_2 g$ must be equal to the monomial

$$-\prod_{(k,l)\in F} x_{kl}.$$

This implies that $g$ is a monomial. Since no variable divides every term in $\theta_C$, that monomial must be a unit. □

## 3.5 Generalizations

### Low-Rank Matrices

One natural direction to generalize these results would be to fix $r > 1$, and find conditions for a matrix to be rank or nonnegative rank $r$ and have nonnegative entries that sum to 1. One obvious consequence of our results is that any matrix completable to rank 1 is trivially completable as a higher-rank matrix. It is harder to provide tighter conditions, however, even in the smallest examples.

**Example 3.5.1** ($r = 2, m = n = 3$)**.** *There are two polynomials constraining the entries of a $3 \times 3$ probability matrix of rank 2: the determinant must be zero, and the sum of the entries must be 1. The variety of matrices with these properties has dimension 7.*

*In terms of the matroid, there are two distinct bases up to row permutation and transpose: the set of size* 7 *obtained by omitting adjacent edges, and the set obtained by omitting non-adjacent edges. To find completions, we substitute* $X$ *and* $R - X$ *for the missing entries, where* $R = 1 - (a + b + c + d + e + f + g)$*:*

$$\textit{Base 1:} \begin{pmatrix} a & b & c \\ d & e & f \\ g & X & R-X \end{pmatrix} \qquad \textit{Base 2:} \begin{pmatrix} a & b & c \\ d & X & f \\ g & e & R-X \end{pmatrix}.$$

*Since the sum is now fixed at* 1*, we only need to check that there is a value of* $X$ *in* $[0, R]$ *so that the determinant is* 0*. The determinant in base 1 gives a linear equation in* $X$*, while the determinant in base* 2 *is a quadratic; the solutions to each are:*

$$X = \frac{g(bf - ce) + R(ae - bd)}{(ae - bd) + (af - cd)}$$

$$X = \frac{(aR + bd - cg) \pm \sqrt{(aR + bd - cg)^2 - 4a(b(dR - fg) + e(af - cd))}}{2a}$$

*Substituting the values* $(.07, .09, .09, .12, .15, .04, .16)$ *for the known coordinates of the matrix yields a completion for Base 1, but since the discriminant of the Base 2 determinant is negative, no probability completion is possible.*

Since the determinantal matroid is not fully understood for $r > 1$, the results on completability to probability matrices of rank 1 will be difficult to generalize to higher rank. From the statistics viewpoint, it would be more interesting to study completability to probability matrices of nonnegative rank at most $r$, because the $r$-th mixture model of two discrete random variables is the semialgebraic set of matrices of nonnegative rank at most $r$. If a nonnegative matrix has rank $0, 1$, or 2, then its nonnegative rank is equal to its rank. Hence, in Example 3.5.1, we simultaneously address the question of completing a partial matrix to a probability matrix of nonnegative rank 2.

If $r \geq 3$, then matrices of nonnegative rank at most $r$ form a complicated semialgebraic set. For $r = 3$, a semialgebraic description of this set is given in [28, Theorem 3.1]. Partial matrices that are completable to probability matrices of nonnegative rank at most 3, are coordinate projections of this semialgebraic set. To find all probability completions of nonnegative rank 3 of a partial matrix, one has to find all probability completions of rank 3 and then take the intersection with the semialgebraic set of matrices of nonnegative rank at most 3.

## General Tensors

The generalization to higher-dimensional tensors brings several challenges. Theorem 3.2.3 gives a partial result characterizing diagonal tensors. However, the nice bipartite graph structure we had for matrices becomes $k$-partite hypergraphs with $k$-hyperedges; notions

like connectivity and acyclicity will need to be modified. So, while any rank-1 matrix completability problem was reducible to a diagonal case, the tensor case does not seem to be reducible in the same way. We record here the results for the smallest case distinct from matrices:

**Example 3.5.2** ($2 \times 2 \times 2$ Tensors)**.** *The variety of* $2 \times 2 \times 2$ *tensors whose entries sum to* 1 *is* 3*-dimensional. For this example, we will only consider the independent sets of the algebraic matroid corresponding to this variety. We use the octahedral symmetry group of the cube to look only at orbits of independent sets:*

1. *(Size* 1*) Any singleton, e.g.* $p_{000}$*. The only condition is* $p_{000} \leq 1$*.*
2. *(Size* 2*) Three orbits of pairs:*
   a) $p_{000}, p_{001}$*:* $p_{000} + p_{001} \leq 1$*.*
   b) $p_{000}, p_{011}$*:* $\sqrt{p_{000}} + \sqrt{p_{011}} \leq 1$*.*
   c) $p_{000}, p_{111}$*:* $\sqrt[3]{p_{000}} + \sqrt[3]{p_{111}} \leq 1$*.*
3. *(Size* 3*) Three orbits of triples:*
   a) $p_{000}, p_{001}, p_{010}$*:* $p_{000} + p_{001} + p_{010} + (p_{001}p_{010}/p_{000}) \leq 1$*.*
   b) $p_{000}, p_{001}, p_{110}$*:* $\sqrt{p_{000} + p_{001}} + \sqrt{p_{110} + p_{001}p_{110}/p_{000}} \leq 1$*.*
   c) $p_{000}, p_{101}, p_{011}$*: The tensor is completable if and only if the equation*
   $$x^3 + (p_{000} + p_{101} + p_{011} - 1)x^2 + (p_{000}p_{101} + p_{000}p_{011} + p_{101}p_{011})x + p_{000}p_{101}p_{011} = 0$$
   *has a root in the interval* $[0, 1]$*.*

*In this small example, most partial tensors reduced to a case we knew how to handle; however, 3c does not have a simple semialgebraic description. This equation was obtained by adding* $x$ *as an entry then completing using minors, and summing the entries to* 1*.*

With this chapter, the matroid for rank-1 matrices summing to 1 has been completely described. Moving forward, tensors and higher-rank matrices summing to 1 present new challenges with a similar flavor. Having explored an application to statistics, the next chapter will be devoted to biology.

# Chapter 4

# Biology: Chemical Reaction Matroid

This chapter is based on the paper *Algebraic Systems Biology: A Case Study for the Wnt Pathway* [20], written jointly with Elizabeth Gross, Heather Harrington, and Bernd Sturmfels. Our goal in this chapter is to employ a comprehensive battery of algebraic tools to study the shuttle model for the Wnt pathway. While many of these tools are not matroidal in nature – for example, polyhedral analysis and numerical algebraic geometry – the chapter also makes extensive use of matroids. In particular, the algebraic matroid associated to the model is featured in Section 4.4, and also aids our analysis in Section 4.6. Using matroids in an algebraic study of a chemical reaction network shows how the algebraic matroid can give us compelling combinatorial tools to understand a polynomial system's structure.

The theory of biochemical reaction networks is fundamental for systems biology [27, 54]. It is based on a wide range of mathematical fields, including dynamical systems, numerical analysis, optimization, combinatorics, probability, and, last but not least, algebraic geometry. There are numerous articles that use algebraic geometry in the study of biochemical reaction networks, especially those arising from mass action kinetics. A tiny selection is [8, 14, 24, 40, 49].

We here perform a detailed analysis of one specific system, namely the shuttle model for the Wnt signaling pathway, introduced recently by MacLean, Rosen, Byrne, and Harrington [37]. Our aim is twofold: to demonstrate how biology can lead to interesting questions in algebraic geometry and to apply state-of-the-art techniques from computational algebra to biology.

The dynamical system we study consists of the following 19 ordinary differential equations. Their derivation and the relevant background from biology will be presented in Section 4.1.

$$
\begin{array}{rcl}
\dot{x}_1 & = & -k_1x_1 + k_2x_2 \\
\dot{x}_2 & = & k_1x_1 - (k_2 + k_{26})x_2 + k_{27}x_3 - k_3x_2x_4 + (k_4 + k_5)x_{14} \\
\dot{x}_3 & = & k_{26}x_2 - k_{27}x_3 - k_{14}x_3x_6 + (k_{15} + k_{16})x_{15} \\
\dot{x}_4 & = & -k_3x_2x_4 - k_9x_4x_{10} + k_4x_{14} + k_8x_{16} + (k_{10} + k_{11})x_{18} \\
\dot{x}_5 & = & -k_{28}x_5 + k_{29}x_7 - k_6x_5x_8 + k_5x_{14} + k_7x_{16} \\
\dot{x}_6 & = & -k_{14}x_3x_6 - k_{20}x_6x_{11} + k_{15}x_{15} + k_{19}x_{17} + (k_{21} + k_{22})x_{19} \\
\dot{x}_7 & = & k_{28}x_5 - k_{29}x_7 - k_{17}x_7x_9 + k_{16}x_{15} + k_{18}x_{17} \\
\dot{x}_8 = -\dot{x}_{16} & = & -k_6x_5x_8 + (k_7 + k_8)x_{16} \\
\dot{x}_9 = -\dot{x}_{17} & = & -k_{17}x_7x_9 + (k_{18} + k_{19})x_{17} \\
\dot{x}_{10} & = & k_{12} - (k_{13} + k_{30})x_{10} - k_9x_4x_{10} + k_{31}x_{11} + k_{10}x_{18} \\
\dot{x}_{11} & = & -k_{23}x_{11} + k_{30}x_{10} - k_{31}x_{11} - k_{20}x_6x_{11} - k_{24}x_{11}x_{12} + k_{25}x_{13} + k_{21}x_{19} \\
\dot{x}_{12} = -\dot{x}_{13} & = & -k_{24}x_{11}x_{12} + k_{25}x_{13} \\
\dot{x}_{14} & = & k_3x_2x_4 - (k_4 + k_5)x_{14} \\
\dot{x}_{15} & = & k_{14}x_3x_6 - (k_{15} + k_{16})x_{15} \\
\dot{x}_{18} & = & k_9x_4x_{10} - (k_{10} + k_{11})x_{18} \\
\dot{x}_{19} & = & k_{20}x_6x_{11} - (k_{21} + k_{22})x_{19}
\end{array}
\tag{4.1}
$$

The quantity $x_i$ is a differentiable function of an unknown $t$, representing time, and $\dot{x}_i(t)$ is the derivative of that function. This dynamical system has five linear conservation laws:

$$
\begin{array}{rcl}
0 & = & (x_1 + x_2 + x_3 + x_{14} + x_{15}) - c_1 \\
0 & = & (x_4 + x_5 + x_6 + x_7 + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19}) - c_2 \\
0 & = & (x_8 + x_{16}) - c_3 \\
0 & = & (x_9 + x_{17}) - c_4 \\
0 & = & (x_{12} + x_{13}) - c_5
\end{array}
\tag{4.2}
$$

The 31 quantities $k_i$ are the rate constants of the chemical reactions, and the five $c_i$ are the conserved quantities. Both of these are regarded as parameters, so we have 36 parameters in total. Our object of interest is the *steady state variety*, which is the common zero set of the right hand sides of (4.1) and (4.2). This variety lives in $K^{19}$, where $K$ is an algebraically closed field that contains the rational numbers $\mathbb{Q}$ as well as the 36 parameters $k_i$ and $c_i$. If these parameters are fixed to be particular real numbers then we can take $K = \mathbb{C}$, the field of complex numbers. If it is preferable to regard $\mathbf{k} = (k_1, \ldots, k_{31})$ and $\mathbf{c} = (c_1, \ldots, c_5)$ as vectors of unknowns, then $K = \overline{\mathbb{Q}(\mathbf{k}, \mathbf{c})}$ is the algebraic closure of the rational function field. In this latter setting, when all parameters are generic, we shall derive the following result:

**Theorem 4.0.3.** *The polynomials in (4.1)–(4.2) have* 9 *distinct zeros in* $K^{19}$ *when* $K = \overline{\mathbb{Q}(\mathbf{k}, \mathbf{c})}$.

By analyzing the steady state variety, we can better understand the model, which is nonlinear, and thus the biological system. The aim is to predict the system's behavior, offer biological insight, and determine what data are required to verify or reject the model. Here is a list of questions one might ask about our model from the perspective of systems biology.

## Biological Problems

These are labeled according to the section that will address them.

3. *For what real positive rate parameters and conserved quantities does the system exhibit multistationarity?* This question is commonly asked when using a dynamical system for modeling a real-world phenomenon. When modeling a process that experimentally appears to have more than one stable equilibrium, multistationary models are preferred.

4. *Suppose we can measure only a subset of the species concentrations. Which subsets can lead to model rejection?* If all species are measurable at steady state, then we can substitute data into the system (4.1), and check that all expressions $\dot{x}_i$ are close to zero. If only some $x_i$ are known, we still want to be able to evaluate models with the available data.

5. *Give a complete description of the stoichiometric compatibility classes for the chemical reaction network.* A stoichiometric compatibility class is the set of all points accessible from a given state via the reactions in the system. This question relates more closely to the dynamics of the system, but also has ramifications for the set of all steady states.

6. *What information does species concentration data give us for parameter estimation?* In particular, are the parameters identifiable? Identifiability means that having many measurements of the concentrations $\mathbf{x}$ can determine the reaction rate constants $\mathbf{k}$. If not identifiable, we will explore algebraic constraints imposed by the species concentration data. This question is relevant for complete and partial steady-state data (usually noisy).

These questions are open challenges for medium to large models in systems biology and medicine [27, 54]. The book chapter [36] illustrates standard mathematical and statistical methods for addressing these questions, with Wnt signaling as a case study. Here, we examine these questions from the perspective of algebraic geometry. The aim is to provide insight into global behavior by applying tools from nonlinear algebra to synthetic and systems biology. Below are the algebraic problems underlying the four biological problems listed above.

## Algebraic Problems

3. Describe the set of points $(\mathbf{k}, \mathbf{c}) \in \mathbb{R}^{31}_{>0} \times \mathbb{R}^{5}_{>0}$ such that the polynomials (4.1)-(4.2) have two or more positive zeros $\mathbf{x} \in \mathbb{R}^{19}_{>0}$. When is there only one? Identify the discriminant.

4. Which projections of the variety defined by (4.1) into coordinate subspaces of $K^{19}$ are surjective? Equivalently, describe the algebraic matroid on the ground set $\{x_1, \ldots, x_{19}\}$.

5. The conservation relations (4.2) specify a linear map $\chi : \mathbb{R}^{19} \to \mathbb{R}^5$, $\mathbf{x} \mapsto \mathbf{c}$. Describe all the convex polyhedra $\chi^{-1}(\mathbf{c}) \cap \mathbb{R}^{19}_{\geq 0}$ where $\mathbf{c}$ runs over the points in the open orthant $\mathbb{R}^5_{>0}$.

6. a. *Complete data:* Describe the matroid on the ground set $\{k_1, k_2, \ldots, k_{31}\}$ that is defined by the linear forms on the right hand sides of (4.1), for fixed steady-state concentrations.

b. *Partial steady-state data without noise:* Repeat the analysis after eliminating some of the $\mathbf{x}$-coordinates.

c. *Partial steady-state data with noise:* For the remaining $\mathbf{x}$-coordinates, suppose that we have data which are *approximately* on the projected steady state variety. Determine a parameter vector $(\mathbf{k}, \mathbf{c})$ that best fits the data.

In this paper we shall address these questions, and several related ones, after explaining the various ingredients. A particular focus is the exchange between the algebraic formulation and its biological counterpart. Our presentation is organized as follows.

In Section 4.1 we review the basics on the Wnt signaling pathway, we recall the shuttle model of MacLean *et al.* [37], and we derive the dynamical system (4.1)–(4.2). In Section 4.2 we establish Theorem 4.0.3, and we examine the set of all steady states. This is here regarded as a complex algebraic variety in an affine space of dimension $55 = 19 + 31 + 5$ with coordinates $(\mathbf{x}, \mathbf{k}, \mathbf{c})$.

In Sections 4.3, 4.4, 4.5, and 4.6 we address the four problems stated above. The numbers of the problems refer to the respective sections. Each section starts out with an explanation of how the biological problem and the algebraic problem are related. The rationale behind Section 4.3 is likely to be familiar, given that multistationarity has been discussed widely in the literature; see e.g. [8, 40]. On the other hand, in Section 4.4 we use the language of algebraic matroids, which are new tools for chemical reaction networks. Section 4.5 characterizes the polyhedral geometry encoded in the conservation relations (4.2). This is a case study in the spirit of [49, Figure 1]. Section 4.6 addresses the problems of parameter identifiability and parameter estimation. Finally, in Section 4.7 we return to the biology, and we discuss what our findings might imply for the study of Wnt signaling and other systems.

## 4.1 From Biology to Algebra

Cellular decisions such as cell division, specialization and cell death are governed by a rich repertoire of complex signals that are produced by other cells and/or stimuli. In order for a cell to come to an appropriate decision, it must *sense* its external environment, communicate this information to the nucleus, and respond by regulating genes and producing relevant proteins. Signaling molecules called ligands, external to the cell, can bind to proteins called receptors, initializing the propagation of information within the cell by molecular interactions and modifications (e.g. phosphorylation). This signal may be relayed from the cytoplasm into the nucleus via molecules and the cell responds by activation or deactivation of gene(s) that control, for example, cell fate. The complex interplay of molecules involved in this information transmission is called a signaling transduction pathway. Although many signaling pathways have been defined biochemically, much is still not understood about them or how a signal results in a particular cellular response. Mathematical models constructed at different

scales of molecular complexity may help unravel the central mechanisms that govern cellular decisions, and their analysis may inform and guide testable hypotheses and therapies.

This chapter will focus on the canonical Wnt signaling pathway, which is involved in cellular processes, both during development and in adult tissues. This includes stem cells. Dysfunction of this pathway has been linked to neurodegenerative diseases and cancer. Consequently, Wnt signaling has been widely studied in various organisms, including amphibians and mammals. Researchers are interested in how the extracellular ligand Wnt affects the protein $\beta$-catenin, which plays a pivotal role in turning genes on and off in the nucleus.

The molecular interactions within the Wnt signaling pathway are not yet fully understood. This has led to the development and analysis of many mathematical models. The Wnt shuttle model [37] includes an abstraction of the signal transduction pathway (via activation/inactivation of molecules) described above. The model also takes into account molecules that exist, interact and move between different compartments in the cell (e.g., cytoplasm and nucleus). Biologists understand the Wnt system as either *Wnt off* or *Wnt on*. However, such a scenario is rarely binary (i.e., different concentration levels of Wnt may exist) and inherently depends on spatial movement of molecules. The Wnt shuttle model includes complex interactions with nonlinearities arising in the equations. In particular, it includes both the Wnt off and Wnt on scenarios, by adjusting initial conditions or parameter values. The biology needed to understand the model can be described as follows. See also Table 4.1.

*Wnt off:* When cells do not sense the extracellular ligand Wnt, $\beta$-catenin is degraded (broken down). The degradation of $\beta$-catenin is partially dependent on a group of molecules (Axin, APC and GSK-3) that form the *destruction complex*. Crucially, the break down of $\beta$-catenin occurs when the destruction complex is in an active state; modification to the destruction complex by proteins, called phosphatases, changes it from inactive to active. Additionally, $\beta$-catenin can degrade independent of the destruction complex. Synthesis of $\beta$-catenin occurs at a constant rate.

*Wnt on:* When receptors on the surface of a cell bind to Wnt, the Wnt signaling transduction pathway is initiated. This enables $\beta$-catenin to move into the nucleus where it binds with transcription factors that regulate genes. This signal propagation is mediated by the following molecular interactions. After Wnt stimulus, the protein Dishevelled is activated near the membrane. This in turn inactivates the destruction complex, thereby preventing the destruction of $\beta$-catenin, allowing it to accumulate in the cytoplasm through natural synthesis. Throughout the molecular interactions in the signaling pathway, intermediate complexes can form (e.g., $\beta$-catenin bound with Dishevelled).

*Space:* The location of molecules plays a pivotal role: $\beta$-catenin moves between the cytoplasm and the nucleus (to reach target genes and regulate them). Dishevelled and molecules that form the destruction complex shuttle between the nucleus and the cytoplasm. However, it is assumed that only the inactive destruction complex can shuttle (since in the cytoplasm it would be bound to $\beta$-catenin). Phosphatases exist in both the nucleus and the cytoplasm but the movement across compartments is not included in the model. Symmetry of reactions is assumed if the species exist in both compartments. Intermediate complexes

are assumed to be short-lived, or not large enough for movement across compartments.

The Wnt shuttle model of [37] has 19 species whose interactions can be framed as biochemical reactions. These species correspond to variables $x_1, \ldots, x_{19}$ in our dynamical system (4.1). Namely, $x_i$ represents the concentration of the species that is listed in the $i$th row in Table 4.1.

| Variable | Species | Symbol |
|---|---|---|
| | **Dishevelled** | $\boldsymbol{D}$ |
| $x_1$ | Dishevelled in cytoplasm (inactive) | $D_i$ |
| $x_2$ | Dishevelled in cytoplasm (active) | $D_a$ |
| $x_3$ | Dishevelled in nucleus (active) | $D_{an}$ |
| | **Destruction complex (APC/Axin/GSK3$\beta$)** | $\boldsymbol{Y}$ |
| $x_4$ | Destruction complex in cytoplasm (active) | $Y_a$ |
| $x_5$ | Destruction complex in cytoplasm (inactive) | $Y_i$ |
| $x_6$ | Destruction complex in nucleus (active) | $Y_{an}$ |
| $x_7$ | Destruction complex in nucleus (inactive) | $Y_{in}$ |
| | **Phosphatase** | $\boldsymbol{P}$ |
| $x_8$ | Phosphatase in cytoplasm | $P$ |
| $x_9$ | Phosphatase in nucleus | $P_n$ |
| | **$\boldsymbol{\beta}$− catenin** | $\boldsymbol{x}$ |
| $x_{10}$ | $\beta$-catenin in cytoplasm | $x$ |
| $x_{11}$ | $\beta$-catenin in nucleus | $x_n$ |
| | **Transcription Factor** | $\boldsymbol{T}$ |
| $x_{12}$ | TCF (gene transcription in nucleus) | $T$ |
| | **Intermediate complex** | $\boldsymbol{C}$ |
| $x_{13}$ | Transcription complex, $\beta$-catenin: TCF in nucleus | $C_{xT}$ |
| $x_{14}$ | Intermediate complex, $\beta$-catenin: dishevelled in cytoplasm | $C_{YD}$ |
| $x_{15}$ | Intermediate complex, destruction complex: dishevelled in nucleus | $C_{YDn}$ |
| $x_{16}$ | Intermediate complex, destruction complex: phosphatase in cytoplasm | $C_{YP}$ |
| $x_{17}$ | Intermediate complex, destruction complex: phosphatase in nucleus | $C_{YPn}$ |
| $x_{18}$ | Intermediate complex, $\beta$-catenin: destruction complex in cytoplasm | $C_{xY}$ |
| $x_{19}$ | Intermediate complex, $\beta$-catenin: destruction complex in nucleus | $C_{xYn}$ |

Table 4.1: The 19 species in the Wnt shuttle model.

The second column in Table 4.1 indicates the biological meaning of the 19 species. The symbols in the last column are those used in the presentation of the Wnt shuttle model in [37].

The 19 species in the model interact according to the 31 reactions given in Table 4.2. Each reaction comes with a rate constant $k_i$. These are the coordinates of our parameter vector $\mathbf{k}$.

| **Reaction** | **Explanation** |
|---|---|
| $x_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} x_2$ | (In)activation of dishevelled, depends on Wnt |
| $x_2 + x_4 \underset{k_4}{\overset{k_3}{\rightleftharpoons}} x_{14} \xrightarrow{k_5} x_2 + x_5$ | Destruction complex active $\to$ inactive |
| $x_5 + x_8 \underset{k_7}{\overset{k_6}{\rightleftharpoons}} x_{16} \xrightarrow{k_8} x_4 + x_8$ | Destruction complex inactive $\to$ active |
| $x_4 + x_{10} \underset{k_{10}}{\overset{k_9}{\rightleftharpoons}} x_{18} \xrightarrow{k_{11}} x_4 + \emptyset$ | Destruction complex-dependent $\beta$-catenin degradation |
| $\emptyset \xrightarrow{k_{12}} x_{10}$ | $\beta$-catenin production |
| $x_{10} \xrightarrow{k_{13}} \emptyset$ | Destruction complex-independent $\beta$-catenin degradation |
| $x_3 + x_6 \underset{k_{15}}{\overset{k_{14}}{\rightleftharpoons}} x_{15} \xrightarrow{k_{16}} x_3 + x_7$ | Destruction complex active $\to$ inactive (nucleus) |
| $x_7 + x_9 \underset{k_{18}}{\overset{k_{17}}{\rightleftharpoons}} x_{17} \xrightarrow{k_{19}} x_6 + x_9$ | Destruction complex inactive $\to$ active (nucleus) |
| $x_6 + x_{11} \underset{k_{21}}{\overset{k_{20}}{\rightleftharpoons}} x_{19} \xrightarrow{k_{22}} x_6 + \emptyset$ | Destruction complex-dependent $\beta$-catenin degradation (nucleus) |
| $x_{11} \xrightarrow{k_{23}} \emptyset$ | Destruction complex-independent $\beta$-catenin degradation (nucleus) |
| $x_{11} + x_{12} \underset{k_{25}}{\overset{k_{24}}{\rightleftharpoons}} x_{13}$ | $\beta$-catenin binding to TCF (nucleus) |
| $x_2 \underset{k_{27}}{\overset{k_{26}}{\rightleftharpoons}} x_3$ | Shuttling of active dishevelled |
| $x_5 \underset{k_{29}}{\overset{k_{28}}{\rightleftharpoons}} x_7$ | Shuttling of inactive-form destruction complex |
| $x_{10} \underset{k_{31}}{\overset{k_{30}}{\rightleftharpoons}} x_{11}$ | Shuttling of $\beta$-catenin |

Table 4.2: The 31 reactions in the Wnt shuttle model.

The 31 reactions in Table 4.2 translate into a dynamical system $\dot{\mathbf{x}} = \Psi(\mathbf{x}; \mathbf{k})$. Here $\Psi$ is a vector-valued function of the vectors of species concentrations $\mathbf{x}$ and rate constants $\mathbf{k}$. The choice of $\Psi$ is up to the modeler. In our study, we assume that $\Psi$ represents the *law of mass action* [27, §2.1.1]. This is precisely what is used in [37] for the Wnt shuttle model. The resulting dynamical system is (4.1). We refer to [8, 14, 24, 40, 49] and their many references for mass action kinetics and its variants. In summary, Table 4.2 translates into the dynamical system (4.1) under the law of mass action. The five relations in (4.2) constitute a basis for the linear space of conservation relations of the model in Table 4.2 assuming mass action kinetics.

We refer to $x_1, \ldots, x_{19}$ as the *species concentrations*, $k_1, \ldots, k_{31}$ as the *rate parameters*, and $c_1, \ldots, c_5$ as the *conserved quantities*. We write $\mathbf{x}, \mathbf{k}$ and $\mathbf{c}$ for the vectors with these coordinates. As is customary in algebraic geometry, we take the coordinates in the complex

numbers $\mathbb{C}$, or possibly in some other algebraically closed field $K$ containing the rationals $\mathbb{Q}$.

Our aim is to understand the relationships between $\mathbf{x}, \mathbf{k}$ and $\mathbf{c}$ in the Wnt shuttle model. To this end, we introduce the *steady state variety* $\mathcal{S} \subset \mathbb{C}^{55}$. This is the set of all points $(\mathbf{x}, \mathbf{k}, \mathbf{c})$ that satisfy the equations $\dot{x}_1 = \ldots = \dot{x}_{19} = 0$ in (4.1) along with the five conservation laws in (4.2). We write our ambient affine space as $\mathbb{C}^{55} = \mathbb{C}^{19}_{\mathbf{x}} \times \mathbb{C}^{31}_{\mathbf{k}} \times \mathbb{C}^{5}_{\mathbf{c}}$. This emphasizes the distinction between the species concentrations, rate parameters, and conserved quantities.

## 4.2 Ideals, Varieties, and Nine Points

We write $I$ for the ideal in the polynomial ring $\mathbb{Q}[\mathbf{x}, \mathbf{k}] = \mathbb{Q}[x_1, \ldots x_{19}, k_1, \ldots k_{31}]$ that is generated by the 19 polynomials $\dot{x}_i$ on the right hand side of (4.1). Five of these generators are redundant. Indeed, the conservation relations (4.2) give the following identities mod $I$:

$$\begin{gathered}\dot{x}_1 + \dot{x}_2 + \dot{x}_3 + \dot{x}_{14} + \dot{x}_{15} \;=\; \dot{x}_8 + \dot{x}_{16} \;=\; \dot{x}_9 + \dot{x}_{17} \;=\; \dot{x}_{12} + \dot{x}_{13} \;= \\ \dot{x}_4 + \dot{x}_5 + \dot{x}_6 + \dot{x}_7 + \dot{x}_{14} + \dot{x}_{15} + \dot{x}_{16} + \dot{x}_{17} + \dot{x}_{18} + \dot{x}_{19} \;=\; 0.\end{gathered}$$

For instance, the polynomials $\dot{x}_{13}, \dot{x}_{15}, \dot{x}_{16}, \dot{x}_{17}$ and $\dot{x}_{19}$ are redundant because they can be expressed as negated sums of other generators of $I$. Hence $I$ is generated by 14 polynomials. The variety $V(I)$ lives in the 50-dimensional affine space $\mathbb{C}^{19}_{\mathbf{x}} \times \mathbb{C}^{31}_{\mathbf{k}}$, and it is isomorphic to the steady state variety $\mathcal{S} \subset \mathbb{C}^{55}$. A direct computation using the computer algebra package `Macaulay2` [18] shows that $V(I)$ has dimension 36. Hence the affine ideal $I$ is a complete intersection in $\mathbb{Q}[\mathbf{x}, \mathbf{k}]$. Furthermore, using `Macaulay2` we can verify the following lemma.

**Lemma 4.2.1.** *The ideal $I$ admits the non-trivial decomposition $I = I_m \cap I_e$, where $I_e = I : \langle x_1 \rangle$ and $I_m = I + \langle x_1 \rangle$, both of these components have codimension 14, and $I_e$ is a prime ideal.*

The ideal $I_m$ is called the *main component*, while $I_e$ is called the *extinction component*, since it reflects those steady states where a number of the reactants "run out." Both of these ideals live in $\mathbb{Q}[\mathbf{x}, \mathbf{k}]$, and we now present explicit generators. The extinction component equals

$$\begin{gathered}I_e = \langle x_1, x_2, x_3, x_5, x_7, x_{14}, x_{15}, x_{16}, x_{17}, k_{30}x_{10} - (k_{23} + k_{31})x_{11} - k_{22}x_{19}, \\ k_{13}x_{10} + k_{23}x_{11} + k_{11}x_{18} + k_{22}x_{19} - k_{12}, k_{24}x_{11}x_{12} - k_{25}x_{13}, \\ k_{20}x_6x_{11} - (k_{21} + k_{22})x_{19}, k_9x_4x_{10} - (k_{10} + k_{11})x_{18}\rangle.\end{gathered}$$

The ideal $I_e$ is found to be prime in $\mathbb{Q}[\mathbf{x},\mathbf{k}]$. The main component equals

$$\begin{aligned} I_m \;=\; \langle & k_{16}x_{15}-k_{19}x_{17}, k_5x_{14}-k_8x_{16}, k_{30}x_{10}-(k_{23}+k_{31})x_{11}-k_{22}x_{19}, \\ & k_{13}x_{10}+k_{23}x_{11}+k_{11}x_{18}+k_{22}x_{19}-k_{12}, k_{28}x_5-k_{29}x_7, k_{26}x_2-k_{27}x_3, \\ & k_1x_1-k_2x_2, k_{24}x_{11}x_{12}-k_{25}x_{13}, k_{20}x_6x_{11}-(k_{21}+k_{22})x_{19}, \\ & k_9x_4x_{10}-(k_{10}+k_{11})x_{18}, k_{17}x_7x_9-(k_{18}+k_{19})x_{17}, k_6x_5x_8-(k_7+k_8)x_{16}, \\ & k_{14}x_3x_6-k_{15}x_{15}-k_{19}x_{17}, k_3x_2x_4-k_4x_{14}-k_8x_{16}, \\ & (k_4k_6k_8k_{14}k_{16}k_{18}k_{26}k_{29}+k_5k_6k_8k_{14}k_{16}k_{18}k_{26}k_{29}+ \\ & k_4k_6k_8k_{14}k_{16}k_{19}k_{26}k_{29}+k_5k_6k_8k_{14}k_{16}k_{19}k_{26}k_{29})k_1x_6x_8 \\ & -(k_3k_5k_7k_{15}k_{17}k_{19}k_{27}k_{28}+k_3k_5k_8k_{15}k_{17}k_{19}k_{27}k_{28} \\ & +k_3k_5k_7k_{16}k_{17}k_{19}k_{27}k_{28}+k_3k_5k_8k_{16}k_{17}k_{19}k_{27}k_{28})k_1x_4x_9\rangle. \end{aligned}$$

This ideal is not prime in $\mathbb{Q}[\mathbf{x},\mathbf{k}]$. For instance, the variable $k_1$ is a zerodivisor modulo $I_m$, as seen from the last generator. Removing the factor $k_1$ from the last generator yields the quotient ideal $I_m : \langle k_1\rangle$. However, even that ideal still has several associated primes. All of these prime ideals, except for one, contain some of the rate constants $k_i$.

That special component is characterized in the following proposition. Given any ideal $J \subset \mathbb{Q}[\mathbf{x},\mathbf{k}]$, we write $\widetilde{J} = \mathbb{Q}(\mathbf{k})[\mathbf{x}]J$ for its extension to the polynomial ring $\mathbb{Q}(\mathbf{k})[\mathbf{x}]$ in the unknowns $x_1,\ldots,x_{19}$ over the field of rational functions in the parameters $k_1,\ldots,k_{31}$.

**Proposition 4.2.2.** *The ideal $J_m = \widetilde{I_m} \cap \mathbb{Q}[\mathbf{x},\mathbf{k}]$ is prime. Its irreducible variety $V(J_m) \subset \mathbb{C}^{50}$ has dimension 36; it is the unique component of $V(I_m)$ that maps dominantly onto $\mathbb{C}^{31}_{\mathbf{k}}$.*

*Proof.* The ideal $\widetilde{I_m}$ has the same generators as $I_m$ but now regarded as polynomials in $\mathbf{x}$ with coefficients in $\mathbb{Q}(\mathbf{k})$. Symbolic computation in the ring $\mathbb{Q}(\mathbf{k})[\mathbf{x}]$ reveals that $\widetilde{I_m}$ is a prime ideal. This implies that $J_m$ is a prime ideal in $\mathbb{Q}[\mathbf{x},\mathbf{k}]$, and hence $V(J_m)$ is irreducible. The dimension statement follows from the result of Lemma 4.2.1 that $I_m$ is a complete intersection. This ensures that $V(I_m)$ has no lower-dimensional components, by Krull's Principal Ideal Theorem. Finally, $V(J_m)$ maps dominantly onto $\mathbb{C}^{31}_{\mathbf{k}}$ because $J_m \cap \mathbb{Q}[\mathbf{k}] = \{0\}$. □

**Corollary 4.2.3.** *The ideal $\widetilde{I}$ is radical, and it is the intersection of two primes in $\mathbb{Q}(\mathbf{k})[\mathbf{x}]$:*

$$\widetilde{I} \;=\; \widetilde{I_e} \cap \widetilde{I_m}. \tag{4.3}$$

*Proof.* This follows directly from Proposition 4.2.2 and the primality of $I_e$ in Lemma 4.2.1. □

The decomposition has the following geometric interpretation. We now work over the field $K = \overline{\mathbb{Q}(\mathbf{k})}$. All rate constants are taken to be generic. Then $V(\widetilde{I})$ is the 5-dimensional variety of all steady states in $K^{19}$. This variety is the union of two irreducible components,

$$V(\widetilde{I}) \;=\; V(\widetilde{I_e}) \cup V(\widetilde{I_m}),$$

where each component is 5-dimensional. The first component lies inside the 10-dimensional coordinate subspace $V(x_1, x_2, x_3, x_5, x_7, x_{14}, x_{15}, x_{16}, x_{17})$. Hence it is disjoint from the hyperplane defined by the first conservation relation $x_1 + x_2 + x_3 + x_{14} + x_{15} = c_1$. In other words, $V(\widetilde{I_e})$ is mapped into a coordinate hyperplane under the map $\chi : K^{19} \to K^5, \mathbf{x} \mapsto \mathbf{c}$.

On the other hand, the second component $V(\widetilde{I_m})$ maps dominantly onto $K^5$ under $\chi$. Theorem 4.0.3 states that the generic fiber of this map consists of 9 reduced points. Equivalently,

$$\chi^{-1}(\mathbf{c}) \cap V(\widetilde{I}) \,=\, \chi^{-1}(\mathbf{c}) \cap V(\widetilde{I_m}) \tag{4.4}$$

is a set of nine points in $K^{19}$. We are now prepared to argue that this is indeed the case.

*Computational Proof of Theorem 4.0.3.* We consider the ideal of the variety (4.4) in the polynomial ring $\mathbb{Q}(\mathbf{k}, \mathbf{c})[\mathbf{x}]$. This polynomial ring has 19 variables, and all 36 parameters are now scalars in the coefficient field. This ideal is generated by the right hand sides of (4.1) and (4.2). Performing a Gröbner basis computation in this polynomial ring verifies that our ideal is zero-dimensional and has length 9. Hence (4.4) is a reduced affine scheme of length 9 in $K^{19}$.

Fast numerical verification of this result is obtained by replacing the coordinates of $\mathbf{k}$ and $\mathbf{c}$ with generic (random rational) values. In `Macaulay2` one finds, with probability 1, that the resulting ideals in $\mathbb{Q}[\mathbf{x}]$ are radical of length 9. We also verified this result via *numerical algebraic geometry*, using the two software packages `Bertini` [4] and `PHCpack` [53]. □

## 4.3 Multistationarity and its Discriminant

This section centers around Question 3 from the Introduction: *For what real positive rate parameters and conserved quantities does the system exhibit multistationarity?* This is commonly asked about biochemical reaction networks and about dynamical systems in general.

Mathematically, this is a problem of *real* algebraic geometry. Writing $\mathcal{S}$ for the steady state variety in $\mathbb{C}^{55}$, we are interested in the fibers of the map $\pi_{\mathbf{k},\mathbf{c}} : \mathcal{S} \cap \mathbb{R}^{55}_{>0} \to \mathbb{R}^{31}_{>0,\mathbf{k}} \times \mathbb{R}^{5}_{>0,\mathbf{c}}$. According to Theorem 4.0.3, the general fiber consists of 9 *complex* points $\mathbf{x} \in \mathbb{C}^{19}_{\mathbf{x}}$, when the map $\pi_{\mathbf{k},\mathbf{c}}$ is taken over $\mathbb{C}$. But here we take it over the reals $\mathbb{R}$ or over the positive reals $\mathbb{R}_{>0}$.

In our application to biology, we only care about concentration vectors $\mathbf{x}$ whose coordinates are real and positive. Thus we wish to stratify $\mathbb{R}^{31}_{>0,\mathbf{k}} \times \mathbb{R}^{5}_{>0,\mathbf{c}}$ according to the cardinality of

$$\pi^{-1}_{\mathbf{k},\mathbf{c}}(\mathbf{k}, \mathbf{c}) \,=\, \big\{\, (\mathbf{x}, \mathbf{k}', \mathbf{c}') \in \mathcal{S} \,\cap\, \mathbb{R}^{55}_{>0} \;:\; \mathbf{k}' = \mathbf{k} \text{ and } \mathbf{c}' = \mathbf{c} \big\}. \tag{4.5}$$

This stratification comes from a decomposition of the 36-dimensional orthant $\mathbb{R}^{31}_{>0,\mathbf{k}} \times \mathbb{R}^{5}_{>0,\mathbf{c}}$ into connected open semialgebraic subsets. The walls in this decomposition are given by the *discriminant* $\Delta$, a giant polynomial in the 36 unknowns $(\mathbf{k}, \mathbf{c})$ that is to be defined later.

We begin with the following result on what is possible with regard to real positive solutions.

**Theorem 4.3.1.** *Consider the polynomial system in (4.1)–(4.2) where all parameters $k_i$ and $c_j$ are positive real numbers. The set (4.5) of positive real solutions can have* $1, 2,$ *or* $3$ *elements.*

*Proof.* For random choices of $(\mathbf{k}, \mathbf{c}) = (k_1, \ldots, k_{31}, c_1, \ldots, c_5)$ in the orthant $\mathbb{R}^{36}_{>0}$, our polynomial system has 9 complex solutions, by Theorem 4.0.3. For the following two special choices of the 36 parameter values, all 9 solutions are real. First, take $(\mathbf{k}, \mathbf{c})$ to be the vector

$$(1.7182818, 53.2659, 3.4134082, 0.61409879, 0.61409879, 3.4134082, 0.98168436, 0.98168436, 92.331732, 0.86466471, 79.9512906, 97.932525, 1, 3.2654672, 0.61699064, 0.61699064, 37.913879, 0.86466471, 0.86466471, 4.7267833, 0.17182818, 0.68292191, 1, 0.55950727, 1.0117639, 1.7182818, 1.7182818, 0.99326205, 0.99326205, 5.9744464, 1, 4.9951026, 16.4733784, 1.6006340000000001, 1.2089126, 2.7756596399999998).$$

The resulting system has three positive solutions $\mathbf{x} \in \mathbb{R}^{19}_{>0}$. Next, let $(\mathbf{k}', \mathbf{c}')$ be the vector

$$(0.948166, 7.45086, 5.72974, 3.96947, 7.21145, 7.8761, 1.87614, 8.11372, 6.21862, 5.24801, 3.10707, 1.08146, 5.22133, 5.84158, .911392, 4.28788, 4.81201, 9.67849, 1.34452, 7.38597, 6.64451, 7.10229, 8.57942, 5.79076, 6.33244, 1.53916, 1.39658, 0.81673, 5.8434, 3.86223, 7.22696, 1.45438, 3.36482, 6.06453, 4.82045, 3.6014).$$

Here, one solution to our system is positive. By connecting the two parameter points above with a general curve in $\mathbb{R}^{36}_{>0}$, and by examining in-between points $(\mathbf{k}'', \mathbf{c}'')$, we can construct a system with two positive solutions. All computations were carried out using `Bertini` [4]. □

**Remark 4.3.2.** *At present, we do not know whether the number of real positive solutions can be larger than three. We suspect that this is impossible, but we currently cannot prove it.*

The difficulty lies in the fact that the stratification of $\mathbb{R}^{36}_{>0}$ is extremely complicated. In computer algebra, the derivation of such stratifications is known as the problem of *real root classification*. For a sample of recent studies in this direction see [7, 13, 47]. Real root classification is challenging even when the number of parameters is 3 or 4; clearly, 36 parameters is out of the question. The stratification of $\mathbb{R}^{36}_{>0}$ by behavior of (4.5) has way too many cells.

While symbolic techniques for real root classification are infeasible for our system, we can use numerical algebraic geometry [19] to gain insight into the stratification of $\mathbb{R}^{36}_{>0}$. *Coefficient-parameter homotopies* [41] can solve the steady state polynomial system (4.1)-(4.2) for multiple choices of $(\mathbf{k}, \mathbf{c})$ quickly. For our computations we use `Bertini.m2`. This is the `Bertini` interface for `Macaulay2`, as described in [2]. Each system has 19 equations in 19 unknowns and, for random $(\mathbf{k}, \mathbf{c})$, each system has 9 complex solutions. Such a system can be solved in less than one second using the `bertiniParameterHomotopy` function from `Bertini.m2`.

Below we describe the following experiment. We sample $10,000$ parameter vectors $(\mathbf{k}, \mathbf{c})$ from two different probability distributions on $\mathbb{R}^{36}_{>0}$. In each case we report the observed

frequencies for the number of real solutions and number of positive solutions. We then follow these experiments with a specialized sampling scheme for testing numerical robustness.

*Uniform sampling scheme:* Here we choose $(\mathbf{k}, \mathbf{c})$ uniformly from the cube $(0.0, 100.0)^{36}$. Sampling 10,000 parameter vectors from this scheme and solving the steady state system for each of these parameter vectors in `Bertini`, we obtained $9,992$ solutions sets that contained 9 complex points. Solution sets with less than 9 points occur when some paths in the coefficient-parameter homotopy fail. We call solution sets with 9 solutions *good*.

*Integer sampling scheme:* Here we select $(\mathbf{k}, \mathbf{c})$ uniformly from $\{1, 2, 3\}^{36}$. Sampling 10,000 parameter vectors according to this scheme and solving the corresponding steady state system returned $9,963$ good solution sets. Below is a table that records how many of the good solution sets had $9, 7, 5, 3$ real solutions; all solution sets had 1 positive real solution.

| # of real solutions | 9 | 7 | 5 | 3 |
|---|---|---|---|---|
| Freq. for Uniform Sampling | 5,760 | 3,675 | 544 | 13 |
| Freq. for Integer Sampling | 2,138 | 5,181 | 2,522 | 122 |

Table 4.3: Frequencies for the sampling schemes.

These computations indicate that for most parameter vectors in $(0, 100)^{36}$ we will see only one positive solution to the steady state system. But while the set of parameter vectors that result in multiple steady states is not very large, we can give evidence that multistationarity is preserved under small perturbations. This is our next point.

*Testing Robustness:* Let $(\mathbf{k}^*, \mathbf{c}^*)$ be the first point in the proof of Theorem 4.3.1. For each index $i \in \{1, \ldots, 19\}$ we choose $y_i$ uniformly from $(-0.03 \cdot k_i^*, 0.03 \cdot k_i^*)$ then set $k_i = k_i^* + y_i$. We ran the same process for the $c_i$. Sampling $10,000$ parameter vectors this way and solving the corresponding steady state systems returned $10,000$ good solution sets, as follows:

| # of real solutions | Freq. | | # of pos. solutions | Freq. |
|---|---|---|---|---|
| 9 | 9,879 | | 3 | 9,879 |
| 7 | 121 | | 1 | 121 |

Table 4.4: Frequencies for testing robustness scheme.

In the remainder of this section, we properly define the discriminant $\Delta$ that separates the various strata in $\mathbb{R}^{36}_{>0}$. Let $\Delta_{\text{int}}$ denote the Zariski closure in $\mathbb{C}^{31}_{\mathbf{k}} \times \mathbb{C}^{5}_{\mathbf{c}}$ of all parameter vectors $(\mathbf{k}, \mathbf{c})$ for which (4.1)–(4.2) does not have 9 isolated complex solutions and there are no solutions with $x_i = 0$ for some $i$. It can be shown that $\Delta_{\text{int}}$ is a hypersurface that is defined over $\mathbb{Q}$, so it is given by a unique (up to sign) irreducible squarefree polynomial in $\mathbb{Z}[\mathbf{k}, \mathbf{c}]$. We use the symbol $\Delta_{\text{int}}$ also for that polynomial. To be precise, $\Delta_{\text{int}}$ is the discriminant of a number field $L$ with $K \supset L \supset \mathbb{Q}$, namely $L$ is the field of definition of the finite $K$-scheme (4.4).

Next, for any $i \in \{1, 2, \ldots, 19\}$ consider the intersection of the steady state variety $\mathcal{S}$ with the hyperplane $\{x_i = 0\}$. The Zariski closure of the image of $\mathcal{S} \cap \{x_i = 0\}$ under the map $\pi_{\mathbf{k},\mathbf{c}}$ is a hypersurface in $\mathbb{C}^{19}_{\mathbf{k}} \times \mathbb{C}^{31}_{\mathbf{c}}$, defined over $\mathbb{Q}$, and we write $\Delta_{x_i=0}$ for the unique

(up to sign) irreducible polynomial in $\mathbb{Z}[\mathbf{k}, \mathbf{c}]$ that vanishes on that hypersurface. We now define

$$\Delta \;:=\; \Delta_{\text{int}} \cdot \text{lcm}\big(\, \Delta_{x_1=0}\,,\, \Delta_{x_2=0}\,,\, \ldots\,,\, \Delta_{x_{19}=0}\,\big).$$

This product with a least common multiple (lcm) is the *discriminant* for our problem.

**Example 4.3.3.** *The degree of* $\Delta_{\text{int}}$ *as a polynomial only in* $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5)$ *equals* 34. *To illustrate this, we set* $\mathbf{c} = \big(5, 16 + C, \frac{8}{5} - C, \frac{6}{5} + C, 3 - C\big)$ *where* $C$ *is a parameter, and*

$$\mathbf{k} \;=\; \left(\frac{9}{5}, \frac{9}{5}, 3, \frac{2}{3}, \frac{2}{3}, 3, 1, 1, 100, \frac{4}{5}, 80, 100, 1, 3, \frac{2}{3}, \frac{2}{3}, 38, \frac{4}{5}, \frac{4}{5}, 4, \frac{1}{8}, \frac{3}{5}, 1, \frac{1}{2}, 19, \frac{7}{4}, \frac{7}{4}, 1, 1, 5, 1\right).$$

*Under this specialization, the polynomial* $\Delta_{\text{int}}$ *becomes an irreducible polynomial of degree* 34 *in the parameter* $C$. *Its coefficients are enormously large integers. It has* 14 *real roots.*

*For the other factors* $\Delta_{x_i=0}$ *of the discriminant, we find the following specializations:*

$$\begin{array}{c} x_1 \to 0,\ x_2 \to 0,\ x_3 \to 0,\ x_4 \to (C{+}16)(5C{-}8),\ x_5 \to C{+}16,\ x_6 \to (C{+}16)(5C{+}6), \\ x_7 \to C{+}16, x_8 \to 5C - 8,\ x_9 \to 5C + 6,\ x_{10} \to \textit{a quartic}\, q(C),\ x_{11} \to 0,\ x_{12} \to C{-}3, \\ x_{13} \to C{-}3,\ x_{14} \to (C{+}16)(5C{-}8),\ x_{15} \to (C{+}16)(5C{+}6),\ x_{16} \to (C{+}16)(5C{-}8), \\ x_{17} \to (C{+}16)(5C{+}6),\ \ x_{18} \to (C{+}16)(5C{-}8)q(C),\ \ x_{19} \to (C{+}16)(5C{+}6). \end{array} \tag{4.6}$$

*These polynomials have* 8 *distinct real roots in total, so the total number of real roots of the discriminant is* $14 + 8 = 22$. *These are the break points where real root behavior changes:*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $(9,0)$ | $\mathbf{-77.2388}$ | $(9,0)$ | $\mathbf{-16.0000}$ | $(9,0)$ | $-5.28669$ | $(7,0)$ | $-1.57472$ |
| $(9,0)$ | $\mathbf{-1.46506}$ | $(9,0)$ | $-1.34899$ | $(7,0)$ | $-1.29581$ | $(9,0)$ | $\mathbf{-1.20000}$ |
| $(9,1)$ | $\mathbf{-1.19215}$ | $(9,1)$ | $-1.18389$ | $(7,1)$ | $-0.584325$ | $(9,3)$ | $-0.361808$ |
| $(7,3)$ | $0.191039$ | $(5,1)$ | $1.30812$ | $(7,1)$ | $1.33197$ | $(5,1)$ | $\mathbf{1.60000}$ |
| $(5,0)$ | $1.60161$ | $(3,0)$ | $\mathbf{3.0000}$ | $(3,0)$ | $4.26306$ | $(5,0)$ | $11.1174$ |
| $(7,0)$ | $21.4165$ | $(9,0)$ | $\mathbf{310.141}$ | $(9,0)$ | | | |

*In this table, we list all* 22 *roots of the specialized discriminant* $\Delta(C)$. *The eight boldface values of* $C$ *are the roots of (4.6): here one of the coordinates of* $\mathbf{x}$ *becomes zero. At the other* 14 *values of* $C$, *the number of real roots changes. Between any two roots we list the pair* $(r, p)$, *where* $r$ *is the number of real roots and* $p$ *is the number of positive real roots. For instance, for* $-0.361808 < C < 0.191039$, *there are* 7 *real roots of which* 3 *are positive.*

## 4.4 Algebraic Matroids and Parametrizations

Question 4 asks: *Suppose we can measure only a subset of the species concentrations. Which subsets can lead to model rejection?* This issue is important for the Wnt shuttle model because, in the laboratory, only some of the species are measurable by existing techniques.

We shall address Question 4 using *algebraic matroids*. Matroid theory allows us to analyze the structure of relationships among the 19 species in Table 4.1. This first appeared in [37]. We here present an in-depth study of the matroids that govern the Wnt shuttle model.

We are here interested in the matroid that is defined by the prime ideal $P = \widetilde{I_m}$ in $\mathbb{Q}(\mathbf{k})[\mathbf{x}]$. Its ground set $X$ is the set of species concentrations $\{x_1, \ldots, x_{19}\}$. Since $V(\widetilde{I_m})$ is 5-dimensional, each basis consists of five elements in $X$. In our application, bases are the maximal subsets of $X$ that can be specified independently at steady state; they are also the minimal-cardinality sets that can be measured to learn all species concentrations. The rank of a set $Y$ indicates the number of measurements required to learn the concentrations for every element of $Y$. Flats are the full subsets that are specified by any given collection of measurements.

Circuits furnish our answer to Question 4: they are minimal sets of species that can be used to test compatibility of the data with the model. For each circuit $Y$ there is a unique-up-to-scalars relation in $\widetilde{I_m} \cap Q(\mathbf{k})[Y]$, called the *circuit polynomial* of $Y$. If the measurements indicate that this relation is not satisfied, then the model and data are not compatible.

**Proposition 4.4.1.** *The algebraic matroid of $\widetilde{I_m}$ has rank 5. It has 951 circuits, summarized in Table 4.5. Of the 11628 subsets of $X$ of size 5, precisely 2389 are bases. The 2092 bases summarized in Table 4.6 have base degree 1, while the remaining 297 have base degree 2.*

The computation of this matroid was carried out using the methods described in [48]. It was first reported in [37], along with the matroids of alternative models for the Wnt pathway. The idea there was to find subsets of variables that were dependent for different models.

Our matroid analysis here goes beyond [37] in several ways:

1. We keep track of the parameters $\mathbf{k}$. We take our circuit polynomials to have (relatively prime) coefficients in $\mathbb{Z}[\mathbf{k}]$. This gives us a new tool for model rejection, e.g. in situations where only one data point is known but some parameter values are available.

2. We show how circuits can be used in parameter estimation; this will be done in Section 4.7.

3. We use the degree-1 bases to derive rational parametrizations of the variety $V(\widetilde{I_m})$.

We now explain Table 4.5. A circuit polynomial has *type* $(i, j)$ if it contains $i$ species concentrations ($\mathbf{x}$-variables) and $j$ rate parameters ($\mathbf{k}$-variables). The entry in row $i$ and column $j$ in Table 4.5 is the number of circuits of type $(i, j)$. Zero values are omitted for clarity.

**Example 4.4.2.** *There are five circuits of type $(2, 2)$. One of them is $\dot{x}_1 = -k_1x_1 + k_2x_2$. Most of the 951 circuit polynomials in $\widetilde{I_m}$ are more complicated. In particular, they are non-linear in both $\mathbf{x}$ and $\mathbf{y}$. For instance, the unique circuit polynomial of type $(6, 11)$ equals*

$$\begin{aligned}&(-k_{15}k_{17}k_{19}k_{20}k_{25} - k_{16}k_{17}k_{19}k_{20}k_{25})x_7x_9x_{13}\\ &+(k_{14}k_{16}k_{18}k_{21}k_{24} + k_{14}k_{16}k_{19}k_{21}k_{24} + k_{14}k_{16}k_{18}k_{22}k_{24} + k_{14}k_{16}k_{19}k_{22}k_{24})x_3x_{12}x_{19}.\end{aligned}$$

*In Section 4.6, we will consider the role of these nonlinear functions in parameter estimation.*

Given a basis $Y$ of an algebraic matroid, its *base degree* is the length of the generic fiber of the projection of $V(P)$ onto the $Y$-coordinates (cf. [48]). Bases with degree 1 are desirable:

| | 2 | 3 | 4 | 5 | 6 | | 2 | 3 | 4 | 5 | 6 | | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 1 | | | | 12 | | | 13 | 10 | | 22 | | | 8 | 58 | |
| 3 | | 6 | | | | 13 | | | 13 | 15 | 2 | 23 | | | 4 | 56 | 5 |
| 4 | 1 | 5 | | | | 14 | | | 19 | 16 | 1 | 24 | | | | 54 | 14 |
| 5 | | 6 | 1 | | | 15 | | | 17 | 21 | 4 | 25 | | | | 53 | 15 |
| 6 | | 7 | 5 | | | 16 | | | 15 | 11 | 2 | 26 | | | | 8 | 16 |
| 7 | | 5 | 3 | | | 17 | | | 16 | 32 | 9 | 27 | | | 12 | 56 | 16 |
| 8 | | 1 | 11 | 1 | | 18 | | | 4 | 6 | 2 | 28 | | | 2 | | 2 |
| 9 | | 6 | 12 | 3 | | 19 | | | 26 | 36 | 11 | 29 | | | | 29 | 14 |
| 10 | | | 11 | 1 | | 20 | | | 44 | 1 | 1 | 30 | | | | | |
| 11 | | 4 | 7 | 11 | 1 | 21 | | | 26 | 27 | 9 | 31 | | | | | 6 |

Table 4.5: The 951 circuit polynomials, by numbers of unknowns $x_i$ and $k_j$.

**Proposition 4.4.3.** *Let $P \subset K[X]$ be a prime ideal, $Y$ a basis of its algebraic matroid, $|X| = n$, and $|Y| = r$. If $Y$ has base degree 1 then $V(P)$ is a rational variety, and the basic circuits of $Y$ specify a birational map $\varphi_Y : K^r \dashrightarrow K^n$ whose image is Zariski dense in $V(P)$*

*Proof.* For each coordinate $x_i$ in $X \backslash Y$ there exists a circuit containing $Y \cup \{x_i\}$; this is the *basic circuit* of $(Y, x_i)$. Since $Y$ has base degree 1, the generic fiber of the map $V(P) \to K^r$ consists of a unique point. Therefore the circuit polynomial is linear in $x_i$. It has the form

$$p_i(Y) \cdot x_i + q_i(Y), \qquad \text{where } p_i, q_i \in K[Y].$$

The $i$-coordinate of the rational map $\varphi_Y$ equals $x_i$ if $x_i \in Y$ and $-q_i(Y)/p_i(Y)$ if $x_i \notin Y$. $\square$

From Propositions 4.4.1 and 4.4.3, we obtain 2092 rational parametrizations of the variety $V(\widetilde{I_m})$. These are the maps $\varphi_Y : K^5 \dashrightarrow K^{19}$, where $Y$ runs over all bases of base degree 1. Using these $\varphi_Y$, we obtain 2092 representations of the steady state variety (4.4) as a subset of $K^5$, where now $K = \mathbb{Q}(\mathbf{k}, \mathbf{c})$. Namely, we consider the preimages of the five hyperplanes defined by (4.2). These are hypersurfaces in $K^5$ whose intersection represents the nine points in (4.4). We performed the following computation for all 2092 bases $Y = \{y_1, \ldots, y_5\}$ of base degree 1:

1. Substitute $\mathbf{x} = \varphi_Y(y_1, \ldots, y_5)$ into the five linear equations (4.2).

2. Clear the denominators $d_1, \ldots, d_5$ in each equation to get polynomials $h_1, \ldots, h_5$ in $Y$.

3. The saturation ideal $J_Y = \langle h_1, \ldots, h_5 \rangle : \langle d_1 d_2 \cdots d_5 \rangle^\infty$ represents the preimage of (4.4).

Given such a wealth of parametrizations, we seek one where $J_Y$ has desirable properties. We use the following criterion: consider subsets of five of the generators of $J_Y$, compute the *mixed volume* of their Newton polytopes, and fix a subset minimizing that mixed volume.

| *Mixed Volume* | 5 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 20 | 23 | 24 | 25 | 30 | 35 | 42 | 45 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 2 | 416 | 6 | 73 | 50 | 167 | 563 | 751 | 10 | 12 | 6 | 1 | 11 | 12 | 4 | 4 | 4 |

Table 4.6: Reducing the steady state equations to the 2092 bases of base degree 1

In the census of 2092 bases in Table 4.6, that minimum is referred to as the mixed volume of $Y$.

By Bernstein's Theorem, the mixed volume is the number of solutions to a generic system with the five given Newton polytopes. We seek bases $Y$ where this matches the number nine from Theorem 4.0.3. We see that the mixed volume is nine for 416 of the bases in Table 4.6.

**Example 4.4.4.** *The basis $Y = \{x_1, x_4, x_6, x_8, x_{13}\}$ has base degree 1 and mixed volume 9. The remaining variables can be expressed in terms of $Y$ as follows. For brevity, we set*

$$\begin{aligned} r(x_4, x_6) &= k_9k_{11}k_{20}k_{22}x_4x_6 + k_9k_{11}(k_{21}+k_{22})(k_{23}+k_{31})x_4 \\ &+ k_{20}k_{22}(k_{10}+k_{11})(k_{13}+k_{30})x_6 + (k_{10}+k_{11})(k_{21}+k_{22})(k_{13}k_{23}+k_{23}k_{30}+k_{13}k_{31}). \end{aligned}$$

| | | | | | |
|---|---|---|---|---|---|
| $x_2$ | $=$ | $\frac{k_1}{k_2}x_1$ | $x_{12}$ | $=$ | $\frac{r(x_4,x_6)}{k_{12}k_{30}(k_{10}+k_{11})(k_{21}+k_{22})}\frac{k_{25}}{k_{24}}x_{13}$ |
| $x_3$ | $=$ | $\frac{k_1k_{26}}{k_2k_{27}}x_1$ | $x_{14}$ | $=$ | $\frac{k_1k_3}{k_2(k_4+k_5)}x_1x_4$ |
| $x_5$ | $=$ | $\frac{k_1k_3k_5(k_7+k_8)}{k_2k_6k_8(k_4+k_5)}\frac{x_1x_4}{x_8}$ | $x_{15}$ | $=$ | $\frac{k_1k_{14}k_{26}}{k_2k_{27}(k_{15}+k_{16})}x_1x_6$ |
| $x_7$ | $=$ | $\frac{k_1k_3k_5k_{28}(k_7+k_8)}{k_2k_6k_8k_{29}(k_4+k_5)}\frac{x_1x_4}{x_8}$ | $x_{16}$ | $=$ | $\frac{k_1k_3k_5}{k_2k_8(k_4+k_5)}x_1x_4$ |
| $x_9$ | $=$ | $\frac{k_6k_8k_{14}k_{16}k_{26}k_{29}(k_4+k_5)(k_{18}+k_{19})}{k_3k_4k_5k_{17}k_{19}k_{27}k_{28}(k_7+k_8)(k_{15}+k_{16})}\frac{x_6x_8}{x_4}$ | $x_{17}$ | $=$ | $\frac{k_1k_{14}k_{16}k_{26}}{k_2k_{19}k_{27}(k_{15}+k_{16})}x_1x_6$ |
| $x_{10}$ | $=$ | $\frac{k_{12}(k_{10}+k_{11})(k_{20}k_{22}x_6+(k_{21}+k_{22})(k_{23}+k_{31}))}{r(x_4,x_6)}$ | $x_{18}$ | $=$ | $\frac{k_9k_{12}(k_{20}k_{22}x_6+(k_{21}+k_{22})(k_{23}+k_{31}))}{r(x_4,x_6)}x_4$ |
| $x_{11}$ | $=$ | $\frac{k_{12}k_{30}(k_{10}+k_{11})(k_{21}+k_{22})}{r(x_4,x_6)}$ | $x_{19}$ | $=$ | $\frac{k_{12}k_{20}k_{30}(k_{10}+k_{11})}{r(x_4,x_6)}x_6$ |

*This map $\varphi_Y$ is substituted into (4.2), and then we saturate. The resulting ideal $J_Y$ equals*

$$\begin{aligned} \langle &\alpha_1x_6x_8 + \alpha_2x_4 + \alpha_3x_6, && \alpha_4x_1x_6 + \alpha_5x_1 + \alpha_6x_8 + \alpha_7, \\ &\alpha_8x_1x_4 + \alpha_9x_8 + \alpha_{10}, && \alpha_{11}x_4x_6x_{13} + \alpha_{12}x_4x_{13} + \alpha_{13}x_6x_{13} + \alpha_{14}x_{13} + \alpha_{15}, \\ &\alpha_{16}x_4x_6^2 + \alpha_{17}x_6^3 + \alpha_{18}x_4x_6 + && \alpha_{19}x_6^2 + \alpha_{20}x_8^2 + \alpha_{21}x_1 + \alpha_{22}x_4 + \alpha_{23}x_6 + \alpha_{24}x_8 + \alpha_{25}\rangle, \end{aligned}$$

*where the $\alpha_1, \ldots, \alpha_{25}$ are certain explicit rational functions in the* $\mathbf{k}$*-parameters.*

## 4.5 Polyhedral Geometry

Dynamics of the system while not at steady state cannot typically be studied with algebraic methods. One exception is the set of all possible states accessible from a given set of initial values via the chemical reactions in the model. This set is called a *stoichiometric compatibility class* in the biochemistry literature. Mathematically, these classes are convex polyhedra. We determine them all for the Wnt shuttle model. This resolves Problem 5 from the Introduction.

The conservation relations (4.2) define a linear map $\chi$ from the orthant of concentrations $\mathbb{R}^{19}_{\geq 0}$ to the orthant of conserved quantities $\mathbb{R}^{5}_{\geq 0}$. We express this projection as a $5\times 19$-matrix:

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{18} \\ x_{19} \end{pmatrix} \tag{4.7}$$

Let $P_{\mathbf{c}}$ denote the fiber of the map $\chi$ for $\mathbf{c} \in \mathbb{R}^5_{\geq 0}$. This is known in the biochemical literature as the *invariant polyhedron* or the *stoichiometric compatibility class* of the given $\mathbf{x}$; see e.g. [49, (3)]. The fiber over the origin is $P_{\mathbf{0}} = \mathbb{R}_{\geq 0}\{\mathbf{e}_{10}, \mathbf{e_{11}}\}$, the two-dimensional orthant formed by all positive linear combinations of $\mathbf{e}_{10}$ and $\mathbf{e}_{11}$. If $\mathbf{c} \in \mathbb{R}^5_{\geq 0}$ is an interior point, then $P_{\mathbf{c}}$ is a 14-dimensional convex polyhedron of the form $P_{\mathbf{0}} \times \tilde{P}_{\mathbf{c}}$ where $\tilde{P}_{\mathbf{c}}$ is a 12-dimensional (compact) polytope. Two vectors $\mathbf{c}$ and $\mathbf{c}'$ are considered *equivalent* if their invariant polyhedra $P_{\mathbf{c}}$ and $P_{\mathbf{c}'}$ have the same normal fan. This property is much stronger than being combinatorially isomorphic. The equivalence classes are relatively open polyhedral cones, and they define a partition of $\mathbb{R}^5_{\geq 0}$. This partition is the *chamber complex* of the matrix (4.7). For a low-dimensional illustration, see [49, Figure 1]. Informally speaking, the chamber complex classifies the possible boundary behaviors of our dynamical system.

**Proposition 4.5.1.** *The chamber complex of our* $5\times 19$*-matrix divides* $\mathbb{R}^5_{\geq 0}$ *into* 19 *maximal cones. It is the product of a ray,* $\mathbb{R}_{\geq 0}$*, and the cone over a subdivision of the tetrahedron. That subdivision consists of* 18 *smaller tetrahedra and* 1 *bipyramid, described in detail below.*

*Proof.* The product structure arises because the matrix has two blocks after permuting columns, an upper left $4\times 17$ block and a lower right $1\times 2$ block $(1\ 1)$. Our task is to compute the chamber decomposition of $\mathbb{R}^4_{\geq 0}$ defined by the $4 \times 17$-block. After deleting zero columns and multiple columns, we are left with a $4 \times 7$-matrix, given by the seven left columns in

$$M \quad = \quad \begin{array}{c} \begin{array}{cccccccccccc} a & b & c & d & e & f & g & h & i & j & k & l \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}.$$

The correspondence between the seven left columns of $M$ and the columns of (4.7) is as follows:

$$a = \{x_4, x_5, x_6, x_7, x_{18}, x_{19}\}, \quad b = \{x_{14}, x_{15}\}, \quad c = \{x_{16}\},$$
$$d = \{x_{17}\}, \quad e = \{x_1, x_2, x_3\}, \quad f = \{x_8\}, \quad g = \{x_9\}.$$

The remaining columns of $M$ are additional vertices in the subdivision.

The following table lists the 19 maximal chambers. For each chamber we list the extreme rays and the facet-defining inequalities. For instance, the chamber in $\mathbb{R}^5_{\geq 0}$ denoted by $efjk$ is the orthant spanned by the columns $e$, $f$, $j$ and $k$ of the matrix $M$ times the ray $(0,0,0,0,1)^T$. It is defined by $c_5 \geq 0$ together with the four listed inequalities: $c_4 \geq 0$, $\min(c_1, c_3) \geq c_2 \geq c_4$.

| | |
|---|---|
| $abcd$ | $\{c_4, c_3, c_1, c_2 - c_4 - c_3 - c_1\}$ |
| $bcdl$ | $\{c_2 - c_3 - c_1, c_2 - c_4 - c_1, c_2 - c_4 - c_3, -c_2 + c_4 + c_3 + c_1\}$ |
| $efgk$ | $\{c_2, -c_2 + c_4, -c_2 + c_3, -c_2 + c_1\}$ |
| $bcjl$ | $\{c_4, -c_2 + c_1 + c_3, c_2 - c_3 - c_4, c_2 - c_1 - c_4\}$ |
| $bdil$ | $\{c_3, -c_2 + c_4 + c_1, c_2 - c_4 - c_3, c_2 - c_3 - c_1\}$ |
| $beij$ | $\{c_3, c_4, c_1 - c_2, c_2 - c_3 - c_4\}$ |
| $cdhl$ | $\{c_1, -c_2 + c_3 + c_4, c_2 - c_4 - c_3, c_2 - c_4 - c_1\}$ |
| $cfhj$ | $\{c_4, c_1, -c_2 + c_3, c_2 - c_4 - c_1\}$ |
| $dghi$ | $\{c_1, c_3, -c_2 + c_4, c_2 - c_1 - c_3\}$ |
| $egik$ | $\{c_3, -c_2 + c_4, c_2 - c_3, -c_2 + c_1\}$ |
| $fghk$ | $\{c_1, -c_1 + c_2, -c_2 + c_3, -c_2 + c_4\}$ |
| $efjk$ | $\{c_4, c_1 - c_2, c_2 - c_4, -c_2 + c_3\}$ |
| $bijl$ | $\{c_2 - c_1, -c_2 + c_1 + c_3, -c_2 + c_4 + c_1, c_2 - c_3 - c_4\}$ |
| $chjl$ | $\{c_2 - c_3, -c_2 + c_4 + c_3, -c_2 + c_3 + c_1, c_2 - c_4 - c_1\}$ |
| $dhil$ | $\{c_2 - c_4, -c_2 + c_4 + c_1, -c_2 + c_3 + c_4, c_2 - c_1 - c_3\}$ |
| $ghik$ | $\{c_4 - c_2, c_2 - c_3, c_2 - c_1, -c_2 + c_1 + c_3\}$ |
| $eijk$ | $\{c_2 - c_4, c_2 - c_3, c_1 - c_2, -c_2 + c_3 + c_4\}$ |
| $fhjk$ | $\{c_2 - c_4, c_2 - c_1, -c_2 + c_3, -c_2 + c_4 + c_1\}$ |
| $hijkl$ | $\{c_2 - c_4, c_2 - c_3, c_2 - c_1, -c_2 + c_4 + c_3, -c_2 + c_4 + c_1, -c_2 + c_3 + c_1\}$ |

Interpreting the columns of $M$ as homogeneous coordinates, the table describes a subdivision of the standard tetrahedron into 18 tetrahedra and one bipyramid $hijkl$. These cells use the 12 vertices $a, b, \ldots, l$. The reader is invited to check that this subdivision has precisely 39 edges and 47 triangles, so the Euler characteristic is correct: $12 - 39 + 47 - 19 = 1$. □

We shall prove the following result about the Wnt shuttle model.

**Proposition 4.5.2.** *Suppose that the rate constants $k_i$ and the conserved quantities $c_j$ are all strictly positive. Then no steady states exist on the boundary of the invariant polyhedron $P_{\mathbf{c}}$.*

*Proof.* Consider the two components $I_m$ and $I_e$ of the steady state ideal $I$ given in Lemma 4.2.1. We intersect each of the two varieties with the affine-linear space defined by the

conservation relations (4.2) for some $\mathbf{c} \in \mathbb{R}^5_{>0}$. We claim that all solutions $\mathbf{x}$ satisfy $x_i \neq 0$ for $i = 1, 2, \ldots, 19$.

For the main component $V(I_m)$, we prove this assertion with the help of the parametrization $\varphi_Y$ from Example 4.4.4. If the values of $x_1, x_4, x_6, x_8, x_{13}$ and of the expression $r(x_4, x_6)$ are nonzero, then each coordinate of $\varphi_Y$ is nonzero. We next observe that $r(x_4, x_6) > 0$ for any $\mathbf{k} > 0$ and $\mathbf{x} \geq 0$. A case analysis, using binomial relations in the ideal $I_m$, reveals that if any of $x_1, x_4, x_6, x_8, x_{13}$ are zero, some coordinate of $\mathbf{c}$ is forced to zero as well:

$$\begin{array}{rcrcr}
x_1 = 0 & \Rightarrow & x_2, x_3, x_{14}, x_{15} = 0 & \Rightarrow & c_1 = 0, \\
x_{13} = 0 & \Rightarrow & x_{12} = 0 & \Rightarrow & c_5 = 0, \\
x_4 = 0 & \Rightarrow & x_5, x_6, x_7, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19} = 0 & \Rightarrow & c_2 = 0, \\
 & \text{or} & x_8, x_{16} = 0 & \Rightarrow & c_3 = 0, \\
x_6 = 0 & \Rightarrow & x_9, x_{17} = 0 & \Rightarrow & c_4 = 0, \\
 & \text{or} & x_4 = 0 & \Rightarrow & c_2 \text{ or } c_3 = 0, \\
x_8 = 0 & \Rightarrow & x_{16} = 0 & \Rightarrow & c_3 = 0.
\end{array}$$

It remains to consider the extinction component. Its ideal $I_e$ contains the set $b \cup l = \{x_1, x_2, x_3, x_{14}, x_{15}\}$. The corresponding columns of the matrix in (4.7) are the only columns with a nonzero entry in the fourth row. This implies that $c_4 = 0$ holds for every steady state in $V(I_e)$. We conclude that there are no steady states on the boundary of the polyhedron $P_{\mathbf{c}}$. $\square$

**Remark 4.5.3.** *In this proof we did not need the detailed description of the chamber complex, because of the special combinatorial structure in the Wnt shuttle model. In general, when studying chemical reaction networks that arise in systems biology, an analysis like Proposition 4.5.1 is requisite for gaining information about possible zero coordinates in the steady states.*

## 4.6 Parameter Estimation

Question 6 asks: *What information does species concentration data give us for parameter estimation?* This question is of particular importance to experimentalists, as species concentrations depend on initial conditions, whereas parameter values are intrinsic to the biological process being modeled. Identifiability of parameters has been studied in many contexts, notably in statistics [16] and in biological modeling [39]. Sometimes, as in [39], parameters are determined from complete time-course data of the dynamical system, making a differential algebra approach desirable. In the present treatment, we focus on the steady state variety, so we consider data collection only at steady state. We assume that there is a true but unknown parameter vector $\mathbf{k}^* \in \mathbb{R}^{31}$ of rate constants, and our data are sampled from the positive real points $\mathbf{x}$ on the variety in $\mathbb{R}^{19}$ that is defined by the 19 polynomials in (4.1).

### Complete Species Information.

The first algebraic question we answer: To what extent is the true parameter vector $\mathbf{k}^*$ determined by points on its steady state variety?

To address this question, we form the polynomial matrix $F(\mathbf{x})$ of format $19 \times 31$ whose entries are the coefficients of the right-hand sides of (4.1), regarded as linear forms in $\mathbf{k}$. With this notation, our dynamical system (4.1) can be written in matrix-vector product form as

$$\dot{\mathbf{x}} \;=\; F(\mathbf{x}) \cdot \mathbf{k}.$$

Our data points are sampled from

$$\big\{\, \mathbf{x} \in \mathbb{R}^{19}_{>0} \;:\; F(\mathbf{x}) \cdot \mathbf{k}^* = \mathbf{0} \,\big\}. \tag{4.8}$$

Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ denote generic data points in (4.8). The set of all parameter vectors $\mathbf{k}$ that are compatible with these data is a linear subspace of $\mathbb{R}^{31}$, namely it is the intersection

$$\text{kernel}(F(\mathbf{x}_1)) \,\cap\, \text{kernel}(F(\mathbf{x}_2)) \,\cap\, \text{kernel}(F(\mathbf{x}_3)) \,\cap\, \cdots \tag{4.9}$$

The best we can hope to recover from sampling data is the following subspace containing $\mathbf{k}^*$:

$$\bigcap_{\mathbf{x}\,\text{in}\,(4.8)} \text{kernel}(F(\mathbf{x})) \;\;\subset\;\; \mathbb{R}^{31}. \tag{4.10}$$

We refer to (4.10) as the *space of parameters compatible with* $\mathbf{k}^*$. A direct computation reveals:

**Proposition 4.6.1.** *The space of all parameters compatible with* $\mathbf{k}^*$ *is a* 14*-dimensional subspace of* $\mathbb{R}^{31}$*. If* $\mathbf{x}$ *is generic then the kernel of* $F(\mathbf{x})$ *is a* 17*-dimensional subspace of* $\mathbb{R}^{31}$*.*

This has the following noteworthy consequence for our biological application:

**Corollary 4.6.2.** *The parameters of the Wnt shuttle model are not identifiable from steady state data, but there are* 14 *degrees of freedom in recovering the true parameter vector* $\mathbf{k}^*$*.*

Our next step is to gain a more precise understanding of the subspaces in Proposition 4.6.1. To do this, we shall return to the combinatorial setting of matroid theory. We introduce two matroids on the 31 reactions in Table 4.2. The common ground set is $K = \{k_1, k_2, \ldots, k_{31}\}$. The *one-point matroid* $\mathcal{M}_{\text{one}}$ is the rank 17 matroid on $K$ defined by the linear subspace $\text{kernel}(F(\mathbf{x}))$ of $\mathbb{R}^{31}$ where $\mathbf{x} \in \mathbb{R}^{19}$ is generic. The *parameter matroid* $\mathcal{M}_{\text{par}}$ is the rank 14 matroid on $K$ defined by the space (4.10) of all parameters compatible with a generic $\mathbf{k}^*$. The following result, obtained by calculations, reflects the block structure of the matrix $F(\mathbf{x})$.

**Proposition 4.6.3.** *The one-point matroid* $\mathcal{M}_{\text{one}}$ *is the graphic matroid of the graph shown in Figure 4.1 a). Its seven connected components are matroids of ranks* $3, 3, 7, 1, 1, 1, 1$*. The rank* 14 *parameter matroid* $\mathcal{M}_{\text{par}}$ *is obtained from* $\mathcal{M}_{\text{one}}$ *by specializing the rank* 7 *component to the rank* 4 *matroid on* 11 *elements whose affine representation is shown in Figure 4.1 b).*

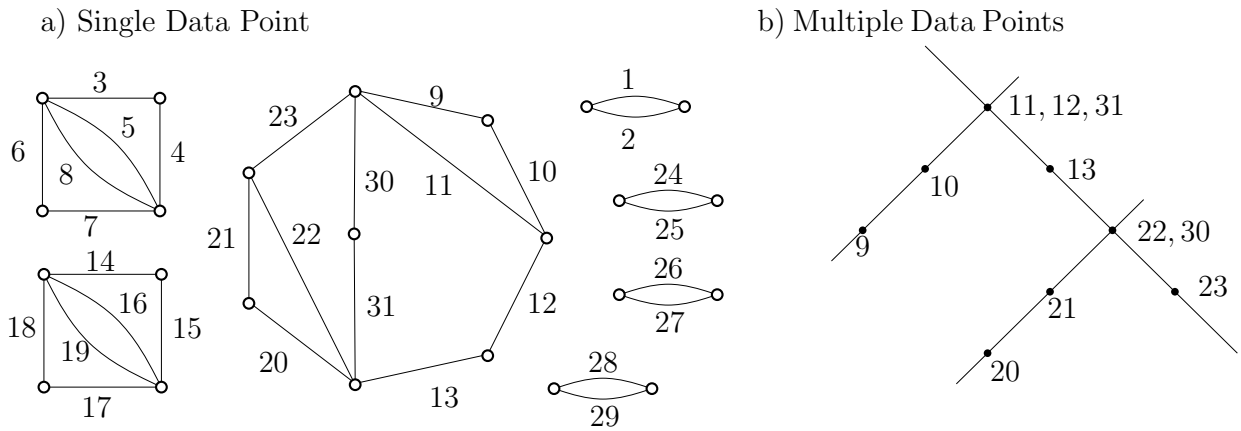



Figure 4.1: Graphic rep. of $\mathcal{M}_{\text{one}}$, and affine rep. of the rank 4 component of $\mathcal{M}_{\text{par}}$.

This characterizes the combinatorial constraints imposed on the parameters $\mathbf{k}$ by measuring the species concentrations at steady state. For a single measurement $\mathbf{x}$, the result on $\mathcal{M}_{\text{one}}$ tells us that the $19 \times 31$-matrix $F(\mathbf{x})$ has rank $14 = 31 - \text{rank}(\mathcal{M}_{\text{one}})$. After row operations, it block-decomposes into two matrices of format $3 \times 6$, one matrix of format $4 \times 11$, and four matrices of format $1 \times 2$. Each of these seven matrices is row-equivalent to the node-edge *cycle matrix* of a directed graph, with underlying undirected graph as in Figure 4.1 (a).

Consider the graph with edges $9, 10, 11, 12, 13, 20, 21, 22, 23, 30, 31$. The cycle $\{22, 23, 30, 31\}$ reveals that our measurement $\mathbf{x}$ imposes <u>one</u> linear constraint on $k_{22}, k_{23}, k_{30}, k_{31}$. If we take further measurements, as in (4.9), then six of the seven blocks of $F(\mathbf{x})$ remain unchanged. Only the $4 \times 11$-block of $F(\mathbf{x})$ must be enlarged, to a $7 \times 11$-matrix. The rows of that new matrix specify the affine-linear dependencies among 11 points in $\mathbb{R}^3$. That point configuration is depicted in Figure 4.1 (b). For instance, the points $\{9, 10, 11\}$ are collinear, the points $\{20, 21, 22\}$ are collinear, but these two lines are skew in $\mathbb{R}^3$. From the other line we see that that repeated measurements at steady state impose <u>two</u> linear constraints on $k_{22}, k_{23}, k_{30}, k_{31}$.

## Circuit Data.

The second question we address in this section: *Given partial species concentration data, is any information about parameters available?* In Section 7.1, all 19 concentrations $x_i$ were available for a steady state. In what follows, we suppose that $x_i$ can only be measured for indices $i$ in a subset of the species, say $C \subset \{1, \ldots, 19\}$. In our analysis, it will be useful to take advantage of the rank 5 algebraic matroid in Proposition 4.4.1, since that matroid governs dependencies among the coordinates $x_1, \ldots, x_{19}$ at steady states.

We here focus on the special case when $C$ is one of the 951 circuits of the algebraic

matroid of $\widetilde{I_m}$. Let $f_C$ be the corresponding circuit polynomial, as in Table 4.5. We regard $f_C$ as a polynomial in $\mathbf{x}$ whose coefficients are polynomials in $\mathbb{Q}[\mathbf{k}]$. Suppose that $f_C$ has $r$ monomials $\mathbf{x}^{a_1}, \ldots, \mathbf{x}^{a_r}$. We write $F_C \in \mathbb{Q}[\mathbf{k}]^r$ for the vector of coefficients, so our circuit polynomial is the dot product $f_C(\mathbf{k}, \mathbf{x}) = F_C(\mathbf{k}) \cdot (\mathbf{x}^{a_1}, \ldots, \mathbf{x}^{a_r})$. We write $\mathcal{V}_C \subset \mathbb{R}^r$ for the algebraic variety parametrized by $F_C(\mathbf{k})$. Thus $\mathcal{V}_C$ is the Zariski closure in $\mathbb{R}^r$ of the set $\{F_C(\mathbf{k}') : \mathbf{k}' \in \mathbb{R}^{31}\}$.

Our idea for parameter recovery is this: rather than looking for $\mathbf{k}$ compatible with the true parameter $\mathbf{k}^*$, we seek a point $\mathbf{y} = F_C(\mathbf{k})$ in $\mathcal{V}_C$ that is compatible with $F_C(\mathbf{k}^*)$. And, only later do we compute a preimage of $\mathbf{y}$ under the map $\mathbb{R}^{31} \to \mathbb{R}^r$ given by $F_C$. Most interesting is the case when $\mathcal{V}_C$ is a proper subvariety of $\mathbb{R}^r$. Direct computations yield the following:

**Proposition 4.6.4.** *For precisely* 288 *of the* 951 *circuits $C$ of the algebraic matroid of the steady state ideal $\widetilde{I_m}$, the coefficient variety $\mathcal{V}_C$ is a proper subvariety in its ambient space $\mathbb{R}^r$. In each of these cases, the defining ideal of $\mathcal{V}_C$ is of one of the following four types:*

$$\langle y_2y_6 - y_3y_5 \rangle \tag{4.11}$$

$$\langle y_5y_6 - 2y_3y_7, y_5^2 - 4y_2y_7, y_3y_5 - 2y_2y_6, y_2y_6^2 - y_3^2y_7 \rangle \tag{4.12}$$

$$\langle y_3y_5^2 - y_2y_5y_6 + y_1y_6^2 \rangle \tag{4.13}$$

$$\langle 2y_3y_4 - y_2y_5, y_2y_3 - 2y_1y_5, y_2^2 - 4y_1y_4 \rangle \tag{4.14}$$

**Example 4.6.5.** *Consider the circuit $C = \{6, 10, 18\}$. The circuit polynomial $f_C$ equals*

$$\begin{gathered}(k_{13}k_{20}k_{22} + k_{20}k_{22}k_{30}) \cdot x_6x_{10} \;+\; k_{11}k_{20}k_{22} \cdot x_6x_{18} \;-\; k_{12}k_{20}k_{22} \cdot x_6 \\ +(k_{13}k_{21}k_{23} + k_{13}k_{22}k_{23} + k_{21}k_{23}k_{30} + k_{22}k_{23}k_{30} + k_{13}k_{21}k_{31} + k_{13}k_{22}k_{31}) \cdot x_{10} \\ +(k_{11}k_{21}k_{23} + k_{11}k_{22}k_{23} + k_{11}k_{21}k_{31} + k_{11}k_{22}k_{31}) \cdot x_{18} \\ -\,(k_{12}k_{21}k_{23} + k_{12}k_{22}k_{23} + k_{12}k_{21}k_{31} + k_{12}k_{22}k_{31}).\end{gathered}$$

*Here $r = 6$ and we write $F_C(\mathbf{k}) = (y_1, y_2, y_3, y_4, y_5, y_6)$ for the vector of coefficient polynomials. The variety $\mathcal{V}_C$ is the hypersurface in $\mathbb{R}^5$ defined by the equation $y_2y_6 = y_3y_5$.*

We now sample data points $\mathbf{x}_i$ from the model with the true (but unknown) parameter vector $\mathbf{k}^*$. Each such point defines a hyperplane $\{\mathbf{y} \in \mathbb{R}^r : \mathbf{y} \cdot (\mathbf{x}_1^{a_1}, \ldots, \mathbf{x}_r^{a_r}) = 0\}$. The parameter estimation problem is to find the intersection of these data hyperplanes with the variety $\mathcal{V}_C$. That intersection contains the point $\mathbf{y}^* = F_C(\mathbf{k}^*)$, which is what we now aim to recover.

## Noisy Circuit Data.

The final question we consider in this section is: *Given partial species concentration data with noise, is any information about parameters available?*

As in Section 7.2, we fix a circuit $C$ of the algebraic matroid in Section 4.4, and we assume that we can only measure the concentrations $x_j$ where $j \in C$. Each measurement

$\mathbf{x}_i \in \mathbb{R}^C$ still defines a hyperplane $\mathbf{y} \cdot (\mathbf{x}_i^{a_1}, \ldots, \mathbf{x}_i^{a_r}) = 0$ in the space $\mathbb{R}^r$. But now the true vector $\mathbf{y}^* = F_C(\mathbf{k}^*)$ is not exactly on that hyperplane, but only close to it. Hence, if we take $s$ repeated measurements, with $s > r$, the intersection of these hyperplanes should be empty.

We propose to find the best fit by solving the following least squares optimization problem:

$$\text{Minimize} \quad \sum_{i=1}^{s} \big( \mathbf{y} \cdot (\mathbf{x}_i^{a_1}, \ldots, \mathbf{x}_i^{a_r}) \big)^2 \quad \text{subject to } \mathbf{y} \in \mathcal{V}_C \cap \mathbb{S}^{r-1}, \tag{4.15}$$

where $\mathbb{S}^{r-1} = \{\mathbf{y} \in \mathbb{R}^r : y_1^2 + y_2^2 + \cdots + y_r^2 = 1\}$ denotes the unit sphere. When the variety $\mathcal{V}_C$ is the full ambient space $\mathbb{R}^r$, this is a familiar regression problem, namely, to find the hyperplane through the origin that best approximates $s$ given points in $\mathbb{R}^r$. Here "best" means that the sum of the squared distances of the $s$ points to the hyperplane is minimized. This happens for 663 of the 951 circuits $C$, and in that case we can apply standard techniques.

However, for the 288 circuits $C$ identified in Proposition 4.6.4, the problem is more interesting. Here the hyperplanes under consideration are constrained to live in a proper subvariety. In that case we need some algebraic geometry to reliably find the global optimum in (4.15).

Our problem is to minimize a quadratic function over the real affine variety $\mathcal{V}_C \cap \mathbb{S}^{r-1}$. The quadratic objective function is generic because the $\mathbf{x}_i$ are sampled with noise. The intrinsic algebraic complexity of our optimization problem was studied by Draisma et al. in [10]. That complexity measure is the *ED degree* of $\mathcal{V}_C \cap \mathbb{S}^{r-1}$, which is the number of solutions in $\mathbb{C}^r$ to the critical equations of (4.15). Here, by ED degree we mean the ED degree of $\mathcal{V}_C \cap \mathbb{S}^{r-1}$, when considered in generic coordinates. This was called the *generic ED degree* in [44].

We illustrate our algebraic approach by working out the first instance (4.11) in Proposition 4.6.4.

**Example 4.6.6.** *Suppose we are given $s$ noisy measurements of the concentrations $x_6, x_{10}, x_{18}$. In order to find the best fit for the parameters $\mathbf{k}$, we employ the circuit polynomial $f_C$ in Example 4.6.5. We compute $\mathbf{y} \in \mathbb{R}^6$ by solving the corresponding optimization problem (4.16). This problem is to minimize a random quadratic form subject to two quadratic constraints*

$$y_2 y_6 - y_3 y_5 \;=\; y_1^2 + y_2^2 + y_3^2 + y_4^2 + y_5^2 + y_6^2 - 1 \;=\; 0. \tag{4.16}$$

*We solve this problem using the method of Lagrange multipliers. This leads to a system of polynomial equations in $\mathbf{y}$. Using saturation, we remove the singular locus of (4.16), which is the circle $\{\mathbf{y} \in \mathbb{R}^6 : y_1^2 + y_4^2 - 1 = y_2 = y_3 = y_5 = y_6 = 0\}$. The resulting ideal has precisely 40 zeros in $\mathbb{C}^6$. In the language of [10, 44], the generic ED degree of the variety (4.16) equals 40.*

## 4.7 From Algebra to Biology

The aims of this chapter have been: (1) to demonstrate how biology can lead to interesting questions in algebraic geometry, and (2) to apply new techniques from computational algebra in biology. So far, our tour through (numerical) algebraic geometry, polyhedral geometry and combinatorics has demonstrated the range of mathematical questions to explore. In this section, we will focus on translating our analysis into applicable considerations for the research cycle in systems biology, which is illustrated in Figure 4.7. In what follows we discuss some concrete applications and results pertaining to the steps (a), (b) and (c) in Figure 4.7.

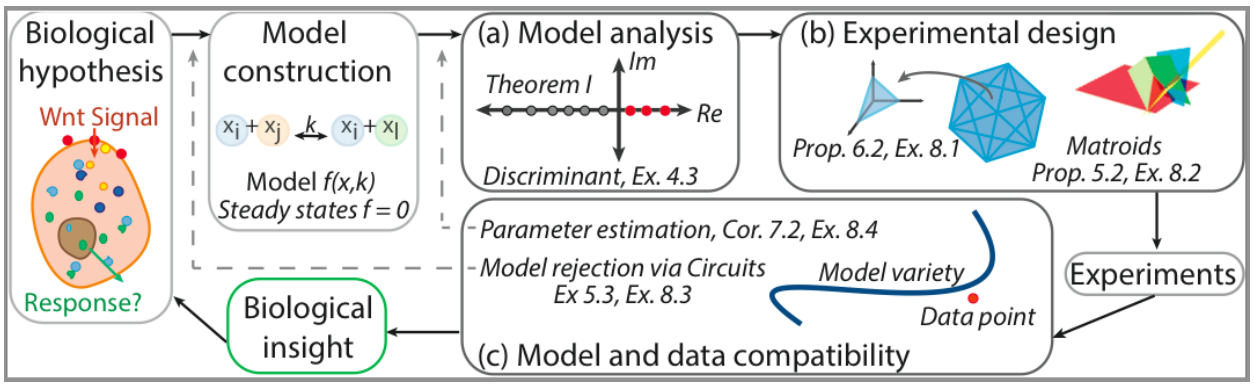



Figure 4.2: Schematic diagram for systems biology research.

**Analysis of the Model**: Before any experiments are performed, our techniques inform the modeler of the global steady-state properties of the model. The number of real solutions to system (4.1)–(4.2), stated in Theorem 4.0.3, governs the number of observable steady states. Various sampling schemes demonstrated that *most* parameter values lead to only one observable steady state. We produced a set of parameter values and conserved quantities with three real solutions, and two solutions are also attainable. If the "true" parameters $\mathbf{k}^*$ and $\mathbf{c}^*$ admit multiple real solutions, then multistationarity of the system is theoretically possible.

If multiple states are observed experimentally, then the model must be capable of multistationarity. In the Wnt shuttle model, the system is capable of multiple steady states; however, based on parameter sampling, the frequency of this occurrence is low, and parameters in this regime are somewhat stable under perturbation. The discriminant of the system is a polynomial of degree 34 in $\mathbf{c}$, and our analysis along a single line in $\mathbf{c}$-space illustrates the high degree of complexity inherent in the full stratification of the 36-dimensional parameter space.

**Experimental Design**: In Section 4.5, the combinatorial structure of the various stoichiometric compatibility classes was fully characterized. As the conserved quantities $\mathbf{c} = (c_1, \ldots, c_5)$ range over all positive real values, the set of all compatible species-concentration

vectors $\mathbf{x}$ will take one of 19 polyhedral shapes $P_{\mathbf{c}}$. This may find application in identifying multiple steady state solutions for specific rate constants $\mathbf{k}$. A natural choice for initial conditions when performing experiments is on or near the vertices of the 14-dimensional polyhedron $P_{\mathbf{c}}$.

**Example 4.7.1.** *Suppose the conserved quantities vector lies in the bipyramid, e.g.* $\mathbf{c} = (1, 2, 2, 2, 3)$. *The preimage of* $\mathbf{c}$ *in* $\mathbf{x}$*-space is a product of the orthant* $\mathbb{R}_{\geq 0}\{\mathbf{e}_{10}, \mathbf{e}_{11}\}$ *and a 12-dimensional polytope with* 400 *vertices:* $(1, 0, 0, 2, 0, 0, 0, 2, 2, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0)$, *and* 399 *of its permutations. This product is the polyhedron* $P_{\mathbf{c}}$. *If we have control over initial conditions, beginning near the vertices positions us to find interesting systems behavior.*

In the laboratory, the experimentalist makes choices of what to measure and what not to measure. For instance, measuring a particular $x_i$ may be infeasible, or there may be a situation in which measuring concentration $x_i$ can preclude measuring concentration $x_j$.

For every strategy, we fix a *cost vector*, listing the costs of making each measurement. We use the symbol $N$ to indicate infeasible measurements. Suppose there are two different ways to run the experiment; then we have a $2 \times 19$ *cost matrix* $P$, whose rows are cost vectors for each experiment. We multiply $P$ by the 0-1-incidence matrix for the 951 circuits of Proposition 4.4.1. That matrix has a 1 in row $i$ and column $j$ if circuit $j$ contains species $i$, and 0 otherwise. The product is a matrix of size $2 \times 951$. For $N \to \infty$, the $2 \times 951$ matrix has a finite entry in position $(i, j)$ precisely when the strategy $i$ can measure the circuit $j$. Minimizing over those finite cost entries selects the most cost-effective experiment to measure a circuit.

**Example 4.7.2.** *Suppose that none of the intermediate complexes* $x_{13}, \ldots, x_{19}$ *are measurable, and that we are able to measure only one Phosphatase concentration (*$x_4$ *or* $x_8$*) in each experimental setup. A corresponding cost matrix might look like*

$$P = \begin{bmatrix} 1 & 1 & 1 & N & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & N & N & N & N & N & N & N \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & N & 1 & 1 & 1 & 1 & N & N & N & N & N & N & N \end{bmatrix}$$

*Multiplying by the circuit support matrix of size* $19 \times 951$ *reveals* 82 *feasible experiments:* 50 *using the first row of* $P$, *and* 32 *using the second. With more refined cost assignment, this would decide not only feasibility but also optimal cost. In this way, the matroid allows us to choose cost-minimal experiments to obtain meaningful information for the model.*

**Model and data compatibility**: After an experiment is performed, the task of the modeler is to test the data with the model. One possible outcome is *model rejection.* If the data are compatible, then another outcome is *parameter estimation.* Both may provide insights for biology. The role of algebraic geometry is seen in [19, 21] and shown in the next two examples.

**Example 4.7.3** (Model Rejection)**.** *Suppose that rate parameters* $k_i$ *are all known to be* 1, *and that we have collected data for variables* $x_1, x_4, x_{14}$. *The circuit polynomial is* $k_1 k_3 x_1 x_4 +$

*$(-k_2k_4 - k_2k_5)x_{14}$, which specializes to $x_1x_4 - 2x_{14}$. If the evaluation of the positive quantity $|x_1x_4 - 2x_{14}|$ lies above a threshold $\epsilon$, then we can reject the model as not matching the data.*

Every circuit polynomial of the matroid is a *steady state invariant*; depending on which experiment was performed, the collection of measured variables must contain some circuit. Even if one can measure all 19 species at steady state, it is not possible to recover all 31 kinetic rate constants, but we do have relationships that must be satisfied among parameters [36].

**Example 4.7.4** (Parameter Estimation)**.** *Suppose that rate parameters are unknown, and that we have collected data for $x_6, x_{10}, x_{18}$. The corresponding circuit polynomial $f_C$ is shown in Example 4.6.5. We know that the coefficients of $f_C$ satisfy the constraint $y_2y_6 = y_3y_5$. Suppose our experiments lead to the following ten measurements for the vector $(x_6, x_{10}, x_{18})$:*

$$\{(.715335, 4.06778, 14.6806), (.390982, 4.83152, 6.08251), (.706539, 4.98107, 3.83617),$$
$$(.14316, 4.30851, 12.5809), (.995583, 4.01222, 15), (.413817, 4.08114, 14.902), (.232206, 3.38274, 23.3162),$$
$$(.219045, 5.06008, 3.67175), (.704106, 3.52804, 21.1037), (.648732, 3.6505, 19.7008)\}$$

*The data lead us to the following function to optimize in (4.15):*

$$\begin{gathered}57.2345y_1^2 + 376.181y_1y_2 + 801.672y_2^2 - 27.5625y_1y_3 - 96.4429y_2y_3\\ +3.36521y_3^2 + 179.49y_1y_4 + 564.034y_2y_4 - 42.729y_3y_4 + 178.839y_4^2 + 564.034y_1y_5\\ +2424.31y_2y_5 - 144.7y_3y_5 + 1054.49y_4y_5 + 2263.2y_5^2 - 42.729y_1y_6\\ -144.7y_2y_6 + 10.339y_3y_6 - 83.8072y_4y_6 - 269.749y_5y_6 + 10y_6^2\end{gathered}$$

The global minimum of this quadratic form on the codimension 2 variety (4.16) has coordinates

$$y_1 = 0.183472,\ y_2 = 0.152416,\ y_3 = 0.959232,\ y_4 = 0.038042,\ y_5 = 0.00335267,\ y_6 = 0.211.$$

Given these values, one now has three degrees of freedom in estimating the nine parameters $k_i$ that appear in the circuit polynomial $f_C$. The other ten coordinates of $\mathbf{k}$ are unspecified.

The main agenda of this chapter was to show the range of algebraic tools that can be used to analyze chemical reaction networks. In the process, algebraic matroids showed themselves to be a vital part of the applied algebraic geometry toolkit.

# Chapter 5

# Further Directions

This thesis has aimed to provide the reader with background and tools to compute algebraic matroids for varieties arising in nature. The explorations in Chapters 3 and 4 then gave particular examples where these tools were put to effective use. At the conclusion of this dissertation, a plethora of questions remain, both about sharpening our tools and our understanding, as well as new areas where the computation of algebraic matroids may be useful. We will compile all open questions from throughout the text, along with other directions not yet mentioned, that we hope to explore in future work.

I. **Base Degrees and Circuit Polynomials**.

1. Given a prime ideal of a certain form, for example, ideals generated by $k$ forms of degree $d$ and variable support of size $r$, what is the expected distribution of base degrees and circuit degrees?
2. What constraints are there on this collection of degrees? (Page 8)
3. Specifically, what is the full set of Chow polytopes allowed given a particular matroid?
4. Are there "local" properties of the matroid polytope that we can use to find bases of high degree without computing the full list of base degrees? Since high-degree bases have demonstrated nice symmetry (e.g. in the Plackett-Luce matroid, and the matroid of $Gr(3,6)$), this may give important insights.
5. Describe the prime ideals for which the set of circuits are a minimal generating set for the ideal. When does a proper subset of circuits generate the ideal?
6. Fix some prime ideal $P$. Suppose the set of circuit polynomials do not generate $P$. How many components will be in the ideal they do generate?

II. **Algebraic Representability**:

1. Is the class of algebraic matroids closed under taking duals? This problem is a long-standing one, but some progress may be possible with the help of computations.
2. In particular, we might answer the simpler question: Is the Tic-Tac-Toe matroid algebraic? The dual of this matroid is known to be non-algebraic [22], so sampling from the right space and computing the algebraic matroid may turn up a counterexample.
3. Suppose $\mathcal{M}$ is a certain algebraic, non-linear matroid. What is the minimal integer $d$ such that an algebraic representation exists for $\mathcal{M}$ with all circuit polynomials having degree at most $d$? What is the smallest field that can serve as the ground field for this representation?

III. **Geometry**.

1. What hyperplane sections induce a truncation of the algebraic matroid? Intuition dictates that almost all hyperplanes would lead to a truncation; characterizing the subvariety of the projective space of hyperplanes that would induce a different matroid seems very interesting but nontrivial. (Page 46)
2. The Grassmannian is a geometric object whose Plücker ideal has fascinating combinatorics. We saw that the high-degree base of $\mathcal{M}(Gr(3,6))$ was highly symmetric. Why does this symmetry translate into a higher degree fiber? Are there any other Grassmannians with unusually high-degree projections? (Page 22)
3. The non-matroidal locus is the set of points where the coordinate projections line up with the tangent space of the variety. For a variety of dimension $r$ and degree $d$, what number of components of what degree does the non-matroidal locus have?

IV. **Statistics**.

1. Characterize the matroid for the Plackett-Luce model $PL_n$ for all $n$, as we did for $n = 4$. (Page 15)
2. Characterize the matroid of the variety of rank 2 matrices with entries summing to 1, as we did for rank 1 matrices. (Page 47)
3. Characterize the matroid of the variety of rank 1 tensors with entries summing to 1, as we did for rank 1 matrices. (Page 49)
4. Gaussoids were introduced in [34] to study Gaussian graphical models. The graphs themselves give rise to a graphic matroid. Finally, these models also give rise to ideals relating the various correlations, which in turn produce algebraic matroids. What is the relationship among the gaussoid, the graphical matroid and the algebraic matroid of a graphical model?

V. **Biology**.

Two distinct matroids arise in the definition of a chemical reaction network: first, the directed graph of the network defines a graphical matroid. second, the ordinary differential equations define an algebraic matroid. What properties do these two matroids share?

# Bibliography


[1] David Avis and Komei Fukuda. "Reverse Search for Enumeration". In: *Discrete Applied Mathematics* 65 (1993), pp. 21–46.

[2] Daniel J. Bates, Elizabeth Gross, Anton Leykin, and Jose Rodriguez. "Bertini for Macaulay2". arXiv:1310.3297. 2014.

[3] Daniel J. Bates, Jonathan D. Hauenstein, Andrew J. Sommese, and Charles W. Wampler. *Bertini: Software for Numerical Algebraic Geometry*. Available at bertini.nd.edu with permanent doi: dx.doi.org/10.7274/R0H41PB5.

[4] Daniel J. Bates, Andrew J. Sommese, Jonathan D. Hauenstein, and Charles W. Wampler. *Numerically Solving Polynomial Systems with Bertini*. Vol. 25. Software, environments, tools. SIAM, 2013.

[5] Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White, and Günter M. Ziegler. *Oriented matroids*. Second. Vol. 46. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press, 1999.

[6] Endre Boros, Khaled Elbassioni, Vladimir Gurvich, and Leonid Khachiyan. "Algorithms for enumerating circuits in matroids". In: *Algorithms and computation*. Vol. 2906. Lecture Notes in Comput. Sci. Berlin: Springer, 2003, pp. 485–494.

[7] Changbo Chen, James H. Davenport, Marc Moreno Maza, Bican Xia, and Rong Xiao. "Computing with semi-algebraic sets: Relaxation techniques and effective boundaries". In: *J. Symbolic Comput.* 52 (2013), pp. 72–96.

[8] Gheorghe Craciun and Martin Feinberg. "Multiple equilibria in complex chemical reaction networks: semiopen mass action systems". In: *SIAM J. Appl. Math.* 70.6 (2010), pp. 1859–1877.

[9] Michele D'Adderio and Luca Moci. "Arithmetic matroids, the Tutte polynomial and toric arrangements". In: *Adv. Math.* 232 (2013), pp. 335–367.

[10] Jan Draisma, Emil Horobet, Giorgio Ottaviani, Bernd Sturmfels, and Rekha Thomas. "The Euclidean distance degree of an algebraic variety". In: *Foundations of Computational Mathematics* (to appear). arXiv:1309.0049.

[11] Andreas Dress and Lászlo Lovász. "On some combinatorial properties of algebraic matroids". In: *Combinatorica* 7.1 (1987), pp. 39–48.

[12] Andreas Dress and Walter Wenzel. "Valuated matroids". In: *Adv. Math.* 93.2 (1992), pp. 214–250.

[13] Jean-Charles Faugère, Guillaume Moroz, Fabrice Rouillier, and Mohab Safey El Din. "Classification of the perspective-three-point problem, discriminant variety and real solving polynomial systems of inequalities". In: *ISSAC 2008*. ACM, New York, 2008, pp. 79–86.

[14] Elisenda Feliu and Carsten Wiuf. "Variable elimination in post-translational modification reaction networks with mass-action kinetics". In: *J. Math. Biol.* 66.1-2 (2013), pp. 281–310.

[15] Alex Fink and Luca Moci. "Matroids over a ring". In: *ArXiv e-prints* (2012). arXiv: `1209.6571`.

[16] Luis David García-Puente, Sonja Petrović, and Seth Sullivant. "Graphical models". In: *J. Softw. Algebra Geom.* 5 (2013), pp. 1–7.

[17] Ewgenij Gawrilow and Michael Joswig. "polymake: a Framework for Analyzing Convex Polytopes". In: *Polytopes — Combinatorics and Computation*. Ed. by Gil Kalai and Günter M. Ziegler. Birkhäuser, 2000, pp. 43–74.

[18] Daniel R. Grayson and Michael E. Stillman. *Macaulay2, a software system for research in algebraic geometry*. Available at `http://www.math.uiuc.edu/Macaulay2/`.

[19] Elizabeth Gross, Brent Davis, Kenneth Ho, Dan Bates, and Heather Harrington. "Model selection using numerical algebraic geometry". in preparation, 2015.

[20] Elizabeth Gross, Heather A. Harrington, Zvi Rosen, and Bernd Sturmfels. "Algebraic Systems Biology: A Case Study for the Wnt Pathway". arXiv:1502.03188. 2015.

[21] Heather Harrington, Kenneth Ho, Thomas Thorne, and Michael Stumpf. "Parameter-free model discrimination criterion based on steady-state coplanarity". In: *Proc. Natl. Acad. Sci.* 109 (2012), pp. 15746–15751.

[22] Winfried Hochstättler. "About the Tic-Tac-Toe Matroid". Available at `http://e-archive.informatik.uni-koeln.de/272/`. 1997.

[23] Aubrey Ingleton and Roger Main. "Non-algebraic matroids exist". In: *Bull. London Math. Soc.* 7 (1975), pp. 144–146.

[24] Robert L. Karp, Mercedes Pérez Millán, Tathagata Dasgupta, Alicia Dickenstein, and Jeremy Gunawardena. "Complex-linear invariants of biochemical networks". In: *J. Theoret. Biol.* 311 (2012), pp. 130–138.

[25] Franz Király, Zvi Rosen, and Louis Theran. "Algebraic Matroids with Graph Symmetry". arXiv:1312.3777. 2013.

[26] Franz Király, Louis Theran, and Ryota Tomioka. "The Algebraic Combinatorial Approach for Low-Rank Matrix Completion". In: *Journal of Machine Learning Research* (to appear). http://arxiv.org/abs/1211.4116.

[27] Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald, Hans Lehrach, and Ralf Herwig. *Systems biology.* John Wiley & Sons, 2013.

[28] Kaie Kubjas, Elina Robeva, and Bernd Sturmfels. "Fixed points of the EM algorithm and nonnegative rank boundaries". In: *Ann. Statist.* 43.1 (2015), pp. 422–461.

[29] Kaie Kubjas and Zvi Rosen. "Matrix Completion for the Independence Model". arXiv:1407.3254. 2014.

[30] Bernt Lindström. "A reduction of algebraic representations of matroids". In: *Proc. Amer. Math. Soc.* 100.2 (1987), pp. 388–389.

[31] Bernt Lindström. "Matroids, algebraic and nonalgebraic". In: *Algebraic, extremal and metric combinatorics, 1986 (Montreal, PQ, 1986).* Vol. 131. London Math. Soc. Lecture Note Ser. Cambridge: Cambridge Univ. Press, 1988, pp. 166–174.

[32] Bernt Lindström. "The non-Pappus matroid is algebraic". In: *Ars Combin.* 16.B (1983), pp. 95–96.

[33] Bernt Lindström. "The non-Pappus matroid is algebraic over any finite field". In: *Utilitas Math.* 30 (1986), pp. 53–55.

[34] Radim Lněnička and František Matúš. "On Gaussian conditional independent structures". In: *Kybernetika (Prague)* 43.3 (2007), pp. 327–342.

[35] Saunders Mac Lane. "A lattice formulation for transcendence degrees and $p$-bases". In: *Duke Math. J.* 4.3 (1938).

[36] Adam MacLean, Heather A Harrington, Michael PH Stumpf, and Helen M Byrne. "Mathematical and statistical techniques for systems medicine: The Wnt signaling pathway as a case study". arXiv:1502.01902, to appear in the series "Methods in Molecular Biology". 2015.

[37] Adam MacLean, Zvi Rosen, Helen. Byrne, and Heather A. Harrington. "Parameter-free methods distinguish Wnt pathway models and guide design of experiments". arXiv:1409.0269. 2014.

[38] Arjun Manrai and Jeremy Gunawardena. "The geometry of multisite phosphorylation". In: *Biophys J* 95 (2008), pp. 5533–43.

[39] Nicolette Meshkat and Seth Sullivant. "Identifiable reparametrizations of linear compartment models". In: *Journal of Symbolic Computation* 63 (2014), pp. 46–67.

[40] Mercedes Pérez Millán, Alicia Dickenstein, Anne Shiu, and Carsten Conradi. "Chemical reaction systems with toric steady states". In: *Bulletin of Mathematical Biology* 74.5 (2012), pp. 1027–1065.

[41] Alexander P Morgan and Andrew J Sommese. "Coefficient-parameter polynomial continuation". In: *Applied Mathematics and Computation* 29.2 (1989), pp. 123–160.

[42] Jiawang Nie, Pablo A. Parrilo, and Bernd Sturmfels. "Semidefinite Representation of the $k$-Ellipse". In: *Algorithms in Algebraic Geometry*. Ed. by Alicia Dickenstein, Frank-Olaf Schreyer, and Andrew J. Sommese. Vol. 146. The IMA Volumes in Mathematics and its Applications. Springer, New York, 2008, pp. 117–132.

[43] Hidefumi Ohsugi and Takayuki Hibi. "Toric ideals and their circuits". In: *J. Commut. Algebra* 5.2 (2013).

[44] Giorgio Ottaviani, Pierre-Jean Spaenlehauer, and Bernd Sturmfels. "Exact solutions in structured low-rank approximation". In: *SIAM Journal on Matrix Analysis and Applications* 35.4 (2014), pp. 1521–1542.

[45] James Oxley. *Matroid theory*. Second. Oxford Graduate Texts in Mathematics. Oxford University Press, Oxford, 2011.

[46] M. J. Piff. "Some Problems in Combinatorial Theory". University of Oxford. PhD thesis. University of Oxford, 1972.

[47] Jose Rodriguez and Xiaoxian Tang. "Mathematical and statistical techniques for systems medicine: The Wnt signaling pathway as a case study". arXiv:1501.00334. 2015.

[48] Zvi Rosen. "Computing Algebraic Matroids". arXiv:1403.8148. 2014.

[49] Anne Shiu and Bernd Sturmfels. "Siphons in chemical reaction networks". In: *Bulletin of mathematical biology* 72.6 (2010), pp. 1448–1463.

[50] William A. Stein et al. *Sage Mathematics Software (Version 6.0)*. `http://www.sagemath.org`. The Sage Development Team. 2005.

[51] Bernd Sturmfels. *Gröbner bases and convex polytopes*. Vol. 8. University Lecture Series. American Mathematical Society, Providence, RI, 1996.

[52] Bernd Sturmfels and Volkmar Welker. "Commutative algebra of statistical ranking". In: *J. Algebra* 361 (2012), pp. 264–286.

[53] Jan Verschelde. "Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation". In: *ACM Transactions on Mathematical Software (TOMS)* 25.2 (1999), pp. 251–276.

[54] Eberhard O Voit. *A first course in systems biology*. Garland Science, 2012.

[55] Bartel Leendert van der Waerden. *Algebra. Vol. I*. Translated from the second German edition. New York: Springer-Verlag, 1991.

[56] Dominic J. A. Welsh. *Matroid theory*. Vol. 8. LMS Monographs. Academic Press, London-New York, 1976.