Decomposing Matrices, Tensors, and Images

by

Elina Mihaylova Robeva

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 Benjamin Recht
Professor Martin Olsson

Spring 2016

Decomposing Matrices, Tensors, and Images

Abstract

Decomposing Matrices, Tensors, and Images

by

Elina Mihaylova Robeva

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Bernd Sturmfels, Chair

In this thesis we apply techniques from algebraic geometry to problems arising from optimization and statistics. In particular, we consider data that takes the form of a matrix, a tensor or an image, and we study how to decompose it so as to find additional and seemingly hidden information about its origin and formation. We show that the practical uses of such decompositions are complemented by appealing algebraic and geometric structure.

In Chapter 2 of this thesis we focus on matrix shaped data. The singular value decomposition, which lies at the core of modern algorithms and can be found efficiently, is not always enough to capture the structure of the data. Often times the matrix at hand as well as the elements of its decomposition are required to have a certain positivity structure, and we need to design algorithms and theory to exploit this structure. Statistical mixture models, for instance, are based on finding a nonnegative decomposition of a nonnegative matrix. We study the algebraic and geometric properties of such decompositions in Section 2.1. Another type of decomposition of a nonnegative matrix, which is useful in convex optimization as well as quantum information theory, is positive semidefinite decomposition. Here we require the elements of the decomposition to be positive semidefinite matrices of a given size. We explore this notion in Section 2.2. One of the most appealing properties of a nonnegative matrix is that we can think of it in terms of a pair of nested polyhedra. We rely on this geometric interpretation when studying nonnegative and positive semidefinite decompositions.

In Chapters 3 and 4 we turn our attention to data in the shape of a tensor. It is even more crucial in this case than in the matrix case to find a decomposition, not only because it provides hidden information about the data, but also because it allows us to store the tensor more concisely. However, one of the biggest obstacles in the field is that finding a decomposition of a general tensor is NP-hard. Inspired by the spectral theorem and the singular value decomposition for matrices, we study tensors whose decomposition consists of elements with an orthogonality structure. We call such tensors orthogonally decomposable, or odeco. One of their best properties is that, like matrices, odeco tensors can be decomposed efficiently. In Chapter 3 we study the spectral properties of such tensors. We give a formula for their eigenvectors and singular vector tuples. We note that computing these for a general

tensor is hard both algebraically and computationally. In Chapter 4 we study the variety of orthogonally decomposable tensors, and we give polynomial equations that cut it out. We do this by showing that a tensor is orthogonally decomposable if and only if a given algebra that arises from it is associative, yet another appealing property of odeco tensors. Despite all of these appealing properties, odeco tensors constitute a very low-dimensional variety. This is why in Section 4.2 we conclude our study of tensors by generalizing the notion of orthogonally decomposable tensors to that of frame decomposable tensors, which now cover the space of all tensors.

In Chapter 5 we study super-resolution imaging. The aim here is, given a low-resolution blurred image, to increase the resolution and remove the blur. This is achieved by decomposing the image into a sum of simpler images, one for each point source of light. We encode the locations of the point sources of light and their intensities in a discrete measure, and propose a convex optimization problem in the space of measures to find this unknown measure. We show that in the absence of noise and in the case of a one-dimensional image, the global optimum of this optimization problem recovers the true locations.

Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my advisor, Bernd, for his endless support, motivation, understanding, and inspiration. His undivided attention, constant encouragement, and personal example have been instrumental in my academic development. Over the entire course of my PhD Bernd's guidance and attention kept me extremely motivated, productive, and excited about my work. His support and advice went beyond research, and I am grateful for his complete understanding during difficult times for me. I admire Bernd's ability to match people who are good at working together, and to suggest to them interesting problems to work on. He has built a really great community and I am grateful to be part of it. I truly believe that Bernd is the best possible advisor, and I feel extremely privileged to have have been able to work with him.

Further, I would like to thank Ben for giving me the opportunity to be part of his group. He was always very enthusiastic during our work on super-resolution imaging together with Geoff, and it was a pleasure being part of it. Our math meetings were always very exciting and motivating. Ben gave me lots of guidance and advice, and I am very grateful for being part of his community.

I would also like to thank Jan for the really interesting and beautiful collaboration together with Ada and Emil. It was a pleasure visiting Eindhoven and talking about math. I am also grateful for numerous helpful discussions with Jan, related to many of the questions that came up in my research. I would also like to thank Kristian for all the interesting and inspiring discussions that we had during my visit in Norway, and during many conferences. I would also like to thank Caroline for inviting me to conferences, and for many interesting discussions about math and academia.

I would like to thank Kaie for two wonderful collaborations. I am grateful for how well we were able to work together and for how nicely our mathematical languages match. I would like to thank Bernd for introducing us to each other, and for helping create a great research team that lead to great friendship. Similarly, I would also like to thank Anna for being a wonderful collaborator and friend.

I would like to thank the Berkeley Graduate division for the UC Berkeley Graduate Fellowship which supported me during my first two years. I would also like to thank the Max Plank Institute for Mathematics in Bonn, Germany, and the National Institute of Mathematical Sciences (NIMS) in Daejeon, Korea, both of which supported me while I was visiting.

I would also like to thank my family for the immense support they have provided during my PhD. I am so grateful for all the love and care that I have received from my mom. They have played an instrumental role in my well-being and work. She has always been able to sense exactly what I need and to implement it as quickly as possible. I would like to thank Geoff for his endless love and support. During the last three years, Geoff's appearance in my life made me happier and more excited than I had ever been, and had a tremendous positive impact on my research. Our project together was a pleasure, and I am very grateful that we are not only a great team in life, but also in research. I would like to thank my dad for

always being such a loving, down-to-earth, logical, and positive person to talk to. He has always been able to give me great advice in all areas of life. I am so grateful to have become part of Londa, Robert, Geoff, and Jon's family. They have always been as nice as one could be to me, and have served as the best possible example both academically and in life.

I would like to thank all my friends and colleagues for always being there for me, and for being able to share happiness, celebrations, adventures, as well as hard times.

Finally, I would like to thank Bernd once again for going over my thesis multiple times and helping me improve it tremendously.

Chapter 1

Introduction

When we observe a signal, it is often useful to decompose it into simpler meaningful parts. This allows us to discover additional seemingly hidden information about its origin.

In the famous Netflix Prize problem [152], we observe a partially filled matrix with each row corresponding to a user, each column corresponding to a movie, and each entry indicating the rating a given user assigns to a given movie. Up to a small error this matrix can be written as the sum of a few rank-one matrices. It turns out that these correspond to the different traits users have, such as whether they like romance, what age they are, or whether they like animated movies. Finding these rank-one matrices is what allows Netflix to predict users' future ratings based on the ratings they have provided in the past.

In astronomy, we often observe a very low-resolution blurred picture of distant galaxies. In order to increase the resolution and recover the exact locations of the stars, we express the picture as the sum of several simpler pictures, one for each star. The field of super-resolution imaging provides tools to find such a decomposition.

In both of the above examples, the task at hand is achieved by decomposing the given signal into a sum of simpler parts. Developing theory and algorithms for finding such decompositions is a major topic in statistics and computer science. The goal of this chapter is to introduce to several different types of matrix, tensor, and image decompositions, which are further studied in this thesis.

1.1 Matrices

Matrix decompositions lie at the core of modern algorithms in scientific computing. Given a real $m \times n$ matrix M, its *rank* is the smallest number r such that there exist matrices $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{r \times n}$ satisfying

$$M = AB.$$

Equivalently, the rank of M is the smallest number r for which there exist vectors $a_1, \ldots, a_m \in \mathbb{R}^r$ (corresponding to the rows of A above) and $b_1, \ldots, b_n \in \mathbb{R}^r$ (corresponding to the columns

of B above) such that

$$M_{ij} = \langle a_i, b_j \rangle, \text{ for all } 1 \leq i \leq m, 1 \leq j \leq n.$$

In practice, the observed matrices often have nonnegative entries (for example, contingency tables or probability distributions). Moreover, the desired decomposition often imposes additional structure on the vectors $a_1, \ldots, a_m, b_1, \ldots, b_n$. For instance, they may be required to lie in a *cone*, such as the nonnegative orthant (nonnegative rank) or the cone of positive semidefinite matrices (positive semidefinite rank). These two different versions of matrix decomposition are studied in Chapter 2.

One of the most important and useful decompositions of a matrix is the *singular value decomposition*. Given $M \in \mathbb{R}^{m \times n}$ we decompose it as

$$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where $u_1, \ldots, u_r \in \mathbb{R}^m$ are orthonormal, $v_1, \ldots, v_r \in \mathbb{R}^n$ are orthonormal, and $\sigma_1, \ldots, \sigma_r$ are nonnegative real numbers. The singular value decomposition can be computed very efficiently, which is why it is the most commonly used decomposition. In Chapters 3 and 4 we study a generalization of this decomposition to tensors, which has many appealing properties.

1.1.1 Nonnegative matrices and nested polytopes

An important feature of nonnegative matrices is that they can be represented by pairs of nested polyhedra. Here, a polyhedron is a finite intersection of closed halfspaces.

Definition 1.1.1. *Let $P \subseteq Q \subseteq \mathbb{R}^{d-1}$ be two nested polyhedra such that P is bounded, i.e. it is a polytope. Assume that P has vertex description*

$$P = conv(v_1, \ldots, v_m)$$

for some $v_1, \ldots, v_m \in \mathbb{R}^{d-1}$ and Q has facet description

$$Q = \{x \in \mathbb{R}^{d-1} | \langle x, w_j \rangle \leq z_j, \forall j = 1, \ldots, n\}$$

for some $w_1, \ldots, w_n \in \mathbb{R}^{d-1}$ and $z_1, \ldots, z_n \in \mathbb{R}$. The slack matrix *of the pair P, Q, denoted by $S_{P,Q}$, is the $m \times n$ nonnegative matrix whose i, j-th entry is*

$$[S_{P,Q}]_{i,j} = z_j - \langle v_i, w_j \rangle.$$

We remark here that the condition that $P \subseteq Q$ is equivalent to the condition $\langle v_i, w_j \rangle \leq z_j$, for every i and j. Therefore, the matrix $S_{P,Q}$ is nonnegative. We also remark that for given P and Q, one can define their slack matrix in many different ways, however all

properties of interest in this thesis such as rank, nonnegative rank, and positive semidefinite rank are preserved regardless of which way we define $S_{P,Q}$. Moreover, the vectors $(-v_1, 1), \ldots, (-v_p, 1), (w_1, z_1), \ldots, (w_q, z_q) \in \mathbb{R}^d$ give a rank-d factorization of $S_{P,Q}$, and therefore $S_{P,Q}$ has rank at most d. In fact, it is easy to see that if P and Q are full-dimensional polyhedra in $\mathbb{R}^{d-1}$, then $S_{P,Q}$ has rank exactly d.

Conversely, given a nonnegative matrix $M \in \mathbb{R}^{m \times n}_{\geq 0}$, we now explain how to construct polytopes P and Q such that $M = S_{P,Q}$. Firstly, if M does not have a zero row, one can rescale its rows so that $M\mathbf{1} = \mathbf{1}$. This rescaling will not change the properties of M that we study. In particular, it doesn't change its rank, nonnegative rank, or positive semidefinite rank.

Lemma 1.1.2 (Lemma 4.1 in [65]). *Let $M \in \mathbb{R}^{m \times n}_{\geq 0}$ be a nonnegative matrix and assume that $M\mathbf{1} = \mathbf{1}$. Let $d = \operatorname{rank}(M)$. Then there exist polytopes $P \subseteq Q \subseteq \mathbb{R}^{d-1}$ such that M is the slack matrix of the pair P, Q.*

We give a brief outline of the proof of this lemma. First, we show that we can find a factorization $M = AB$ such that $A \in \mathbb{R}^{p \times d}$, $B \in \mathbb{R}^{d \times q}$, and in addition $A\mathbf{1} = \mathbf{1}$ and $B\mathbf{1} = \mathbf{1}$. To do that, first choose d linearly independent rows of M and define B to equal the $d \times q$ submatrix of M with these rows. Since M has rank d, then, one can uniquely find the matrix A so that $M = AB$. Now, since the rows of B are a subset of the rows of M and $M\mathbf{1} = \mathbf{1}$, then, $B\mathbf{1} = \mathbf{1}$. Therefore, $\mathbf{1} = M\mathbf{1} = AB\mathbf{1} = A\mathbf{1}$.

Next, we construct the following *cones* $\tilde{P} \subseteq \tilde{Q} \subseteq \mathbb{R}^d$. We define $\tilde{P}$ to be the convex cone spanned by the rows of A and $\tilde{Q}$ to be the convex cone with facets defined by functionals arising from the columns of B. The bounded polytopes P and Q are then obtained by intersecting the cones $\tilde{P}$ and $\tilde{Q}$ with the hyperplane $x_d = 1 - \sum_{i=1}^{d-1} x_i$. See Figure 1.1.

By deriving the explicit definition of P and Q from this construction, we can see that M is the slack matrix of P, Q and that they are both bounded.

1.1.2 Cones and cone rank

A *(convex) cone* in $\mathbb{R}^n$ is a subset $C \subseteq \mathbb{R}^n$ such that $\alpha x + \beta y \in C$ for any $x, y \in C$ and any positive scalars α and β. Our two main examples of cones will be the *nonnegative orthant* $\mathbb{R}^r_{\geq 0}$ inside $\mathbb{R}^r$, and the cone of $k \times k$ symmetric positive semidefinite matrices $\mathcal{S}^k_+$ inside the space $\mathcal{S}^k$ of $k \times k$ symmetric matrices with real entries.

Let $C \subseteq \mathbb{R}^n$ be a convex cone, and assume that $\mathbb{R}^n$ is equipped with an inner product $\langle \cdot, \cdot \rangle$. The *dual cone* to C is the set

$$C^* = \{w \in \mathbb{R}^n | \langle v, w \rangle \geq 0 \text{ for all } v \in C\}.$$

It is also a convex cone. If C is equal to its dual cone, then C is *self-dual*. It is easy to see that the nonnegative orthant $\mathbb{R}^r_{\geq 0}$ is self-dual, where $\mathbb{R}^r$ is equipped with the Euclidean dot product. We claim that the cone of positive semidefinite matrices $\mathcal{S}^k_+$ is also self-dual. Here $\mathcal{S}^k$ is equipped with the *trace inner product* given by $\langle A, B \rangle = \operatorname{trace}(AB)$ for $A, B \in \mathcal{S}^k$.

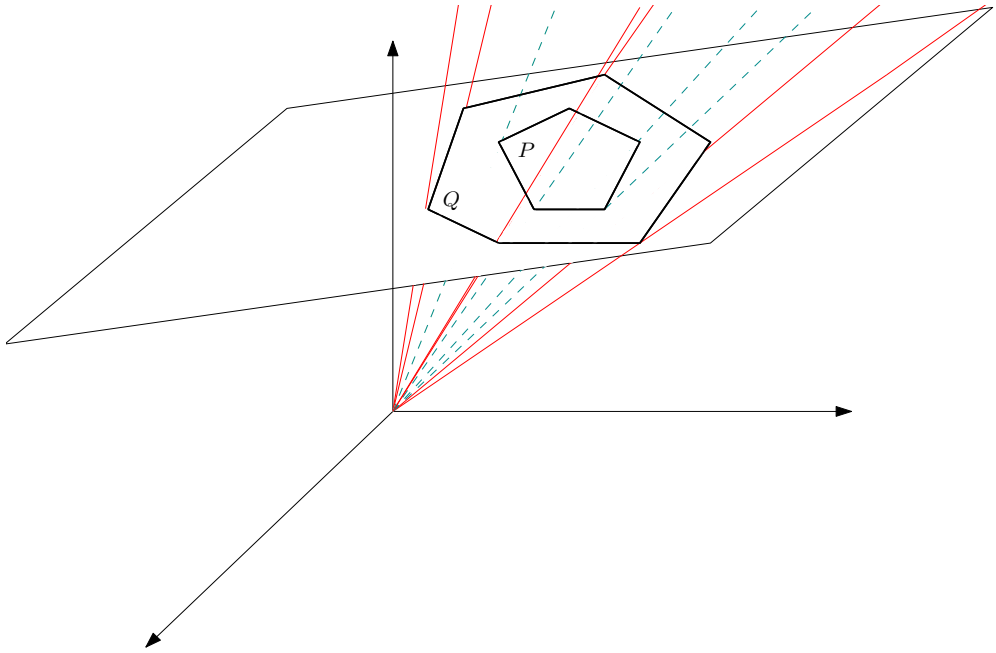


Figure 1.1: The two cones (one in red and one in dashed blue) emerging from the rows of A and the columns of B, and the hyperplane that cuts them. This is how we obtain P and Q from the matrix $M = AB$.

To prove this claim, let $A \in \mathcal{S}^k$ be such that $\text{trace}(AB) \geq 0$ for all $B \in \mathcal{S}^k_+$. Let B be the rank-one matrix $B = vv^T$ where $v \in \mathbb{R}^k$ is any vector. Then,

$$0 \leq \text{trace}(AB) = \text{trace}(Avv^T) = v^T Av.$$

Thus, for every $v \in \mathbb{R}^k$, $v^T Av \geq 0$, i.e. $A \in \mathcal{S}^k_+$. Therefore, $(\mathcal{S}^k_+)^* \subseteq \mathcal{S}^k_+$. Now, let A be any matrix in $\mathcal{S}^k_+$ and let $B \in \mathcal{S}^k_+$. Then, using the spectral theorem and the fact that positive semidefinite matrices have nonnegative eigenvalues, we can write $A = UU^T$ and $B = VV^T$. Thus, $\text{trace}(AB) = \text{trace}(UU^TVV^T) = \text{trace}(((U^TV)^TU^TV)) \geq 0$ since this is a positive semidefinite matrix.

1.1.2.1 Nonnegative rank

Given a nonnegative matrix $M \in \mathbb{R}^{m \times n}_{\geq 0}$, its *nonnegative rank*, denoted by $\text{rank}_+(M)$, is the smallest positive integer r such that there exist vectors $a_1, \ldots, a_m, b_1, \ldots, b_n \in \mathbb{R}^r_{\geq 0}$ with

$$M_{ij} = \langle a_i, b_j \rangle, \quad \forall i, j.$$

Such a decomposition is useful when the application at hand requires that vectors $a_1, \ldots, a_m$, $b_1, \ldots, b_n$ are nonnegative.

Example 1.1.3 (Mixture models)**.** *In statistics, the joint distribution of two seemingly dependent random variables X and Y is sometimes explained by a third hidden random variable Z such that X and Y are independent given Z. For example, consider the random variables*

$$X = \begin{cases} 0 & \text{person is bald} \\ 1 & \text{person has short hair} \\ 2 & \text{person has long hair} \end{cases} \qquad Y = \begin{cases} 0 & \text{person does not watch football} \\ 1 & \text{watches 0 to 2 hours of football per week} \\ 2 & \text{watches } > \text{ 2 hours of football per week.} \end{cases}$$

Now suppose that after asking 1200 people how much football they watch and what length their hair is, we obtain the following contingency table

$$M = \begin{bmatrix} 70 & 65 & 65 \\ 180 & 210 & 210 \\ 170 & 115 & 115 \end{bmatrix},$$

where the rows correspond to the three values X takes and the columns correspond to the three values Y takes. The matrix M has rank 2, so X and Y are dependent. However, if we are to record the answers for men and women separately, we would obtain

$$M = \begin{bmatrix} 70 & 65 & 65 \\ 180 & 210 & 210 \\ 170 & 115 & 115 \end{bmatrix} = \underbrace{\begin{bmatrix} 50 & 25 & 25 \\ 100 & 50 & 50 \\ 150 & 75 & 75 \end{bmatrix}}_{women} + \underbrace{\begin{bmatrix} 20 & 40 & 40 \\ 80 & 160 & 160 \\ 20 & 40 & 40 \end{bmatrix}}_{men}.$$

Since the two summands are rank-one matrices, we see that X and Y are independent given gender. Define the variable Z to equal 1 if the person is female and 2 if the person is male. Then the empirical joint probability distribution of X and Y is the 3×3 matrix

$$P(X,Y) = P(Z=1)P(X,Y|Z=1) + P(Z=2)P(X,Y|Z=2) =$$

$$= \frac{1}{2} \begin{bmatrix} \frac{1}{6} \\ \frac{1}{3} \\ \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2}, \frac{1}{4}, \frac{1}{4} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \frac{1}{6} \\ \frac{2}{3} \\ \frac{1}{6} \end{bmatrix} \begin{bmatrix} \frac{1}{5}, \frac{2}{5}, \frac{2}{5} \end{bmatrix} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{6} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{bmatrix},$$

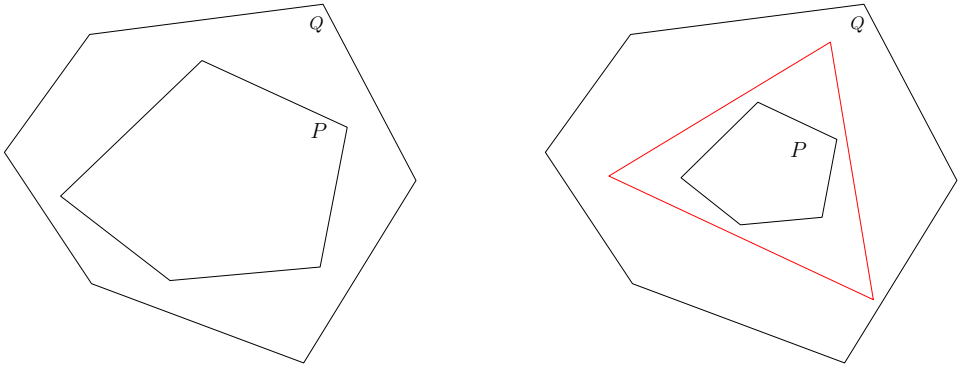
which has nonnegative rank 2. Finding the nonnegative decomposition of this matrix would allows us to learn the hidden parameters of the distribution. In this case we would learn that the amount of hair and the amount of football watched are independent given the gender of a person.

Observe that the rank and nonnegative rank of a matrix $M \in \mathbb{R}^{m \times n}_{\geq 0}$ satisfy the following inequalities

$$\operatorname{rank}(M) \leq \operatorname{rank}_+(M) \leq \min\{m, n\}.$$

We denote by $\mathcal{M}_{d,r}^{m\times n}$ (for short $\mathcal{M}_{d,r}$) the set of nonnegative $m\times n$ matrices of rank at most d and nonnegative rank at most r. We often call $\mathcal{M}_{d,r}$ the r-th *mixture model*. It is a subset of the *variety* $\mathcal{V}_d$ of $m\times n$ matrices of rank at most d, defined by the vanishing of their $(d+1)$-minors. In fact, whenever $d \leq r \leq \min\{m,n\}$, $\mathcal{M}_{d,r}$ is a full-dimensional semialgebraic subset of $\mathcal{V}_d$. In addition to the equations defining $\mathcal{V}_d$ coming from the $(d+1)$-minors, $\mathcal{M}_{d,r}$ is cut out by some polynomial inequalities. In our work [105], described in Chapter 2, we study the set $\mathcal{M}_{d,r}$ for small values of d and r and attempt to find its semialgebraic description. We rely heavily on the following geometric interpretation of nonnegative rank.

Lemma 1.1.4 (Lemma 2.2 in [116]). *Let $M \in \mathbb{R}_{\geq 0}^{m\times n}$ have rank d and let $P \subseteq Q \subset \mathbb{R}^{d-1}$ be obtained as described in Lemma 1.1.2. Then, M has a size r nonnegative factorization if and only if there exists a polytope Δ with r vertices such that $P \subseteq \Delta \subseteq Q$.*



Nonnegative rank greater than 3. Nonnegative rank equal to 3.

Figure 1.2: The figure on the left depicts the two nested polytopes P and Q that arise from a matrix M of nonnegative rank greater than 3 since one cannot fit a triangle in between them. The figure on the right depicts nested polytopes P and Q arising from a matrix M of nonnegative rank 3.

Suppose M has rank 3. Then, the polytopes P and Q lie in $\mathbb{R}^2$ and M has nonnegative rank 3 if and only if we can nest a triangle in between P and Q, as in Figure 1.2. In our work [105], presented in Section 2.1, we exploit this geometric interpretation of nonnegative rank. It is quite interesting to consider the matrices on the *boundary* of $\mathcal{M}_{3,3}$, considered as a subset of $\mathcal{V}_d$. They correspond to pairs of nested polygons $P \subseteq Q \subseteq \mathbb{R}^2$ in between which we can fit a triangle Δ, which cannot be moved in a rigid way while still remaining between P and Q. We give a complete geometric and algebraic characterization of all such matrices.

1.1.2.2 Positive semidefinite rank

Given a nonnegative matrix $M \in \mathbb{R}_{\geq 0}^{m\times n}$, its *positive semidefinite rank* (or *psd rank*) is the smallest positive integer r such that there exist matrices $A_1,\ldots,A_m,B_1,\ldots,B_n \in \mathcal{S}_+^r$ such

that

$$M_{ij} = \langle A_i, B_j \rangle, \text{ for all } 1 \leq i \leq m, 1 \leq j \leq n.$$

Example 1.1.5. *Consider the following* 3×3 *matrix*

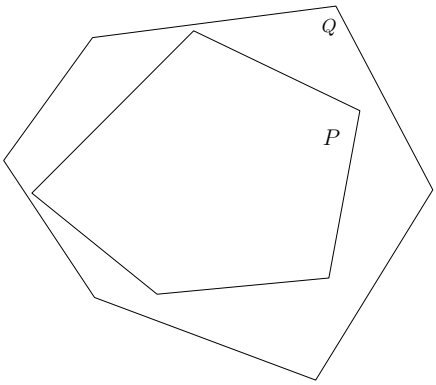
$$M = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

It satisfies $\text{rank}(M) = \text{rank}_+(M) = 3$. *However, its psd rank is equal to 2 since it admits the following size-two psd factorization*

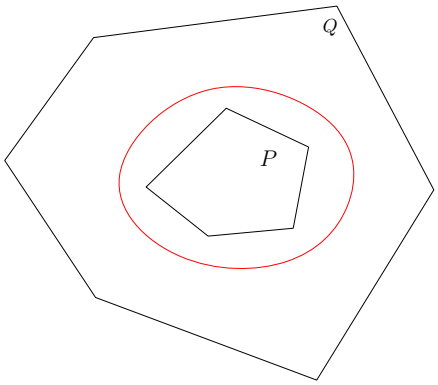
$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad A_3 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$
$$B_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad B_3 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Similarly to nonnegative rank, one can describe positive semidefinite rank via nested polytopes.

Theorem 1.1.6 (Proposition 3.6 in [79])**.** *Let* $M \in \mathbb{R}^{m \times n}_{\geq 0}$ *be a matrix of rank* d *and let* $P \subseteq Q \subseteq \mathbb{R}^{d-1}$ *be obtained as in Lemma 1.1.2. Then,* $\text{rank}_{psd}(M)$ *is the smallest integer* r *for which there exists an affine subspace* $\mathcal{L}$ *of* $\mathcal{S}^r$ *and a linear map* π *such that* $P \subseteq \pi(\mathcal{L} \cap \mathcal{S}^r_+) \subseteq Q$.



Positive semidefinite rank greater than 2.



Positive semidefinite rank equal to 2.

Figure 1.3: The left figure depicts the polytopes coming from a matrix M which has psd rank greater than 2 since one cannot fit an ellipse between P and Q. The right figure shows the polytopes arising from a matrix M which has psd rank 2.

Consider a matrix $M \in \mathbb{R}^{m\times n}_{\geq 0}$ of rank 3. It gives rise to two nested polygons $P \subseteq Q \subset \mathbb{R}^2$. According to Proposition 4.1 in [79], the matrix M has psd rank 2 if and only if we can fit an ellipse between the two polygons P and Q, see Figure 1.3.

Section 2.2 is based on joint work with Kaie Kubjas and Richard Robinson titled *Positive semidefinite rank and nested spectrahedra* [104]. We study the geometry of the space $\mathcal{P}_{d,r}$ of matrices of rank at most d and psd rank at most r. We give a complete semialgebraic description of its *boundary* when $d = 3$ and $r = 2$ and we give partial results towards a description for general rank d and psd rank $r = d - 1$.

1.2 Tensors

A natural generalization of matrices, tensors have direct applications in modern data analysis. Studying the spectral properties and decompositions of tensors is of utmost importance to being able to handle information that comes in more than two dimensions.

A *tensor* T of order d and format $n_1 \times n_2 \times \cdots \times n_d$ is an $n_1 \times n_2 \times \cdots \times n_d$ table with entries in a field $\mathbb{K}$, which will be $\mathbb{R}$ or $\mathbb{C}$ in this thesis. The vector space of tensors of this format is denoted by $\mathbb{K}^{n_1} \otimes \mathbb{K}^{n_2} \otimes \cdots \otimes \mathbb{K}^{n_d}$. Given such a tensor T, its entries are denoted by $T_{i_1,\ldots,i_d}$, where $1 \leq i_j \leq n_j$ for all j. A tensor $T \in \mathbb{K}^n \otimes \cdots \otimes \mathbb{K}^n$ is *symmetric* if $T_{i_1,\ldots,i_d} = T_{i_{\sigma(1)},\ldots,i_{\sigma(d)}}$ for any permutation σ of $\{1, 2, \ldots, d\}$.

1.2.1 Tensor decompositions

Similar to matrix decomposition, tensor decomposition has numerous applications in statistics, neuroscience, signal processing, computer vision, data analysis, and others [101].

A *symmetric tensor* is an $n \times n \times \cdots \times n$ (d times) tensor T such that $T_{i_1\ldots i_d} = T_{i_{\sigma(1)}\cdots i_{\sigma(d)}}$ for any permutation σ on $\{1, \ldots, d\}$. The space of such tensors is denoted by $S^d(\mathbb{K}^n)$. Given a symmetric tensor $T \in S^d(\mathbb{K}^n)$, a *symmetric decomposition* is an expression of the form

$$T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes d},$$

where $v_i \in \mathbb{C}^n$. The smallest r for which such a decomposition exists is the *symmetric rank* (or *Waring rank*) of T. The tensors of the form $v_i^{\otimes d}$ are *rank-one symmetric* tensors.

For an ordinary tensor $T \in \mathbb{K}^{n_1} \otimes \cdots \otimes \mathbb{K}^{n_d}$ a *decomposition* is an expression of the form

$$T = \sum_{i=1}^{r} \sigma_i v_i^{(1)} \otimes \cdots \otimes v_i^{(d)},$$

where $v_1^{(j)}, \ldots, v_d^{(j)} \in \mathbb{K}^{n_j}$. The smallest r for which such a decomposition exists is the *rank* of T, and the tensors of the form $v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}$ are *rank-one* tensors.

Example 1.2.1. *Suppose X_1, X_2, and X_3 are discrete random variables such that X_i takes values in $\{1, 2, \ldots, n_i\}$. Their joint probability distribution is the tensor $P \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_3}$ such that $P_{i_1,i_2,i_3} = \mathbb{P}(x_1 = i_1, X_2 = i_2, X_3 = i_3)$. Now suppose that X_1, X_2, X_3 are independent given a random variable Z taking values in $\{1, \ldots, r\}$. Then, given that $Z = k$, the joint distribution of X_1, X_2, X_3 is a rank one tensor*

$$\mathbb{P}(X_1, X_2, X_3|Z = k) = \mathbb{P}(X_1|Z = k) \otimes \mathbb{P}(X_2|Z = k) \otimes \mathbb{P}(X_3|Z = k),$$

and, therefore the total joint distribution of X_1, X_2, X_3 is a rank r tensor

$$P = \sum_{k=1}^{r} \mathbb{P}(Z = k) \cdot \mathbb{P}(X_1, X_2, X_3|Z = k) =$$

$$= \sum_{k=1}^{r} \mathbb{P}(Z = k) \cdot \mathbb{P}(X_1|Z = k) \otimes \mathbb{P}(X_2|Z = k) \otimes \mathbb{P}(X_3|Z = k).$$

Therefore, if we observe P, finding its decomposition allows us to discover the hidden parameters of the distribution of X_1, X_2, and X_3.

In both the symmetric and the ordinary case, if the elements of the decomposition of T are allowed to have complex entries, there exists a positive integer, called the *generic rank*, such that the set of tensors of this rank is Zariski dense in the set of all tensors of a given format. According to the Alexander-Hirschowitz Theorem [27], when $d \geq 3$, the generic rank of a symmetric tensor $T \in S^d(\mathbb{C}^n)$ equals $\left\lceil \frac{\binom{n+d-1}{d}}{n} \right\rceil$ except in a finite number of cases in which it is one more than this number [2]. For $n \times n$ matrices the generic rank is equal to n and *all* such matrices have rank at most n. However, for $d \geq 3$ there always exist tensors of rank higher than the generic rank.

Finding the decomposition of a given tensor T is one of the most important problems in the field. However, it has been shown that in general, it is an NP-hard problem [92]. Algorithms for it have been proposed by many authors, for example [26, 120]. In this thesis, we focus our attention on a special type of tensors, called *orthogonally decomposable tensors* whose decomposition can be found efficiently.

1.2.2 Orthogonal Tensor Decomposition

The Spectral Theorem states that for any $n \times n$ real symmetric matrix M there exists an orthonormal basis of eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$ with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that

$$M = \begin{bmatrix} | & \cdots & | \\ v_1 & \cdots & v_n \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & v_n \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_n^T & - \end{bmatrix} = \sum_{i=1}^{n} \lambda_i v_i v_i^T = \sum_{i=1}^{n} \lambda_i v_i^{\otimes 2}.$$

We generalize this decomposition to tensors of order $d \geq 3$. A tensor $T \in S^d(\mathbb{R}^n)$ is *symmetric orthogonally decomposable* (or *symmetric odeco*) if it has a decomposition

$$T = \sum_{i=1}^{n} \lambda_i v_i^{\otimes d},$$

where $v_1, \ldots, v_n \in \mathbb{R}^n$ form an orthonormal basis. Their appealing structure allows odeco tensors to be decomposed efficiently, for example via the *tensor power method* [8]. Since there can be at most n orthonormal vectors in $\mathbb{R}^n$, the rank of an odeco tensor is at most n, which is significantly smaller than the rank of a generic tensor $T \in S^d(\mathbb{R}^n)$. What is more, the set of odeco tensors is a strict subset of the set of tensors of rank at most n. However, we can use a procedure, called *whitening*, which allows us to transform a tensor of rank n into an odeco tensor, find the decomposition of the odeco tensor, and then transform back to obtain a decomposition of the original tensor [8].

Orthogonally decomposable tensors can also be defined in the non-symmetric case. Recall first that singular value decomposition allows us to write *any* matrix $M \in \mathbb{R}^m \otimes \mathbb{R}^n$ as

$$M = \sum_{i=1}^{r} \sigma_i v_i^{(1)} (v_i^{(2)})^T = \sum_{i=1}^{r} \sigma_i v_i^{(1)} \otimes v_i^{(2)},$$

where $v_1^{(1)}, \ldots, v_r^{(1)} \in \mathbb{R}^m$ are orthonormal and $v_1^{(2)}, \ldots, v_r^{(2)} \in \mathbb{R}^n$ are orthonormal.

A tensor $T \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ is *orthogonally decomposable* (or *odeco*) if it has a decomposition

$$T = \sum_{i=1}^{n} \sigma_i v_i^{(1)} \otimes \cdots \otimes x_i^{(d)},$$

where $v_1^{(j)}, \ldots, v_n^{(j)} \in \mathbb{R}^{n_j}$ are orthonormal for all $j \in \{1, \ldots, d\}$, $n \leq \max\{n_1, \ldots, n_d\}$, and $\sigma_1, \ldots, \sigma_n \in \mathbb{R}$. Like in the symmetric case, odeco tensors can be decomposed efficiently via an iterative tensor power method [8].

In joint work with Ada Boralevi, Jan Draisma and Emil Horobeţ titled *Orthogonal and unitary tensor decomposition from an algebraic perspective*, and presented in Section 4.1, we find equations defining the variety of orthogonally decomposable tensors. It turns out that odeco tensors correspond very beautifully to associative algebras.

Consider a tensor $T \in S^3(\mathbb{R}^n)$. Let $V = \mathbb{R}^n$ be equipped with the usual inner product. We give V the structure of an algebra arising from the tensor T as follows. For two elements $u, v \in V$, we define their product to be

$$u \star v = T(u, v, \cdot) \in V.$$

Theorem 1.2.2. *The tensor $T \in S^3(\mathbb{R}^n)$ is orthogonally decomposable if and only if the algebra V with product $\star$ arising from T is associative.*

This correspondence gives us more insight into the theory of orthogonally decomposable tensors and allows us to find the equations that define the (real) variety of such tensors. Section 4.1 is dedicated to this work. We generalize this theorem to higher order tensors as well as to tensors which are not necessarily symmetric. We conclude this subsection with an example that illustrates Theorem 1.2.2.

Example 1.2.3. *Let $n = 2$ and fix a basis $\{a, b\}$ of $\mathbb{R}^2$. A $2 \times 2 \times 2$ symmetric tensor T with entries T_{ijk} defines the algebra structure*

$$a \star a = T_{000}a + T_{100}b, \qquad a \star b = T_{100}a + T_{110}b,$$

$$b \star a = T_{100}a + T_{110}b, \qquad b \star b = T_{110}a + T_{111}b.$$

In general this algebra is not associative:

$$b \star (a \star a) = (T_{000}T_{100} + T_{100}T_{110})a + (T_{000}T_{110} + T_{100}T_{111})b,$$

$$(b \star a) \star a = (T_{000}T_{100} + T_{110}T_{100})a + (T_{100}^2 + T_{110}^2)b.$$

It turns out that

$$b \star (a \star a) = (b \star a) \star a \iff T_{000}T_{110} + T_{100}T_{111} = T_{100}^2 + T_{110}^2 \iff T \text{ is odeco.}$$

1.2.3 Decomposing Tensors into Frames

As we mentioned above, even though odeco tensors have appealing properties, including the fact that they can be decomposed efficiently, they constitute a very low-dimensional part of the set of all tensors. In joint work with Luke Oeding and Bernd Sturmfels titled *Decomposing tensors into frames*, and presented in Section 4.2, we generalize the notion of orthogonally decomposable tensors while still imposing extra structure on the decomposition. Instead of an orthonormal basis, we use the more general notion of a *finite unit norm tight frame* (or *funtf*) [29]. A set of vectors $v_1, \ldots, v_r \in \mathbb{R}^n$ forms a funtf if

$$\begin{bmatrix} | & \cdots & | \\ v_1 & \cdots & v_r \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_r^T & - \end{bmatrix} = Id_n, \quad \text{and} \quad ||v_i||^2 = 1, i = 1, \ldots, r,$$

where Id_n is the $n \times n$ identity matrix. A symmetric tensor $T \in S^d(\mathbb{C}^n)$ is *frame decomposable* (or *fradeco*) if it has a decomposition of the form

$$T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes d},$$

where $v_1, \ldots, v_r \in \mathbb{R}^n$ form a funtf. Section 4.2 is dedicated to fradeco tensors. We study the variety of such tensors as well as methods for finding their decompositions.

1.2.4 Spectral theory

Analogous to the definition for symmetric matrices, given a symmetric tensor $T \in \mathbb{S}^d(\mathbb{K}^n)$, a vector $x \in \mathbb{C}^n$ is an *eigenvector* of T with *eigenvalue* $\lambda \in \mathbb{C}$ if

$$T \cdot x^{d-1} = \lambda x.$$

Here $T \cdot x^{d-1}$ is a vector in $\mathbb{C}^n$ which equals the contraction of T with x along $d-1$ of its dimensions. More precisely, its i-th coordinate is

$$(T \cdot x^{d-1})_i := \sum_{i_1,\ldots,i_{d-1}} T_{i_1,\ldots,i_{d-1},i} x_{i_1} \cdots x_{i_{d-1}}.$$

Two eigenvector-eigenvalue pairs (x, λ) and (x', λ') are equivalent if $(x, \lambda) = (tx', t^{d-2}\lambda')$ for some scalar $t \in \mathbb{C}^*$.

The eigenvectors of T can also be described as the set of (nonzero) fixed points of the *tensor power iteration*

$$x \mapsto \frac{T \cdot x^{d-1}}{||T \cdot x^{d-1}||},$$

where $0/||0|| = 0$. Alternatively, they can also be characterized via a variational approach [112]. They are the critical points of the optimization problem

$$\begin{aligned} \text{maximize} \quad & T \cdot x^d \\ \text{such that} \quad & ||x|| = 1, \end{aligned}$$

where $T \cdot x^d \in \mathbb{C}$ is the contraction of T with x along all d of its dimensions. In symbols, $T \cdot x^d = \sum_{i_1,\ldots,i_d} T_{i_1,\ldots,i_d} x_{i_1} \cdots x_{i_d}$. If $x \in \mathbb{C}^n$ is a maximizer of this optimization problem, then the tensor $x^{\otimes d}$ is the best rank-one approximation to T.

Recall that a general symmetric $n \times n$ matrix has exactly n eigenvectors. It was shown in [37] that a general symmetric tensor $T \in S^d(\mathbb{C}^n)$ has finitely many eigenvectors and their number is exactly

$$\frac{(d-1)^n - 1}{d-2}.$$

Now consider a tensor $T \in \mathbb{K}^{n_1} \otimes \cdots \otimes \mathbb{K}^{n_d}$ which is not necessarily symmetric and $n_1, \ldots, n_d$ are not necessarily equal. As with rectangular matrices, we can no longer define eigenvectors. The right notion is now that of singular vector tuples. We define $(x^{(1)}, \ldots, x^{(d)}) \in \mathbb{C}^{n_1} \times \cdots \times \mathbb{C}^{n_d}$ to be a *singular vector tuple* of T if the vectors $x^{(1)}, \ldots, x^{(d)}$ are nonzero, and for every $1 \leq j \leq d$,

$$T(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots, x^{(d)}) \text{ is parallel to } x^{(j)}.$$

In other words, for every j, contracting T along its k-th dimension with $x^{(k)}$ for every $k \neq j$ should yield a vector parallel to $x^{(k)}$. The work [69] shows that a generic tensor has finitely many singular vector tuples and provides a recipe for obtaining their number.

As with eigenvectors, the singular vector tuples of a tensor T are the fixed points of the tensor power iteration

$$(x^{(1)}, \ldots, x^{(d)}) \mapsto \left(\frac{T(\cdot, x^{(2)} \ldots, x^{(d)})}{||T(\cdot, x^{(2)} \ldots, x^{(d)})||}, \ldots, \frac{T(x^{(1)} \ldots, x^{(d-1)}, \cdot)}{||T(x^{(1)} \ldots, x^{(d-1)}, \cdot)||} \right).$$

The set of singular vector tuples is also equal to the set of critical points of the optimization problem

$$\begin{aligned} &\text{maximize} \quad T(x^{(1)}, \ldots, x^{(d)}) \\ &\text{such that } ||x^{(1)}|| = \cdots = ||x^{(d)}|| = 1, \end{aligned}$$

where $T(x^{(1)}, \ldots, x(d))$ is the contraction of T along all of its dimensions with $x^{(1)}, \ldots, x^{(d)}$. If the tuple $(x^{(1)}, \ldots, x^{(d)})$ is a global maximizer of this optimization problem, then the tensor $x^{(1)} \otimes \cdots \otimes x^{(d)}$ is the best rank-one approximation of T [112].

1.2.5 Spectral theory of orthogonally decomposable tensors

Finding the eigenvectors or singular vector tuples of a tensor is in general very hard. However, this is not the case for odeco tensors.

In the paper titled *Orthogonal decomposition of symmetric tensors* I give an explicit description of the eigenvectors of symmetric odeco tensors. Their number equals the number of eigenvectors of a generic tensor of the same format, and they can be expressed as specific linear combinations of the vectors in the decomposition of the given odeco tensor. Section 3.1 is dedicated to this work.

Example 1.2.4. *Let $T \in S^d(\mathbb{R}^n)$ be a symmetric odeco tensor with decomposition $T = \sum_{i=1}^n \lambda_1 v_i^{\otimes d}$ where $v_1, \ldots, v_n$ are orthonormal. Then,*

$$T \cdot v_k^{d-1} = \sum_{i=1}^n \lambda_i \langle v_i, v_k \rangle v_i = \lambda_k v_k.$$

So, each of the vectors $v_1, \ldots, v_n$ is an eigenvector of T. If $d \geq 3$, T has many more eigenvectors, explicitly described in Theorem 3.1.8.

In joint work with Anna Seigal titled *Singular vector tuples of orthogonally decomposable tensors* we give a formula for the singular vector tuples of an ordinary odeco tensor. In this case, the variety of singular vector tuples is not zero-dimensional, contrary to the generic case. We give several examples of this phenomenon and illustrate how the singular vector tuples of a generic tensor degenerate to those of an odeco tensor. Section 3.2 is dedicated to this work.

1.3 Super-Resolution Imaging

Super-resolution imaging, the study of enhancing low-resolution blurred images, has direct applications in numerous fields, including fluorescence microscopy and astronomy. It is one of the first ingredients for studying the macro and micro worlds, namely, it helps us observe them as accurately as possible. Given a low-resolution blurred image of several point-sources, super-resolution aims to find the true locations of the point sources and the intensities of light at each of them.

Mathematically, the *unknowns* are the locations of the point sources, $t_1, \ldots, t_M \in \mathbb{R}^d$, and the intensities at each of them, $c_1, \ldots, c_M \in \mathbb{R}$. It is very convenient to encode these unknown parameters in a discrete measure

$$\mu^* := \sum_{j=1}^{M} c_j \delta_{t_j},$$

where δ_{t_j} is the Dirac delta function centered at t_j. The super-resolution imaging problem is to *recover* μ^* from *observations*

$$x_i = \int f_i(t) d\mu^*(t), \quad i = 1, \ldots, N,$$

where the functions $f_1, \ldots, f_N : \mathbb{R}^d \to \mathbb{R}$ are known.

In this thesis we discuss the following two special cases of this setup.

- The functions $f_1, \ldots, f_N$ are translates of a given function $\psi : \mathbb{R}^d \to \mathbb{R}$, i.e. $f_i(t) = \psi(s_i - t)$ for some $s_1, \ldots, s_N \in \mathbb{R}^d$.
- The functions $f_1, \ldots, f_N$ are various monomials in the entries of $t = (t^{(1)}, \ldots, t^{(d)}) \in \mathbb{R}^d$, i.e. they have the form $f_i(t) = \left(t^{(1)}\right)^{k_{i1}} \cdots \left(t^{(d)}\right)^{k_{id}}$.

The former case is further developed in Subsection 1.3.1 and then in Chapter 5. The latter case is only discussed here in the introduction in Subsection 1.3.2 to serve as a connection with the tensor decomposition problem.

1.3.1 Super-resolution without Separation

We begin by considering the former case. The observations here have the form

$$x_i = \int \psi(s_i - t) d(\mu^*(t)) = \sum_{j=1}^{M} c_i \psi(s_i - t_j).$$

In other words, we observe the value of the function

$$x(s) = \sum_{j=1}^{M} c_i \psi(s - t_j), \tag{1.3.1}$$

at $s = s_1, \ldots, s_N$. Imagine that the unknown locations of the stars in a distant galaxy are $t_1, \ldots, t_M$ and the intensity of light at each of them are $c_1, \ldots, c_M$. Then, the observed signal $x(s)$ has introduced a blur, give by the *pointspread function* ψ, centered at $t_1, \ldots, t_M$.

In fact, for each imaging device (microscope, telescope, camera, even the human eye) there exists a point spread function $\psi : \mathbb{R}^n \to \mathbb{R}$ which blurs each pointsource of light. This phenomenon is due to the diffraction of light and the optics in the imaging device. The locations $s_1, \ldots, s_N$ at which we observe the signal $x(s)$ correspond to the locations of the pixels of an image.

In Figures 1.4 and 1.5 we show an example of a signal in $\mathbb{R}^1$ with 4 unknown pointsources and the Gaussian pointspread function $\psi(t) = e^{-t^2}$.

What we know

- The pointspread function ψ
- Finitely many signal observations:

$$\{x(s_i) | i = 1, \ldots, N\}$$

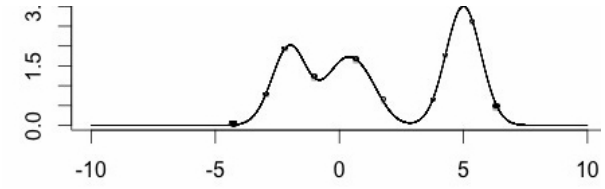


Figure 1.4: We observe the image $x(s)$ at the dotted locations.

What we want to find

- The locations of the point sources: $t_1, \ldots, t_M$
- The intensities: $c_1, \ldots, c_M$.

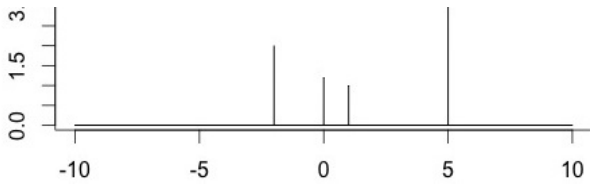


Figure 1.5: A graph of the unknown measure μ^* which encodes $t_1, \ldots, t_M$ and $c_1, \ldots, c_M$.

We now propose how to recover the measure μ^* from the observations $x(s_1), \ldots, x(s_M)$. Let $w(t) = \frac{1}{N} \sum_{i=1}^{N} \psi(s_i - t)$. Consider the optimization problem

$$\begin{aligned} \underset{\mu \geq 0}{\text{minimize}} \quad & \int w(t) \mu(dt) \\ \text{subject to} \quad & x(s_i) = \int \psi(s_i - t) d\mu(t), \quad i = 1, \ldots, N. \end{aligned} \tag{1.3.2}$$

We show that one can recover μ^* (and hence $t_1, \ldots, t_M, c_1, \ldots, c_M$) by solving this problem. In joint work with Geoffrey Schiebinger and Benjamin Recht titled *Superresolution without separation* we prove that the optimization problem (5.1.3) recovers the correct source locations, in the case of a one-dimensional signal and under some determinantal conditions on the point spread function ψ [137].

Much of the mathematical analysis of super-resolution has relied heavily on the assumption that the sources of light are separated by a minimum amount. Our results do not require a separation condition. Instead, we give determinantal conditions on the pointspread func-

tion ψ and we show that the Gaussian pointspread function $\psi(t) = e^{-t^2}$ satisfies these conditions. Section 5 is dedicated to this work.

1.3.2 Super-resolution Imaging as Tensor Decomposition

We conclude the introduction by showing the tight connection between super-resolution imaging and tensor decompositions. We assume that the unknown pointsources $t_1, \ldots, t_M \in \mathbb{C}^d$ and intensities $c_1, \ldots, c_M \in \mathbb{C}$ are encoded in the measure

$$\mu^* = \sum_{i=1}^{M} c_i \delta_{t_i}.$$

We observe the *moments*

$$x_i = x(\mathbf{k}_i) = \int \left(t^{(1)}\right)^{k_{i1}} \cdots \left(t^{(d)}\right)^{k_{id}} d\mu(x), \quad \mathbf{k}_i \in \mathbb{Z}^d, \|\mathbf{k}_i\|_\infty \leq n,$$

for some natural number n. There are $N = (n+1)^d$ possible observations since $\mathbf{k}_i \in \mathbb{Z}^d$ and $\|\mathbf{k}_i\|_\infty \leq n$. Suppose that we make all of these observations, and we wish to recover μ^*.
The *multivariate* version of *Prony's method* [107] accomplishes this task by considering the kernel of the Toeplitz matrix

$$(x(\mathbf{k}_i) - x(\mathbf{k}_j))_{1 \leq i,j \leq N} \in \mathbb{C}^{N \times N}.$$

This kernel provides a set of polynomial equations in the variables $t^{(1)}, \ldots, t^{(d)}$ and the solutions are exactly the unknown pointsource locations $t_1, \ldots, t_M$. Once we have recovered them, finding the point source intensities $c_1, \ldots, c_M$ can be done via solving a system of linear equations.

Now, consider the tensor $T \in (S^{2n}(\mathbb{C}^2))^{\otimes d}$ with the following decomposition

$$T = \sum_{j=1}^{M} c_j \bigotimes_{i=1}^{d} \begin{pmatrix} 1 \\ t_j^{(i)} \end{pmatrix}^{\otimes 2n}.$$

An element of the space $(S^{2n}(\mathbb{C}^2))^{\otimes d}$ has entries indexed by $((k_1, 2n-k_1), \ldots, (k_d, 2n-k_d))$, meaning that in the i-th component $S^{2n}(\mathbb{C}^2)$ there are k_1 ones and $2n-k_1$ twos. If we expand the above expression for T, we see that the entry of T at position $((k_1, 2n-k_1), \ldots, (k_d, 2n-k_d))$ equals

$$T_{((k_1,2n-k_1),\ldots,(k_d,2n-k_d))} = \binom{2n}{n+k_1} \cdots \binom{2n}{n+k_d} x(2\mathbf{n} - \mathbf{k}),$$

where $\mathbf{n} = (n, \ldots, n) \in \mathbb{Z}^d$. This means that observing $x(\mathbf{k}_i)$ for $\mathbf{k}_i \in \mathbb{Z}^d$ and $\|\mathbf{k}_i\|_\infty \leq n$ is equivalent to observing the tensor T, and finding the unknowns $t_1, \ldots, t_M, c_1, \ldots, c_M$ is equivalent to finding a decomposition of T in $(S^{2n}(\mathbb{C}^2))^{\otimes d}$. It would be interesting to study such decompositions in view of the work [107].

Chapter 2

Matrices and Positivity

In this chapter we turn our attention to matrix shaped data. In Section 2.1 we study nonnegative decompositions, and in Section 2.2 we study positive semidefinite matrix decompositions.

2.1 Nonnegative Rank

Mixtures of r independent distributions for two discrete random variables can be represented by matrices of nonnegative rank r. Likelihood inference for the model of such joint distributions leads to problems in real algebraic geometry that are addressed here for the first time. We characterize the set of fixed points of the Expectation Maximization algorithm, and we study the boundary of the space of matrices with nonnegative rank at most 3. Both of these sets correspond to algebraic varieties with many irreducible components. This section represents joint work with Kaie Kubjas and Bernd Sturmfels titled *Fixed Points of the EM Algorithm and Nonnegative Rank Boundaries* [105].

2.1.1 Introduction

The rth mixture model $\mathcal{M}$ of two discrete random variables X and Y expresses the conditional independence statement $X \perp\!\!\!\perp Y \mid Z$, where Z is a hidden (or latent) random variable with r states. Assuming that X and Y have m and n states respectively, their joint distribution is written as an $m \times n$-matrix of nonnegative rank $\leq r$ whose entries sum to 1. This mixture model is also known as the *naive Bayes model*. Its graphical representation is shown in Figure 2.1.

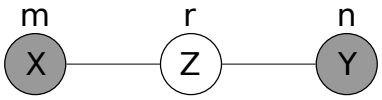


Figure 2.1: Graphical model on two observed variables and one hidden variable

A collection of i.i.d. samples from a joint distribution is recorded in a nonnegative matrix

$$U \;=\; \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{bmatrix}.$$

Here, u_{ij} is the number of observations in the sample with $X = i$ and $Y = j$. The sample size is $u_{++} = \sum_{i,j} u_{ij}$. It is standard practice to fit the model to the data U using the Expectation Maximization (EM) algorithm. Here, fitting means computing the maximum likelihood estimate (MLE). However, it has been pointed out in the literature that EM has several issues (see the next paragraph for details) and one has to be careful when using it. Our goal is to better understand this algorithm by studying its mathematical properties in some detail.

One of the main issues of Expectation Maximization is that it does not provide a certificate for having found the global optimum. The geometry of the algorithm has been a topic for debate among statisticians since the seminal paper of Dempster, Laird and Rubin [49]. Murray [118] responded with a warning for practitioners to be aware of the existence of multiple stationary points. Beale [17] also brought this up, and Fienberg [67] referred to the possibility that the MLE lies on the boundary of the parameter space. A recent discussion of this issue was presented by Zwiernik and Smith [164, §3] in their analysis of inferential problems arising from the semialgebraic geometry of a latent class model. The fact that our model fails to be identifiable was highlighted by Fienberg *et* al. in [68, §4.2.3]. This poses additional difficulties, and it forces us to distinguish between the boundary of the parameter space and the boundary of the model. The image of the former contains the latter.

The EM algorithm aims to maximize the log-likelihood function of the model $\mathcal{M}$. In doing so, it approximates the data matrix U with a product of nonnegative matrices $A \cdot B$ where A has r columns and B has r rows. In Subsection 4.2.2 we review the EM algorithm in our context. Here, it is essentially equivalent to the widely used method of Lee and Seung [111] for *nonnegative matrix factorization*. The nonnegative rank of matrices has been studied from a broad range of perspectives, including computational geometry [1, 43], topology [116], contingency tables [21, 68], complexity theory [115, 156], and convex optimization [64]. We here present the approach from algebraic statistics [54, 122].

Maximum likelihood estimation for the model $\mathcal{M}$ is a non-convex optimization problem. Any algorithm that promises to compute the MLE $\widehat{P}$ will face the following fundamental dichotomy. The optimal matrix $\widehat{P}$ either lies in the relative interior of $\mathcal{M}$ or it lies in the model boundary $\partial\mathcal{M}$.

If $\widehat{P}$ lies in the relative interior of $\mathcal{M}$ then the situation is nice. In this case, $\widehat{P}$ is a critical point for the likelihood function on the manifold of rank r matrices. There are methods by Hauenstein *et* al. [85] for finding the MLE with certificate. The ML degree, which they compute, bounds the number of critical points, and hence all candidates for the global maximizer $\widehat{P}$. However, things are more difficult when $\widehat{P}$ lies in the boundary $\partial\mathcal{M}$.

rank \ size	4×4	5×5	6×6	7×7	8×8
3	4.4%	23%	49%	62%	85%
4		7%	37%	71%	95%
5			10%	55%	96%
6				20%	75%
7					24%

Table 2.1: Percentage of data matrices whose maximum likelihood estimate $\widehat{P}$ lies in the boundary $\partial\mathcal{M}$

In that case, $\widehat{P}$ is generally not a critical point for the likelihood function in the manifold of rank r matrices, and none of the results on ML degrees in [54, 68, 82, 85, 93] are applicable. The present section is the first to address the question of how $\widehat{P}$ varies when it occurs in the boundary $\partial\mathcal{M}$. Table 2.1 underscores the significance of our approach. As the matrix size grows, the boundary case is much more likely to happen for randomly chosen input U. The details for choosing U and the simulation study that generated Table 2.1 will be described in Example 2.1.6.

We now summarize the contents of this section. Subsection 2.1.2 furnishes an introduction to the geometry of the mixture model $\mathcal{M}$ from Figure 2.1. We define the *topological boundary* of $\mathcal{M}$ and the *algebraic boundary* of $\mathcal{M}$, and we explain how these two notions of boundary differ. Concrete numerical examples for 4×4-matrices of rank 3 demonstrate how $\widehat{P}$ behaves as the data U vary.

In Subsection 2.1.3 we review the EM algorithm for the model $\mathcal{M}$, and we identify its fixed points in the parameter space. The main result is the characterization of the set of fixed points in Theorem 2.1.7.

In Subsection 2.1.4 we identify $\mathcal{M}$ with the set of matrices of nonnegative rank at most 3. Theorem 2.1.9 gives a quantifier-free formula for this semialgebraic set. The importance of finding such a formula was already stressed in the articles [4, 5]. The resulting membership test for $\mathcal{M}$ is very fast and can be applied to matrices that contain parameters. The proof of Theorem 2.1.9 is based on the familiar characterization of nonnegative rank in terms of nested polytopes [1, 43, 156], and, in particular, on work of Mond *et al.* [116] on the structure of critical configurations in the plane (shown in Figure 2.5).

In Subsection 2.1.5 we return to Expectation Maximization, and we study the system of equations that characterize the EM fixed points. Proposition 2.1.15 characterizes its solutions in the interior of $\mathcal{M}$. Even in the smallest interesting case, $m = n = 4$ and $r = 3$, the variety of all EM fixed points has a huge number of irreducible components, to be determined and interpreted in Theorem 2.1.19.

The most interesting among these are the 288 components that delineate the topological boundary $\partial\mathcal{M}$ inside the simplex Δ_{15}. These are discussed in Examples 2.1.21 and 2.1.24. Explicit matrices that lie on these components are featured in (2.1.24) and in Examples

2.1.1, 2.1.2 and 2.1.4. In Proposition 2.1.25 we resolve a problem left open in [85, 93] concerning the ML degree arising from $\partial\mathcal{M}$. The main result in Subsection 2.1.6 is Theorem 2.1.23 which characterizes the algebraic boundary of $m \times n$-matrices of nonnegative rank 3. The commutative algebra of the irreducible components in that boundary is the content of Theorem 2.1.26. Corollary 2.1.28 furnishes a quantifier-free semialgebraic formula for $\partial\mathcal{M}$.

The proofs of all lemmas, propositions and corollaries appear in Appendix 2.1.7.1. A review of basic concepts in algebraic geometry is given in Appendix 2.1.7.2. This will help the reader understand the technicalities of our main results. Supplementary materials and software are posted at the website

`http://math.berkeley.edu/~bernd/EM/boundaries.html`

Our readers will find code in `R`, `Macaulay2`, and `Magma` for various sampling experiments, prime decompositions, semialgebraic formulas, and likelihood equations discussed in this section.

The methods presented here are not limited to the matrix model $\mathcal{M}$, but are applicable to a wide range of statistical models for discrete data, especially those used in computational biology [122]. Such models include phylogenetic models [3, 4] and Hidden Markov models [47]. The most immediate generalization is to the rth mixture model of several random variables. It consists of all distributions corresponding to tensors of nonnegative rank at most r. In other words, we replace $m \times n$-matrices by tensors of arbitrary format. The geometry of the case $r = 2$ was studied in depth by Allman *et al.* [5]. For each of these models, there is a natural EM algorithm, with an enormous number of stationary points. The model itself is a complicated semialgebraic set, and the MLE typically occurs on the boundary of that set. For binary tree models this was shown in [164, §3].

This section introduces tools needed to gain a complete understanding of these EM fixed points and model boundaries. We here study them for the graphical model in Figure 2.1. Already in this very simple case, we discovered patterns that are surprisingly rich. Thus, the present work serves as a blueprint for future research in real algebraic geometry that underlies statistical inference.

2.1.2 Model Geometry

We begin with a geometric introduction of the likelihood inference problem to be studied. Let Δ_{mn-1} denote the probability simplex of nonnegative $m \times n$-matrices $P = [p_{ij}]$ with $p_{++} = 1$. Our model $\mathcal{M}$ is the subset of Δ_{mn-1} consisting of all matrices of the form

$$P \;=\; A \cdot \Lambda \cdot B, \tag{2.1.1}$$

where A is a nonnegative $m \times r$-matrix whose columns sum to 1, Λ is a nonnegative $r \times r$ diagonal matrix whose entries sum to 1, and B is a nonnegative $r \times n$-matrix whose rows sum to 1. The triple of parameters (A, Λ, B) represents conditional probabilities for the graphical

model in Figure 2.1. In particular, the kth column of A is the conditional probability distribution of X given that $Z = k$, the kth row of B is the conditional probability distribution given that $Z = k$, and the diagonal of Λ is the probability distribution of Z. The parameter space in which A, Λ, B lie is the convex polytope $\Theta = (\Delta_{m-1})^r \times \Delta_{r-1} \times (\Delta_{n-1})^r$. Our model $\mathcal{M}$ is the image of the trilinear map

$$\phi : \Theta \to \Delta_{mn-1}, \quad (A, \Lambda, B) \mapsto P. \tag{2.1.2}$$

We seek to learn the model parameters (A, Λ, B) by maximizing the likelihood function

$$\binom{u_{++}}{u} \cdot \prod_{i=1}^{m} \prod_{j=1}^{n} p_{ij}^{u_{ij}} \tag{2.1.3}$$

over $\mathcal{M}$. This is equivalent to maximizing the log-likelihood function

$$\ell_U = \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij} \cdot \log\Big(\sum_{k=1}^{r} a_{ik} \lambda_k b_{kj} \Big) \tag{2.1.4}$$

over $\mathcal{M}$. One issue that comes up immediately is that the model parameters are not identifiable:

$$\dim(\Theta) = r(m+n) - r - 1 \quad \text{but} \quad \dim(\mathcal{M}) = r(m+n) - r^2 - 1. \tag{2.1.5}$$

The first expression is the sum of the dimensions of the simplices in the product that defines the parameter space Θ. The second one counts the degrees of freedom in a rank r matrix of format $m \times n$. The typical fiber, i.e. the preimage of a point in the image of (2.1.2), is a semialgebraic set of dimension $r^2 - r$. This is the *space of explanations* whose topology was studied by Mond *et al.* in [116]. Likelihood inference cannot distinguish among points in each fiber, so it is preferable to regard MLE not as an unconstrained optimization problem in Θ but as a constrained optimization problem in $\mathcal{M}$. The aim of this section is to determine its constraints.

Let $\mathcal{V}$ denote the set of real $m \times n$-matrices P of rank $\leq r$ satisfying $p_{++} = 1$. This set is a variety because it is given by the vanishing of a set of polynomials, namely, the $(r+1) \times (r+1)$ minors of the matrix P plus the linear constraint $p_{++} = 1$. A point $P \in \mathcal{M}$ is an *interior point* of $\mathcal{M}$ if there is an open ball $U \subset \Delta_{mn-1}$ that contains P and satisfies $U \cap \mathcal{V} = U \cap \mathcal{M}$. We call $P \in \mathcal{M}$ a *boundary point* of $\mathcal{M}$ if it is not an interior point. The set of all such points is denoted by $\partial\mathcal{M}$ and called the *topological boundary* of $\mathcal{M}$. In other words, $\partial\mathcal{M}$ is the boundary of $\mathcal{M}$ inside $\mathcal{V}$. The variety $\mathcal{V}$ is the Zariski closure of the set $\mathcal{M}$; see Appendix 2.1.7.2. In other words, the set of polynomials that vanish on $\mathcal{M}$ is exactly the same as the set of polynomials that vanish on $\mathcal{V}$. Our model $\mathcal{M}$ is a full-dimensional subset of the variety $\mathcal{V}$ and is given by a set of polynomial inequalities inside $\mathcal{V}$.

Fix U, r, and $P \in \mathcal{M}$ as above. A matrix P is a non-singular point on $\mathcal{V}$ if and only if the rank of P is exactly r. In this case, its tangent space $\mathrm{T}_P(\mathcal{V})$ has dimension $r(m+n) - r^2 - 1$,

which, as expected, equals $\dim(\mathcal{M})$. We call P a *critical point* of the log-likelihood function ℓ_U if $P \in \mathcal{M}$, P is a nonsingular point for $\mathcal{V}$, i.e. $\text{rank}(P) = r$, and the gradient of ℓ_U is orthogonal to the tangent space $\mathrm{T}_P(\mathcal{V})$. Thus, the critical points are the nonnegative real solutions of the various *likelihood equations* derived in [54, 85, 122, 163] to address the MLE problem for $\mathcal{M}$. In other words, the critical points are the solutions obtained by using the Lagrange multipliers method for maximizing the likelihood function over the set $\mathcal{V}$. In the language of algebraic statistics, the critical points are those points in $\mathcal{M}$ that are accounted for by the *ML degree* of the variety $\mathcal{V}$.

Table 2.1 shows that the global maximum $\widehat{P}$ of ℓ_U is often a non-critical point. This means that the MLE lies on the topological boundary $\partial\mathcal{M}$. The ML degree of the variety $\mathcal{V}$ is irrelevant for assessing the algebraic complexity of such $\widehat{P}$. Instead, we need the ML degree of the boundary, as given in Proposition 2.1.25, as well as the ML degrees for the lower-dimensional boundary strata.

The following example illustrates the concepts we have introduced so far and what they mean.

Example 2.1.1. *Fix $m = n = 4$ and $r = 3$. For any integers $a \geq b \geq 0$, consider the data matrix*

$$U_{a,b} \quad = \quad \begin{bmatrix} a & a & b & b \\ a & b & a & b \\ b & a & b & a \\ b & b & a & a \end{bmatrix}. \tag{2.1.6}$$

Note that $\text{rank}(U_{a,b}) \leq 3$. For $a = 1$ and $b = 0$, this is the standard example [43] of a nonnegative matrix whose nonnegative rank exceeds its rank. Thus, $\frac{1}{8}U_{1,0}$ is a probability distribution in $\mathcal{V}\backslash\mathcal{M}$. Within the 2-parameter family (2.1.6), the topological boundary $\partial\mathcal{M}$ is given by the linear equation $b = (\sqrt{2}-1)a$. This follows from the computations in [21, §5] and [116, §5]. We conclude that

$$\tfrac{1}{8(a+b)}U_{a,b} \text{ lies in } \mathcal{V}\backslash\mathcal{M} \quad \text{if and only if} \quad b < (\sqrt{2}-1)a. \tag{2.1.7}$$

For integers $a > b \geq 0$ satisfying (2.1.7), the likelihood function (2.1.3) for $U_{a,b}$ has precisely eight global maxima on our model $\mathcal{M}$. These are the following matrices, each divided by $8(a+b)$:

$$\begin{bmatrix} a & a & b & b \\ v & w & t & u \\ w & v & u & t \\ s & s & r & r \end{bmatrix}, \begin{bmatrix} v & t & w & u \\ a & b & a & b \\ s & r & s & r \\ w & u & v & t \end{bmatrix}, \begin{bmatrix} t & v & u & w \\ r & s & r & s \\ b & a & b & a \\ u & w & t & v \end{bmatrix}, \begin{bmatrix} r & r & s & s \\ t & u & v & w \\ u & t & w & v \\ b & b & a & a \end{bmatrix},$$

$$\begin{bmatrix} a & v & w & s \\ a & w & v & s \\ b & t & u & r \\ b & u & t & r \end{bmatrix}, \begin{bmatrix} v & a & s & w \\ t & b & r & u \\ w & a & s & v \\ u & b & r & t \end{bmatrix}, \begin{bmatrix} t & r & b & u \\ v & s & a & w \\ u & r & b & t \\ w & s & a & v \end{bmatrix}, \begin{bmatrix} r & t & u & b \\ r & u & t & b \\ s & v & w & a \\ s & w & v & a \end{bmatrix}.$$

This claim can be verified by exact symbolic computation, or by validated numerics as in the proof of [85, Theorem 4.4]. Here t is the unique simple real root of the cubic equation

$$\begin{aligned}(6a^3+16a^2b+14ab^2+4b^3)t^3 \;&-\; (20a^4+44a^3b+8ab^3+32a^2b^2)t^2 \\ +\;\; (22a^5+43a^4b+30a^3b^2+7a^2b^3)t \;&-\; (8a^6+16a^5b+10a^4b^2+2a^3b^3) \;\;=\;\; 0.\end{aligned}$$

To fill in the other entries of these nonnegative rank 3 matrices, we use the rational formulas

$$s \,=\, \frac{(a+b)t-a^2}{a},\;\; u \,=\, \frac{tb}{a},\;\; w \,=\, -\frac{t(3a^2+5ab+2b^2)t-4a^3-5a^2b-2ab^2}{2a^3+a^2b},$$

$$r = \frac{2a^2+ab-(a+b)t}{a},\;\; v = \frac{(3a^2{+}5ab{+}2b^2)t^2-(6a^3{+}8a^2b{+}3ab^2)t+6a^3b{+}2a^2b^2{+}4a^4}{2a^3+a^2b}.$$

These formulas represent an exact algebraic solution to the MLE problem in this case. They describe the multivalued map $(a,b) \mapsto \widehat{P}_{a,b}$ from the data to the eight maximum likelihood estimates. This allows us to understand exactly how these solutions behave as the matrix entries a and b vary.

The key point is that the eight global maxima lie in the model boundary $\partial\mathcal{M}$. They are not critical points of ℓ_U on the rank 3 variety $\mathcal{V}$. They will not be found by the methods in [85, 122, 163]. Instead, we used results about the algebraic boundary in Subsection 2.1.5 to derive the eight solutions.

We note that this example can be seen as an extension of [85, Theorem 4.4], which offers a similar parametric analysis for the data set of the "100 Swiss Francs Problem" studied in [68, 163]. ♢

We now introduce the concept of algebraic boundary. Recall that the topological boundary $\partial\mathcal{M}$ of the model $\mathcal{M}$ is a semialgebraic subset inside the probability simplex Δ_{mn-1}. Its dimension is

$$\dim(\partial\mathcal{M}) \;=\; \dim(\mathcal{M})-1 \;=\; rm+rn-r^2-2.$$

Any quantifier-free semialgebraic description of $\partial\mathcal{M}$ will be a complicated Boolean combination of polynomial equations and polynomial inequalities. This can be seen for $r=3$ in Corollary 2.1.28.

To simplify the situation, it is advantageous to relax the inequalities and keep only the equations. This replaces the topological boundary of $\mathcal{M}$ by a much simpler object, namely the algebraic boundary of $\mathcal{M}$. To be precise, we define the *algebraic boundary* to be the Zariski closure $\overline{\partial\mathcal{M}}$ of the topological boundary $\partial\mathcal{M}$. Thus $\overline{\partial\mathcal{M}}$ is a subvariety of codimension 1 inside the variety $\mathcal{V} \subset \mathbb{P}^{mn-1}$. Theorem 2.1.23 will show us that $\overline{\partial\mathcal{M}}$ can have many irreducible components.

The following two-dimensional family of matrices illustrates the results to be achieved in this section. These enable us to discriminate between the topological boundary $\partial\mathcal{M}$ and the algebraic boundary $\overline{\partial\mathcal{M}}$, and to understand how these boundaries sit inside the variety $\mathcal{V}$.

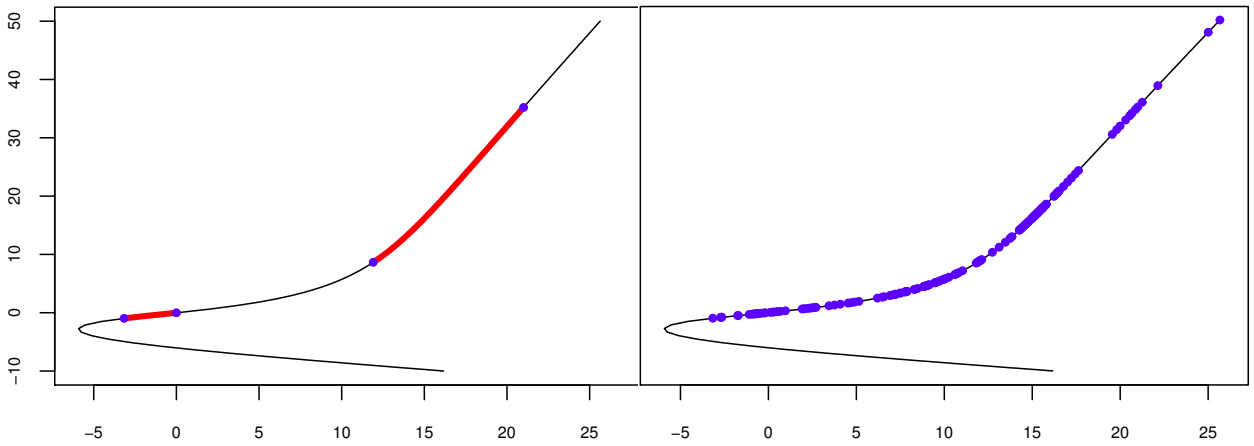


Figure 2.2: In a two-dimensional family of 4×4-matrices, the matrices of rank 3 form a quartic curve. The mixture model, shown in red, has two connected components. Its topological boundary consists of four points (on the left). The algebraic boundary includes many more points (on the right). Currently, there is no known way to obtain the four points on the topological boundary (in the left picture) without first considering all points on the algebraic boundary (in the right picture).

Example 2.1.2. *Consider the following 2-parameter family of 4×4-matrices:*

$$P(x,y) \;=\; \begin{bmatrix} 51 & 9 & 64 & 9 \\ 27 & 63 & 8 & 8 \\ 3 & 34 & 40 & 31 \\ 30 & 25 & 80 & 35 \end{bmatrix} + x \cdot \begin{bmatrix} 1 & 1 & 3 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} + y \cdot \begin{bmatrix} 5 & 4 & 1 & 1 \\ 5 & 1 & 5 & 1 \\ 1 & 5 & 1 & 5 \\ 1 & 1 & 5 & 5 \end{bmatrix}.$$

This was chosen so that $P(0,0)$ lies in a unique component of the topological boundary $\partial\mathcal{M}$. The equation $det(P(x,y)) = 0$ defines a plane curve $\mathcal{C}$ of degree 4. This is the thin black curve shown in Figure 2.2. In our family, this quartic curve $\mathcal{C}$ represents the Zariski closure $\mathcal{V}$ of the model $\mathcal{M}$.

The algebraic boundary $\overline{\partial\mathcal{M}}$ is the variety described in Example 2.1.24. The quartic curve $\mathcal{C}$ meets $\overline{\partial\mathcal{M}}$ in 1618 real points (x,y). Of these 1618 points, precisely 188 satisfy the constraint $P(x,y) \geq 0$. These 188 points are the landmarks for our analysis. They are shown in blue on the right in Figure 2.2. In addition, we mark the unique point where the curve $\mathcal{C}$ intersects the boundary polygon defined by $P(x,y) \geq 0$. This is the leftmost point, defined by $\{det(P(x,y)) = x + 5y + 8 = 0\}$. It equals

$$(-3.161429, -0.967714). \tag{2.1.8}$$

We examined the 187 arcs on $\mathcal{C}$ between consecutive points of $\overline{\partial\mathcal{M}}$ as well as the two arcs at the ends. For each arc we checked whether it lies in $\mathcal{M}$. This was done by a combination of the EM algorithm in Subsection 2.1.3 and Theorem 2.1.9. Precisely 96 of the 189 arcs were

found to lie in $\mathcal{M}$. These form two connected components on the curve $\mathcal{C}$, namely 19 *arcs between (2.1.8) and (0,0), and*

$$76 \text{ arcs between } (11.905773, 8.642630) \text{ and } (21.001324, 35.202110). \tag{2.1.9}$$

These four points represent the topological boundary $\partial\mathcal{M}$. We conclude that, in the 2*-dimensional family $P(x, y)$, the model $\mathcal{M}$ is the union of the two red arcs shown on the left in Figure 2.2.*

Our theory of EM fixed points distinguishes between the (relatively open) red arcs and their blue boundary points. For the MLE problem, the red points are critical while the blue points are not critical. By Table 2.1, the MLE is more likely to be blue than red, for larger values of m and n. ♢

This example demonstrates that the algebraic methods of Subsections 2.1.4, 2.1.5 and 2.1.6 are indispensable when one desires a reliable analysis of model geometries, such as that illustrated in Figure 2.2. To apply a method for finding the critical points of a function, e.g. Lagrange multipliers, the domain of the function needs to be given by equality constraints only. But using only these constraints, one cannot detect the maxima lying on the topological boundary. For finding the critical points of the likelihood function on the topological boundary by using the same methods, one needs to relax the inequality constraints and consider only the equations defining the topological boundary. Therefore, one needs to find the critical points on the algebraic boundary $\overline{\partial\mathcal{M}}$ of the model.

2.1.3 Fixed Points of Expectation Maximization

The EM algorithm is an iterative method for finding local maxima of the likelihood function (2.1.3). It can be viewed as a discrete dynamical system on the polytope $\Theta = (\Delta_{m-1})^r \times \Delta_{r-1} \times (\Delta_{n-1})^r$. We here present the version in [122, §1.3].

Algorithm 1 Function EM(U, r)

Select random $a_1, a_2, \ldots, a_r \in \Delta_{m-1}$, random $\lambda \in \Delta_{r-1}$, and random $b_1, b_2, \ldots, b_r \in \Delta_{n-1}$.
Run the following steps until the entries of the $m\times n$-matrix P converge.
***E-step**: Estimate the $m\times r\times n$-table that represents this expected hidden data:*
  Set $v_{ikj} := \frac{a_{ik}\lambda_k b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij}$ for $i = 1, \ldots, m$, $k = 1, \ldots, r$ and $j = 1, \ldots, n$.
***M-step**: Maximize the likelihood function of the model* ●—●—● for the hidden data:
  Set $\lambda_k := \sum_{i=1}^{m}\sum_{j=1}^{n} v_{ikj}/u_{++}$ for $k = 1, \ldots, r$.
  Set $a_{ik} := (\sum_{j=1}^{n} v_{ikj})/(u_{++}\lambda_k)$ for $k = 1, \ldots, r$ and $i = 1, \ldots, m$.
  Set $b_{kj} := (\sum_{i=1}^{m} v_{ikj})/(u_{++}\lambda_k)$ for $k = 1, \ldots, r$ and $j = 1, \ldots, n$.
***Update** the estimate of the joint distribution for our mixture model* ●—○—●*:*
  Set $p_{ij} := \sum_{k=1}^{r} a_{ik}\lambda_k b_{kj}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.
Return P.

The alternating sequence of E-steps and M-steps defines trajectories in the parameter polytope Θ. The log-likelihood function (2.1.4) is non-decreasing along each trajectory (cf. [122, Theorem 1.15]). In fact, the value can stay the same only at a fixed point of the EM algorithm. See Dempster *et al.* [49] for the general version of EM and its increasing behavior and convergence.

Definition 2.1.3. *An* EM fixed point *for a given table U is any point (A, Λ, B) in the polytope $\Theta = (\Delta_{m-1})^r \times \Delta_{r-1} \times (\Delta_{n-1})^r$ to which the EM algorithm can converge if it is applied to (U, r).*

Every global maximum $\widehat{P}$ of ℓ_U is among the EM fixed points. One hopes that $\widehat{P}$ has a large basin of attraction, and that the initial parameter choice (A, Λ, B) gives a trajectory that converges to $\widehat{P}$. However, this need not be the case, since the EM dynamics on Θ has many fixed points other than $\widehat{P}$. Our aim is to understand all of these.

Example 2.1.4. *The following data matrix is obtained by setting $a = 1, b = 0$ in Example 2.1.1:*

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

Among the EM fixed points for this choice of U with $r = 3$ we find the probability distributions

$$P_1 = \frac{1}{24}\begin{bmatrix} 3 & 3 & 0 & 0 \\ 2 & 0 & 4 & 0 \\ 0 & 2 & 0 & 4 \\ 1 & 1 & 2 & 2 \end{bmatrix}, \quad P_2 = \frac{1}{16}\begin{bmatrix} 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}, \quad \text{and} \quad P_3 = \frac{1}{48}\begin{bmatrix} 4 & 8 & 0 & 0 \\ 3 & 0 & 4 & 5 \\ 5 & 4 & 0 & 3 \\ 0 & 0 & 8 & 4 \end{bmatrix},$$

and their orbits under the symmetry group of U. For instance, the orbit of P_1 is obtained by setting $s = \frac{1}{3}, r = \frac{2}{3}, v = \frac{2}{3}, t = \frac{4}{3}, w = u = 0$ in the eight matrices in Example 2.1.1. Over 98% of our runs with random starting points in Θ converged to one of these eight global maximizers of ℓ_U. Matrices in the orbits of P_2 resp. P_3 were approached only rarely (less than 2%) by the EM algorithm. $\diamondsuit$.

Lemma 2.1.5. *The following are equivalent for a point (A, Λ, B) in the parameter polytope Θ:*

(1) The point (A, Λ, B) is an EM fixed point.

(2) If we start EM with (A, Λ, B) instead of a random point, then EM converges to (A, Λ, B).

(3) The point (A, Λ, B) remains fixed after one completion of the E-step and the M-step.

It is often believed (and actually stated in [122, Theorem 1.5]) that every EM fixed point is a critical point of the log-likelihood function ℓ_U. This statement is not true for the definition of "critical" given in Subsection 2.1.2. In fact, for many instances U, the global maximum $\widehat{P}$ is not critical.

To underscore this important point and its statistical relevance, we tested the EM algorithm on random data matrices U for a range of models with $m = n$. The following example explains Table 2.1.

Example 2.1.6. *In our first simulation, we generated random matrices U from the uniform distribution on Δ_{mn-1} by using `R` and then scaling to get integer entries. For each matrix U, we ran the EM algorithm 2000 times to ensure convergence with high probability to the global maximum $\widehat{P}$ on $\mathcal{M}$. Each run had 2000 steps. We then checked whether $\widehat{P}$ is a critical point of ℓ_U using the rank criterion in [85, (2.3)]. Our results are reported in Table 2.1. The main finding is that, with high probability as the matrix size increases, the MLE $\widehat{P}$ lands on the topological boundary $\partial\mathcal{M}$, and it fails to be critical.*

In a second simulation, we started with matrices $A \in \mathbb{N}^{m\times r}$ and $B \in \mathbb{N}^{r\times n}$ whose entries were sampled uniformly from $\{0, 1, \ldots, 100\}$. We then fixed $P \in \mathcal{M}$ to be the $m\times n$ probability matrix given by AB divided by the sum of its entries. We finally took Tmn samples from the distribution P and recorded the results in an $m\times n$ data matrix U. Thereafter, we applied EM to U. We observed the following. If $T \geq 20$ then the fraction of times the MLE lies in $\partial\mathcal{M}$ is very close to 0. When $T \leq 10$ though, this fraction was higher than the results reported in Table 1. For $T = 10$ and $m = n = 4$, $r = 3$, this fraction was 13%, for $m = n = 5, r = 3$, it was 23%, and for $m = n = 5, r = 4$, it was 17%. Therefore, based on these experiments, in order to have the MLE be a critical point in $\mathcal{M}$, one should have at least 20 times more samples than entries of the matrix. ♢

This brings our attention to the problem of identifying the fixed points of EM. If we could compute all EM fixed points, then this would reveal the global maximizer of ℓ_U. Since a point is EM fixed if and only if it stays fixed after an E-step and an M-step, we can write rational function equations for the EM fixed points in Θ:

$$\lambda_k \;=\; \frac{1}{u_{++}} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{a_{ik}\lambda_k b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij} \quad \text{for all } k,$$

$$a_{ik} \;=\; \frac{1}{\lambda_k u_{++}} \sum_{j=1}^{n} \frac{a_{ik}\lambda_k b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij} \quad \text{for all } i, k,$$

$$b_{kj} \;=\; \frac{1}{\lambda_k u_{++}} \sum_{i=1}^{m} \frac{a_{ik}\lambda_k b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij} \quad \text{for all } k, j.$$

Our goal is to understand the solutions to these equations for a fixed positive matrix U. We seek to find the variety they define in the polytope Θ and the image of that variety in $\mathcal{M}$.

In the EM algorithm we usually start with parameters $a_{ik}, \lambda_k, b_{kj}$ that are strictly positive. The a_{ik} or b_{kl} may become zero in the limit, but the parameters λ_k always remain positive when the u_{ij} are positive since the entries of each column of A and each row of B sum to 1. This justifies that we cancel out the factors λ_k in our equations. After this, the first equation is implied by the other two. Therefore, the set of all EM fixed points is a variety, and it is characterized by

$$a_{ik} \;=\; \frac{1}{u_{++}} \sum_{j=1}^{n} \frac{a_{ik} b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij} \qquad \text{for all } i,k,$$

$$b_{kj} \;=\; \frac{1}{u_{++}} \sum_{i=1}^{m} \frac{a_{ik} b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}} u_{ij} \qquad \text{for all } k,j.$$

Suppose that a denominator $\sum_l a_{il}\lambda_l b_{lj}$ is zero at a point in Θ. Then $a_{ik}b_{kj} = 0$ for all k, and the expression $\frac{a_{ik}b_{kj}}{\sum_{l=1}^{r} a_{il}\lambda_l b_{lj}}$ would be considered 0. Using the identity $p_{ij} = \sum_{l=1}^{r} a_{il}\lambda_l b_{lj}$, we can rewrite our two fixed point equations in the form

$$a_{ik}\big(\sum_{j=1}^{n} \big(u_{++} - \frac{u_{ij}}{p_{ij}}\big) b_{kj}\big) = 0 \text{ for all } k,i \quad \text{and} \quad b_{kj}\big(\sum_{i=1}^{m} \big(u_{++} - \frac{u_{ij}}{p_{ij}}\big) a_{ik}\big) = 0 \text{ for all } k,j. \tag{2.1.10}$$

Let R denote the $m \times n$ matrix with entries $r_{ij} = u_{++} - \frac{u_{ij}}{p_{ij}}$. The matrix R is the gradient of the log-likelihood function $\ell_U(P)$, as seen in [85, (3.1)]. With this, our fixed point equations are

$$a_{ik}\big(\sum_{j=1}^{n} r_{ij} b_{kj}\big) \;=\; 0 \quad \text{for all } k,i \quad \text{and} \quad b_{kj}\big(\sum_{i=1}^{m} r_{ij} a_{ik}\big) \;=\; 0 \quad \text{for all } k,j. \tag{2.1.11}$$

We summarize our discussion in the following theorem, with (2.1.11) rewritten in matrix form.

Theorem 2.1.7. *The variety of EM fixed points in the polytope Θ is defined by the equations*

$$A \star (R \cdot B^T) = 0 \qquad B \star (A^T \cdot R) = 0, \tag{2.1.12}$$

where R is the gradient matrix of the log-likelihood function and $\star$ denotes the Hadamard product. The subset of EM fixed points that are critical points is defined by $R \cdot B^T = 0$ and $A^T \cdot R = 0$.

Proof. Since (2.1.12) is equivalent to (2.1.11), the first sentence is proved by the derivation above. For the second sentence we consider the normal space of the variety $\mathcal{V}$ at a rank r matrix $P = A\Lambda B$. This is the orthogonal complement of the tangent space $\mathrm{T}_P(\mathcal{V})$. The normal space can be expressed as the kernel of the linear map $Q \mapsto (Q \cdot B^T, A^T \cdot Q)$. Hence $R = grad_P(\ell_U)$ is perpendicular to $\mathrm{T}_P(\mathcal{V})$ if and only if $R \cdot B^T = 0$ and $A^T \cdot R = 0$. Therefore, the polynomial equations (2.1.12) define the Zariski closure of the set of parameters for which P is critical. □

The variety defined by (2.1.12) is reducible. In Subsection 2.1.5 we shall present a detailed study of its irreducible components, along with a discussion of their statistical interpretation. As a preview, we here decompose the variety of EM fixed points in the simplest possible case.

Example 2.1.8. *Let $m = n = 2$, $r = 1$, and consider the ideal generated by the cubics in (2.1.12):*

$$\mathcal{F} \;=\; \big\langle a_{11}(r_{11}b_{11}+r_{12}b_{12}),\, a_{21}(r_{21}b_{11}+r_{22}b_{12}),\, b_{11}(a_{11}r_{11}+a_{21}r_{21}),\, b_{12}(a_{11}r_{12}+a_{21}r_{22})\big\rangle.$$

The software `Macaulay2` *[81] computes a primary decomposition into* 12 *components:*

$$\begin{aligned}\mathcal{F} \;=\; & \langle\, r_{11}r_{22}-r_{12}r_{21}, a_{11}r_{11}+a_{21}r_{21}, a_{11}r_{12}+a_{21}r_{22}, b_{11}r_{11}+b_{12}r_{12}, b_{11}r_{21}+b_{12}r_{22}\,\rangle \\ & \cap\, \langle a_{11}, r_{21}, r_{22}\rangle \cap \langle a_{21}, r_{11}, r_{12}\rangle \cap \langle r_{12}, r_{22}, b_{11}\rangle \cap \langle r_{11}, r_{21}, b_{12}\,\rangle \\ & \cap\, \langle a_{11}, r_{22}, b_{11}\rangle \cap \langle a_{11}, r_{21}, b_{12}\rangle \cap \langle a_{21}, r_{12}, b_{11}\rangle \cap \langle a_{21}, r_{11}, b_{12}\rangle \\ & \cap\, \langle a_{11}, a_{21}\rangle \cap \langle b_{11}, b_{12}\rangle \;\cap\; \big(\langle a_{11}, a_{21}\rangle^2 + \langle b_{11}, b_{12}\rangle^2 + \mathcal{F}\big).\end{aligned} \tag{2.1.13}$$

The last primary ideal is embedded. Thus $\mathcal{F}$ is not a radical ideal. Its radical requires an extra generator of degree 5*. The first* 11 *ideals in (2.1.13) are the minimal primes of $\mathcal{F}$. These give the irreducible components of the variety $V(\mathcal{F})$. The first ideal represents the critical points in $\mathcal{M}$.* ♢

2.1.4 Matrices of Nonnegative Rank Three

While the EM algorithm operates in the polytope Θ of model parameters (A, Λ, B), the mixture model $\mathcal{M}$ lives in the simplex $\Delta_{mn-1} \subset \mathbb{R}^{m\times n}$ of all joint distributions. The parametrization ϕ is not identifiable. The topology of its fibers was studied by Mond *et al.* [116], with focus on the first non-trivial case, when the rank r is three. We build on their work to derive a semialgebraic characterization of $\mathcal{M}$. This subsection is self-contained. It can be read independently from our earlier discussion of the EM algorithm. It is aimed at all readers interested in nonnegative matrix factorization, regardless of its statistical relevance.

We now fix $r = 3$. Let A be a real $m \times 3$-matrix with rows $a_1, \ldots, a_m$, and B a real $3 \times n$-matrix with columns $b_1, \ldots, b_n$. The vectors $b_j \in \mathbb{R}^3$ represent points in the projective plane $\mathbb{P}^2$. We view the a_i as elements in the dual space $(\mathbb{R}^3)^*$. These represent lines in $\mathbb{P}^2$. Geometric algebra (a.k.a. Grassmann-Cayley algebra [159]) furnishes two bilinear operations,

$$\vee : \mathbb{R}^3 \times \mathbb{R}^3 \to (\mathbb{R}^3)^* \quad \text{and} \quad \wedge : (\mathbb{R}^3)^* \times (\mathbb{R}^3)^* \to \mathbb{R}^3.$$

These correspond to the classical cross product in 3-space. Geometrically, $a_i \wedge a_j$ is the intersection point of the lines a_i and a_j in $\mathbb{P}^2$, and $b_i \vee b_j$ is the line spanned by the points b_i and b_j in $\mathbb{P}^2$. The pairing $(\mathbb{R}^3)^* \times \mathbb{R}^3 \to \mathbb{R}$ can be denoted by either $\vee$ or $\wedge$. With these conventions, the operations $\vee$ and $\wedge$ are alternating, associative and distributive. For instance, the minor

$$a_i \wedge a_j \wedge a_k \;=\; \det(a_i, a_j, a_k) \tag{2.1.14}$$

vanishes if and only if the lines a_i, a_j and a_k are concurrent. Likewise, the polynomial

$$\begin{aligned}(a_i \wedge a_j) \vee b_{i'} \vee b_{k'} \;\; = \;\; & a_{i1}a_{j2}b_{1i'}b_{2k'} - a_{i1}a_{j2}b_{1k'}b_{2i'} + a_{i1}a_{j3}b_{1i'}b_{3k'} - a_{i1}a_{j3}b_{1k'}b_{3i'} \\ & -a_{i2}a_{j1}b_{1i'}b_{2k'} + a_{i2}a_{j1}b_{1k'}b_{2i'} + a_{i2}a_{j3}b_{2i'}b_{3k'} - a_{i2}a_{j3}b_{2k'}b_{3i'} \\ & -a_{i3}a_{j1}b_{1i'}b_{3k'} + a_{i3}a_{j1}b_{1k'}b_{3i'} - a_{i3}a_{j2}b_{2i'}b_{3k'} + a_{i3}a_{j2}b_{2k'}b_{3i'}\end{aligned} \tag{2.1.15}$$

expresses the condition that the lines a_i and a_j intersect in a point on the line given by $b_{i'}$ and $b_{k'}$. Of special interest is the following formula involving four rows of A and three columns of B:

$$\big(((a_i \wedge a_j) \vee b_{i'}) \wedge a_k\big) \vee \big(((a_i \wedge a_j) \vee b_{j'}) \wedge a_l\big) \vee b_{k'}. \tag{2.1.16}$$

Its expansion is a bihomogeneous polynomial of degree $(6, 3)$ with 330 terms in (A, B).

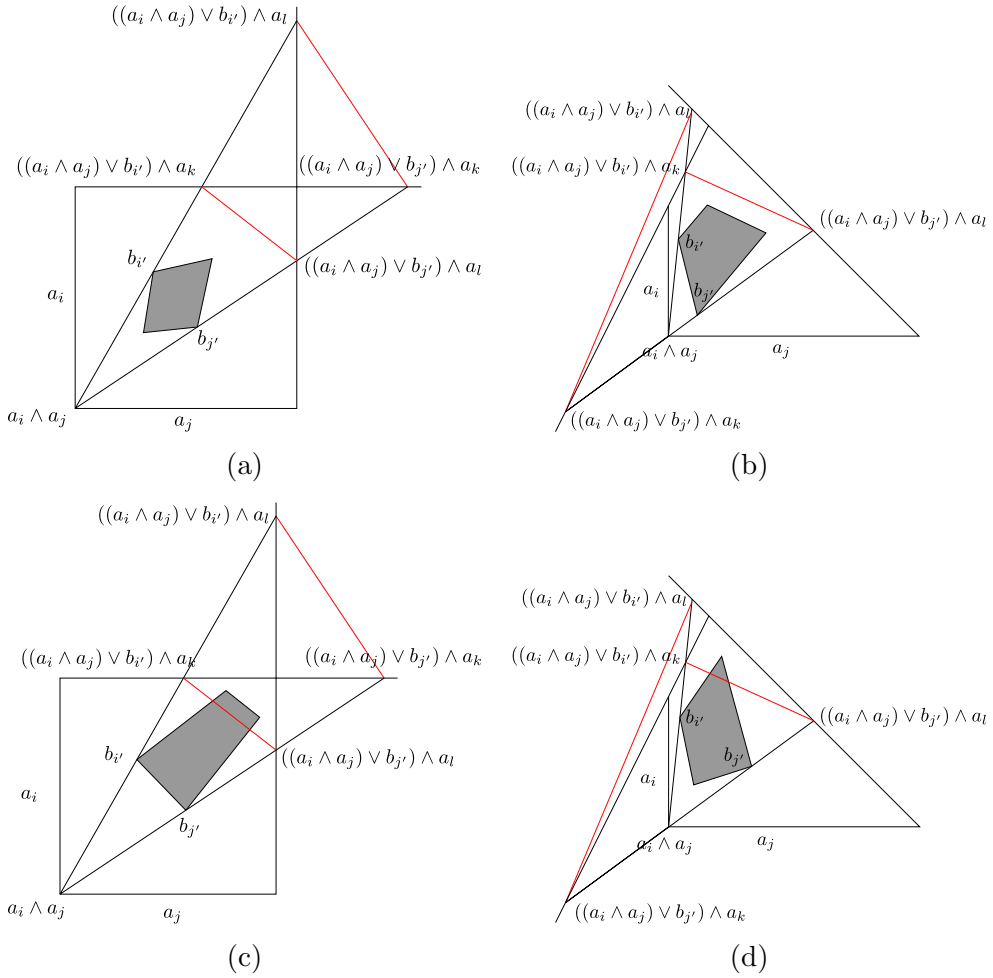


Figure 2.3: In the diagrams (a) and (b), the conditions of Theorem 2.1.9 are satisfied for the chosen i, j, i', j'. In the diagrams (c) and (d), the conditions of Theorem 2.1.9 fail for the chosen i, j, i', j'.

A matrix $P \in \mathbb{R}^{m\times n}$ has nonnegative rank ≤ 3 if it admits a factorization $P = AB$ with A and B nonnegative. The set of such matrices P with $p_{++} = 1$ is precisely the mixture model $\mathcal{M}$ discussed in the earlier subsections. Comparing with (2.1.1), we here subsume the diagonal matrix Λ into either A or B. In what follows, we consider the set $\mathcal{N}$ of pairs (A, B) whose product AB has nonnegative rank ≤ 3. Thus $\mathcal{N}$ is a semialgebraic subset of $\mathbb{R}^{m\times 3} \oplus \mathbb{R}^{3\times n}$. We shall prove:

Theorem 2.1.9. *A pair (A, B) is in $\mathcal{N}$ if and only if $AB \geq 0$ and the following condition holds: Either* $\operatorname{rank}(AB) < 3$, *or* $\operatorname{rank}(AB) = 3$ *and there exist indices $i, j \in [m]$, $i', j' \in [n]$ such that*

sign(2.1.14) is the same or zero for all $k \in [m]\backslash\{i, j\}$
and sign(2.1.15) is the same or zero for all $k' \in [n]\backslash\{i'\}$
and sign((2.1.15)$[i' \to j']$) is the same or zero for all $k' \in [n]\backslash\{j'\}$
and (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] \geq 0$ for all $\{k, l\} \subseteq [m]\backslash\{i, j\}$ and $k' \in [n]\backslash\{i', j'\}$,
or there exist $i, j \in [n]$, $i', j' \in [m]$ such that these conditions hold after swapping A with B^T.

Here, $[m] = \{1, 2, \ldots, m\}$, and the notation $[i' \to j']$ means that the index i' is replaced by the index j' in the preceding expression, and $[k \leftrightarrow l]$ means that k and l are switched.

Theorem 2.1.9 is our main result in Subsection 2.1.4. It gives a finite disjunction of conjunctions of polynomial inequalities in A and B, and thus a quantifier-free first order formula for $\mathcal{N}$. This represents our mixture model as follows: to test whether P lies in $\mathcal{M}$, check whether $\operatorname{rank}(P) \leq 3$; if yes, compute any rank 3 factorization $P = AB$ and check whether (A, B) lies in $\mathcal{N}$. Code for performing these computations in `Macaulay2` is posted on our website.

Theorem 2.1.9 is an algebraic translation of a geometric algorithm. For an illustration see Figure 2.3. In the rest of the subsection, we will study the geometric description of nonnegative rank that leads to the algorithm. Let P be a nonnegative $m \times n$ matrix of rank r. We write $\operatorname{span}(P)$ and $\operatorname{cone}(P)$ for the linear space and the cone spanned by the columns of P, and we define

$$\mathcal{A} = \operatorname{span}(P) \cap \Delta_{m-1} \qquad \text{and} \qquad \mathcal{B} = \operatorname{cone}(P) \cap \Delta_{m-1}. \tag{2.1.17}$$

The matrix P has a size r nonnegative factorization if and only if there exists a polytope Δ with r vertices such that $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$; see Lemma 1.1.4. Without loss of generality, we will assume in the rest of this subsection that the vertices of Δ lie on the boundary of $\mathcal{A}$. We write $\mathcal{M}_r$ for the set of $m \times n$-matrices of nonnegative rank $\leq r$. Here is an illustration that is simpler than Example 2.1.2:

Example 2.1.10. *In [64, §2.7.2], the following family of matrices of rank ≤ 3 is considered:*

$$P(a, b) = \begin{bmatrix} 1-a & 1+a & 1+a & 1-a \\ 1-b & 1-b & 1+b & 1+b \\ 1+a & 1-a & 1-a & 1+a \\ 1+b & 1+b & 1-b & 1-b \end{bmatrix}. \tag{2.1.18}$$

Here, $\mathcal{B}$ is a rectangle and $\mathcal{A} = \{x \in \Delta_3 : x_1 - x_2 + x_3 - x_4 = 0\}$ is a square, see Figure 2.4. Using Theorem 2.1.9, we can check that $P(a, b)$ lies in $\mathcal{M}_3$ if and only if $ab + a + b \leq 1$. $\diamondsuit$

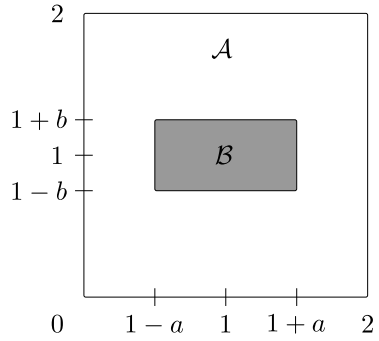


Figure 2.4: The matrix $P(a, b)$ defines a nested pair of rectangles.

Lemma 2.1.11. *A matrix $P \in \mathbb{R}^{m\times n}_{\geq 0}$ of rank r lies in the interior of $\mathcal{M}_r$ if and only if there exists an $(r-1)$-simplex $\Delta \subseteq \mathcal{A}$ such that $\mathcal{B}$ is contained in the interior of Δ. It lies on the boundary of $\mathcal{M}_r$ if and only if every $(r-1)$-simplex Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ contains a vertex of $\mathcal{B}$ on its boundary.*

(a) (b)

Figure 2.5: Critical configurations

For $r = 3$, Mond *et al.* [116] prove the following result. Suppose $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ and every edge of Δ contains a vertex of $\mathcal{B}$. Then, $t\mathcal{B} \subseteq \Delta' \subseteq \mathcal{A}$ for some triangle Δ' and some $t > 1$, unless

(a) an edge of Δ contains an edge of $\mathcal{B}$, or

(b) a vertex of Δ coincides with a vertex of $\mathcal{A}$.

Here the dilate $t\mathcal{B}$ is taken with respect to a point in the interior of $\mathcal{B}$. By Lemma 2.1.11, this means that P lies in the interior of $\mathcal{M}_3^{m\times n}$ unless one of (a) and (b) holds. The conditions (a) and (b) are shown in Figure 2.5. For the proof of this result we refer to [116, Lemma 3.10 and Lemma 4.3].

Corollary 2.1.12. *A matrix $P \in \mathcal{M}_3$ lies on the boundary of $\mathcal{M}_3$ if and only if*

- *P has a zero entry, or*
- *$\operatorname{rank}(P) = 3$ and if Δ is any triangle with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ then every edge of Δ contains a vertex of $\mathcal{B}$, and (a) or (b) holds.*

Corollary 2.1.13. *A matrix $P \in \mathbb{R}^{m\times n}_{\geq 0}$ has nonnegative rank ≤ 3 if and only if*

- *$\operatorname{rank}(P) < 3$, or*
- *$\operatorname{rank}(P) = 3$ and there exists a triangle Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ such that a vertex of Δ coincides with a vertex of $\mathcal{A}$, or*
- *$\operatorname{rank}(P) = 3$ and there exists a triangle Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ such that an edge of Δ contains an edge of $\mathcal{B}$.*

Corollary 2.1.13 provides a geometric algorithm similar to that of Aggarwal *et al.* [1] for checking whether a matrix has nonnegative rank 3. For the algorithm, we need to consider one condition for every vertex of $\mathcal{A}$ and one condition for every edge of $\mathcal{B}$. We now explain these conditions.

Let v be a vertex of $\mathcal{A}$. Let b_1, b_2 be the vertices of $\mathcal{B}$ such that $l_1 = \overline{vb_1}$ and $l_2 = \overline{vb_2}$ support $\mathcal{B}$. Let Δ be the convex hull of v and the other two intersection points of the lines l_1, l_2 with the boundary of $\mathcal{A}$. If $\mathcal{B} \subseteq \Delta$, then P has nonnegative rank 3.

Let l be the line spanned by an edge of $\mathcal{B}$. Let v_1, v_2 be the intersection points of l with $\partial\mathcal{A}$. Let b_1, b_2 be the vertices of $\mathcal{B}$ such that $l_1 = \overline{v_1b_1}$ and $l_2 = \overline{v_2b_2}$ support $\mathcal{B}$. Let v_3 be the intersection point of l_1 and l_2. If $\operatorname{conv}(v_1, v_2, v_3) \subseteq \mathcal{A}$, then P has nonnegative rank 3.

Proof of Theorem 2.1.9. Let $\operatorname{rank}(P) = 3$ and consider any factorization $P = AB$ where $a_1, \ldots, a_m \in (\mathbb{R}^3)^*$ are the row vectors of A and $b_1, \ldots, b_n \in \mathbb{R}^3$ are the column vectors of B. The map $x \mapsto Ax$ identifies $\mathbb{R}^3$ with the common column space of A and P. Under this identification, and by passing from 3-dimensional cones to polygons in $\mathbb{R}^2$, we can assume that the edges of $\mathcal{A}$ are given by $a_1, \ldots, a_m$ and the vertices of $\mathcal{B}$ are given by $b_1, \ldots, b_n$.

To test whether P belongs to $\mathcal{M}_3$, we use the geometric conditions in Corollary 2.1.13. These still involve a quantifier over Δ. Our aim is to translate them into the given quantifier-free formula, referring only to the vertices b_i of $\mathcal{B}$ and the edges a_j of $\mathcal{A}$. First we check with the sign condition on (2.1.14) that the intersection point $a_i \wedge a_j$ defines a vertex of $\mathcal{A}$. Next we verify that the lines $(a_i \wedge a_j) \vee b_{i'}$ and $(a_i \wedge a_j) \vee b_{j'}$ are supporting $\mathcal{B}$, i.e. all vertices of $\mathcal{B}$ lie on the same side of the lines $(a_i \wedge a_j) \vee b_{i'}$ and $(a_i \wedge a_j) \vee b_{j'}$. For this we use the sign conditions on (2.1.15) and (2.1.15)$[i' \to j']$.

Finally we need to check whether all vertices of $\mathcal{B}$ belong to the convex hull of $a_i \wedge a_j$ and the other two intersection points of the lines $(a_i \wedge a_j) \vee b_{i'}$ and $(a_i \wedge a_j) \vee b_{j'}$ with the boundary of $\mathcal{A}$. Fix $\{k, l\} \subseteq [m]\backslash\{i, j\}$. If either the line $(a_i \wedge a_j) \vee b_{i'}$ intersects a_k or the line $(a_i \wedge a_j) \vee b_{j'}$ intersects a_l outside $\mathcal{A}$, then the polygon $\mathcal{B}$ lies completely on one side of the line $(((a_i \wedge a_j) \vee b_{i'}) \wedge a_k) \vee (((a_i \wedge a_j) \vee b_{j'}) \wedge a_l)$. Similarly, if either the line

$(a_i \wedge a_j) \vee b_{i'}$ intersects a_l or the line $(a_i \wedge a_j) \vee b_{j'}$ intersects a_k outside $\mathcal{A}$, then the polygon $\mathcal{B}$ lies completely on one side of the line $(((a_i \wedge a_j) \vee b_{i'}) \wedge a_l) \vee (((a_i \wedge a_j) \vee b_{j'}) \wedge a_k)$. Then the condition (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] \geq 0$ is automatically satisfied for all $k' \in [n] \backslash \{i', j'\}$. If the intersection points $((a_i \wedge a_j) \vee b_{i'}) \wedge a_k$ and $((a_i \wedge a_j) \vee b_{j'}) \wedge a_l$ are on the boundary of $\mathcal{A}$, then the polygon $\mathcal{B}$ is on one side of $(((a_i \wedge a_j) \vee b_{i'}) \wedge a_l) \vee (((a_i \wedge a_j) \vee b_{j'}) \wedge a_k)$. In this case, we use the conditions (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] \geq 0$ to check whether $\mathcal{B}$ is also on one side of the line $(((a_i \wedge a_j) \vee b_{i'}) \wedge a_k) \vee (((a_i \wedge a_j) \vee b_{j'}) \wedge a_l)$. For an illustration see Figure 2.3. □

We wish to reiterate that the semialgebraic formula for our model in Theorem 2.1.9 is quantifier-free. It is a finite Boolean combination of polynomial inequalities with rational coefficients.

Corollary 2.1.14. *If a rational $m \times n$ matrix P has nonnegative rank ≤ 3 then there exists a nonnegative rank ≤ 3 factorization $P = AB$ where all entries of A and B are rational numbers.*

This answers a question of Cohen and Rothblum in [43] for matrices of nonnegative rank 3. It is not known whether this result holds in general. In Subsection 2.1.6 we apply Theorem 2.1.9 to derive the topological boundary and the algebraic boundary of $\mathcal{M}$. Also, using what follows in Subsection 2.1.5, we shall see how these boundaries are detected by the EM algorithm.

2.1.5 Decomposing the variety of EM fixed points

After this in-depth study of the geometry of our model, we now return to the fixed points of Expectation Maximization on $\mathcal{M}$. We fix the polynomial ring $\mathbb{Q}[A, R, B]$ in $mr + mn + rn$ indeterminates a_{ik}, r_{ij}, and b_{kj}. Let $\mathcal{F}$ denote the ideal generated by the entries of the matrices $A \star (R \cdot B^T)$ and $B \star (A^T \cdot R)$ in (2.1.12). Also, let $\mathcal{C}$ denote the ideal generated by the entries of $R \cdot B^T$ and $A^T \cdot R$. Thus $\mathcal{F}$ is generated by $mr + rn$ cubics, $\mathcal{C}$ is generated by $mr + rn$ quadrics, and we have the inclusion $\mathcal{F} \subset \mathcal{C}$. By Theorem 2.1.7, the variety $V(\mathcal{C})$ consists of those parameters A, R, B that correspond to critical points for the log-likelihood function ℓ_U, while the variety $V(\mathcal{F})$ encompasses all the fixed points of the EM algorithm. We are interested in the irreducible components of the varieties $V(\mathcal{F})$ and $V(\mathcal{C})$. These are the zero sets of the minimal primes of $\mathcal{F}$ and $\mathcal{C}$, respectively. More precisely, if $\mathcal{F}$ has minimal primes $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N$, then $V(\mathcal{F}_i)$ are the irreducible components of $V(\mathcal{F})$, and $V(\mathcal{F}) = \bigcup_i V(\mathcal{F}_i)$.

Recall that the matrix R represents the gradient of the log-likelihood function ℓ_U, i.e.

$$r_{ij} \;=\; u_{++} - \frac{u_{ij}}{p_{ij}} \;=\; u_{++} - \frac{u_{ij}}{\sum_k a_{ik}\lambda_k b_{kj}}. \tag{2.1.19}$$

The set of EM-fixed points corresponding to a data matrix $U \in \mathbb{N}^{m \times n}$ is defined by the ideal $\mathcal{F}' \subset \mathbb{Q}[A, B, \Lambda]$ that is obtained from $\mathcal{F}$ by substituting (2.1.19), clearing denominators,

and saturating. Note that $V(\mathcal{F}') = \bigcup_i V(\mathcal{F}'_i)$. So, studying the minimal primes $\mathcal{F}_i$ will help us study the fixed points of EM. A big advantage of considering $\mathcal{F}$ rather than $\mathcal{F}'$ is that $\mathcal{F}$ is much simpler. Also, it does not depend on the data U. This allows a lot of the work in exact MLE using algebraic methods (as in Example 2.1.1) to be done in a preprocessing stage.

There are two important points we wish to make in this subsection:

1. the minimal primes of $\mathcal{F}$ have interesting statistical interpretations, and
2. the non-trivial boundaries of the mixture model $\mathcal{M}$ can be detected from this.

We shall explain these points by working out two cases that are larger than Example 2.1.8.

Example 2.1.8 showed that $\mathcal{F}$ is not radical but has embedded components. Here, we focus on the minimal primes $\mathcal{F}_i$ of $\mathcal{F}$, as these correspond to geometric components of $V(\mathcal{F})$. If $\mathcal{F}_i$ is also a minimal prime of $\mathcal{C}$ then $\mathcal{F}_i$ is a *critical* prime of $\mathcal{F}$. Not every minimal prime of $\mathcal{C}$ is a minimal prime of $\mathcal{F}$. For instance, for $m = n = 2, r = 1$, the ideal $\mathcal{C}$ is the intersection of the first prime in Example 2.1.8 and $\langle a_{11}, a_{21}, b_{11}, b_{12}\rangle$. The latter is not minimal over $\mathcal{F}$. We now generalize this example:

Proposition 2.1.15. *The ideal $\mathcal{C}$ has precisely $r+1$ minimal primes, indexed by $k = 1, \ldots, r+1$:*

$$\mathcal{C} + \big\langle k\text{-minors of } A\big\rangle + \big\langle (m{-}k{+}2)\text{-minors of } R\big\rangle + \big\langle (n{-}m{+}k)\text{-minors of } B\big\rangle \quad \text{if } m \le n,$$
$$\mathcal{C} + \big\langle (m{-}n{+}k)\text{-minors of } A\big\rangle + \big\langle (n{-}k{+}2)\text{-minors of } R\big\rangle + \big\langle k\text{-minors of } B\big\rangle \quad \text{if } m \ge n.$$

Moreover, the ideal $\mathcal{C}$ is radical and, hence, it equals the intersection of its minimal primes.

We refer to Example 2.1.29 for an illustration of Proposition 2.1.15. The proof we give in Appendix 2.1.7.1 relies on methods from representation theory. The duality relation (2.1.26) plays an important role.

We now proceed to our case studies of the minimal primes of the EM fixed ideal $\mathcal{F}$.

Example 2.1.16. *Let $m = n = 3$ and $r = 2$. The ideal $\mathcal{F}$ has 37 minimal primes, in six classes. The first three are the minimal primes of the critical ideal $\mathcal{C}$, as seen in Proposition 2.1.15:*

$$\begin{aligned}
I_1 = \;& \langle r_{23}r_{32} - r_{22}r_{33}, r_{13}r_{32} - r_{12}r_{33}, r_{23}r_{31} - r_{21}r_{33}, r_{22}r_{31} - r_{21}r_{32}, r_{13}r_{31} - r_{11}r_{33}, \\
& r_{12}r_{31} - r_{11}r_{32}, r_{13}r_{22} - r_{12}r_{23}, r_{13}r_{21} - r_{11}r_{23}, r_{12}r_{21} - r_{11}r_{22}, \\
& b_{21}r_{31} + b_{22}r_{32} + b_{23}r_{33}, b_{11}r_{31} + b_{12}r_{32} + b_{13}r_{33}, b_{21}r_{21} + b_{22}r_{22} + b_{23}r_{23}, \\
& b_{11}r_{21} + b_{12}r_{22} + b_{13}r_{23}, a_{12}r_{13} + a_{22}r_{23} + a_{32}r_{33}, a_{11}r_{13} + a_{21}r_{23} + a_{31}r_{33}, \\
& a_{12}r_{12} + a_{22}r_{22} + a_{32}r_{32}, a_{11}r_{12} + a_{21}r_{22} + a_{31}r_{32}, b_{21}r_{11} + b_{22}r_{12} + b_{23}r_{13}, \\
& b_{11}r_{11} + b_{12}r_{12} + b_{13}r_{13}, a_{12}r_{11} + a_{22}r_{21} + a_{32}r_{31}, a_{11}r_{11} + a_{21}r_{21} + a_{31}r_{31}\rangle, \\
I_2 = \;& \langle r_{13}r_{22}r_{31} - r_{12}r_{23}r_{31} - r_{13}r_{21}r_{32} + r_{11}r_{23}r_{32} + r_{12}r_{21}r_{33} - r_{11}r_{22}r_{33}, \\
& b_{21}r_{31} + b_{22}r_{32} + b_{23}r_{33}, b_{11}r_{31} + b_{12}r_{32} + b_{13}r_{33}, b_{21}r_{21} + b_{22}r_{22} + b_{23}r_{23}, \\
& b_{11}r_{21} + b_{12}r_{22} + b_{13}r_{23}, a_{12}r_{13} + a_{22}r_{23} + a_{32}r_{33}, a_{11}r_{13} + a_{21}r_{23} + a_{31}r_{33}, \\
& a_{12}r_{12} + a_{22}r_{22} + a_{32}r_{32}, a_{11}r_{12} + a_{21}r_{22} + a_{31}r_{32}, b_{21}r_{11} + b_{22}r_{12} + b_{23}r_{13}, \\
& b_{11}r_{11} + b_{12}r_{12} + b_{13}r_{13}, a_{12}r_{11} + a_{22}r_{21} + a_{32}r_{31}, a_{11}r_{11} + a_{21}r_{21} + a_{31}r_{31}, \\
& b_{13}b_{22} - b_{12}b_{23}, b_{13}b_{21} - b_{11}b_{23}, b_{12}b_{21} - b_{11}b_{22}, a_{31}a_{22} - a_{21}a_{32}, a_{31}a_{12} - a_{11}a_{32}, a_{21}a_{12} - a_{11}a_{22}\rangle, \\
I_3 = \;& \langle a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}, b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{23}\rangle.
\end{aligned}$$

In addition to these three, $\mathcal{F}$ *has* 12 *non-critical components like*

$$J_1 = \langle a_{11}, a_{21}, r_{31}, r_{32}, r_{33}, r_{13}r_{22} - r_{12}r_{23}, r_{13}r_{21} - r_{11}r_{23}, r_{12}r_{21} - r_{11}r_{22}, \\ b_{21}r_{21}+b_{22}r_{22}+b_{23}r_{23}, b_{21}r_{11}+b_{22}r_{12}+b_{23}r_{13}, a_{12}r_{13}+a_{22}r_{23}, a_{12}r_{12}+a_{22}r_{22}, a_{12}r_{11}+a_{22}r_{21}\rangle,$$

four non-critical components like

$$J_2 = \langle a_{11}, a_{21}, a_{31}, r_{13}r_{22}r_{31} - r_{12}r_{23}r_{31} - r_{13}r_{21}r_{32} + r_{11}r_{23}r_{32} + r_{12}r_{21}r_{33} - r_{11}r_{22}r_{33}, \\ b_{21}r_{21} + b_{22}r_{22} + b_{23}r_{23}, b_{21}r_{11} + b_{22}r_{12} + b_{23}r_{13}, b_{21}r_{31} + b_{22}r_{32} + b_{23}r_{33}, \\ a_{12}r_{13} + a_{22}r_{23} + a_{32}r_{33}, a_{12}r_{12} + a_{22}r_{22} + a_{32}r_{32}, a_{12}r_{11} + a_{22}r_{21} + a_{32}r_{31}\rangle,$$

and 18 *non-critical components like*

$$J_3 = \langle a_{11}, a_{21}, b_{11}, b_{12}, r_{33}, r_{13}r_{22}r_{31} - r_{12}r_{23}r_{31} - r_{13}r_{21}r_{32} + r_{11}r_{23}r_{32}, \\ b_{21}r_{31} + b_{22}r_{32}, b_{21}r_{21} + b_{22}r_{22} + b_{23}r_{23}, b_{21}r_{11} + b_{22}r_{12} + b_{23}r_{13}, \\ a_{12}r_{13} + a_{22}r_{23}, a_{12}r_{12} + a_{22}r_{22} + a_{32}r_{32}, a_{12}r_{11} + a_{22}r_{21} + a_{32}r_{31}\rangle.$$

Each of the 34 *primes* J_1, J_2, J_3 *specifies a face of the polytope* Θ*, as it contains two, three or four of the parameters* a_{ik}, b_{kj}*, and expresses rank constraints on the matrix* $R = [r_{ij}]$*.* $\diamondsuit$

Remark 2.1.17. *Assuming the sample size* u_{++} *to be known, we can recover the data matrix* U *from the gradient* R *using the formula* $U = R \star P + u_{++}P$*. In coordinates, this says*

$$u_{ij} = (r_{ij} + u_{++}) \cdot p_{ij} \qquad \text{for } i \in [m], j \in [n].$$

This formula is obtained by rewriting (2.1.19). Hence, $r_{ij} = 0$ *holds if and only if* $p_{ij} = u_{ij}/u_{++}$*. This can be rephrased as follows. If a minimal prime of* $\mathcal{F}$ *contains the unknown* r_{ij}*, then the corresponding fixed points of the EM algorithm maintain the cell entry* u_{ij} *from the data.*

With this, we can now understand the meaning of the various components in Example 2.1.16. The prime I_1 parametrizes critical points P of rank 2. This represents the behavior of the EM algorithm when run with random starting parameters in the interior of Θ. For special data U, the MLE will be a rank 1 matrix, and such cases are captured by the critical component I_2. The components I_3 and J_2 can be disregarded because each of them contains a column of A. This would force the entries of that column to sum to 0, which is impossible in Θ.

The components J_1 and J_3 describe interesting scenarios that are realized by starting the EM algorithm with parameters on the boundary of the polytope Θ. On the components J_1, the EM algorithm produces an estimate that maintains one of the rows or columns from the data U, and it replaces the remaining table of format 2×3 or 3×2 by its MLE of rank 1. This process amounts to fitting a context specific independence (CSI) model to the data. Following Georgi and Schliep [74], CSI means that independence holds only for some values of the involved variables. Namely, J_1 expresses the constraint that X is independent of Y given that Y is either 1 or 2. Finally, on the components J_3, we have $\text{rank}(A) = \text{rank}(B) = 2$ and $r_{ij} = 0$ for one cell entry (i, j).

Definition 2.1.18. *Let* $\mathcal{F} = \langle A \star (R \cdot B^T), B \star (A^T \cdot R) \rangle$ *be the ideal of EM fixed points. A minimal prime of* $\mathcal{F}$ *is called* relevant *if it contains none of the* mn *polynomials* $p_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}$.

In Example 2.1.8 only the first minimal prime is relevant. In Example 2.1.16 all minimal primes besides I_3 are relevant. Restricting to the relevant minimal primes is justified because the EM algorithm never outputs a matrix containing zeros for positive starting data. Note also that the p_{ij} appear in the denominators in the expressions (2.1.10) that were used in our derivation of $\mathcal{F}$.

Our main result in this subsection is the computation in Theorem 2.1.19. We provide a census of EM fixed points for 4×4-matrices of rank $r = 3$. This is the smallest case where rank can differ from nonnegative rank, and the boundary hypersurfaces (2.1.16) appear.

Theorem 2.1.19. *Let* $m = n = 4$ *and* $r = 3$. *The radical of the EM fixed point ideal* $\mathcal{F}$ *has* 49000 *relevant primes. These come in* 108 *symmetry classes, listed in Table 2.2.*

Proof. We used an approach that mirrors the primary decomposition of binomial ideals [60]. Recall that the EM fixed point ideal equals

$$\begin{aligned} \mathcal{F} \;&=\; \langle\, A \star (R \cdot B^T), B \star (A^T \cdot R) \,\rangle \\ &=\; \langle\, a_{ik}(\sum_{l=1}^{n} r_{il} b_{kl})\,,\, b_{kj}(\sum_{l=1}^{m} r_{lj} a_{lk}) \;:\; k \in [r], i \in [m], j \in [n] \,\rangle. \end{aligned}$$

Any prime ideal containing $\mathcal{F}$ contains either a_{ik} or $\sum_{l=1}^{n} r_{il} b_{kl}$ for any $k \in [r], i \in [m]$, and either b_{kj} or $\sum_{l=1}^{m} r_{lj} a_{lk}$ for any $k \in [r], j \in [n]$. We enumerated all primes containing $\mathcal{F}$ according to the set S of unknowns a_{ik}, b_{kj} they contain. There are 2^{24} subsets and the symmetry group acts on this power set by replacing A with B^T, permuting the rows of A, the columns of B, and the columns of A and the rows of B simultaneously. We picked one representative S from each orbit that is relevant, meaning that we excluded those orbits for which some $p_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}$ lies in the ideal $\langle S \rangle$. For each relevant representative S, we computed the *cellular component* $\mathcal{F}_S = \big((\mathcal{F} + \langle S \rangle) : (\prod S^c)^\infty\big)$, where $S^c = \{a_{11}, \ldots, b_{34}\} \backslash S$. Note that $\mathcal{F}_\emptyset = \mathcal{C}$ is the critical ideal. We next minimalized our cellular decomposition by removing all representatives S such that $\mathcal{F}_T \subset \mathcal{F}_S$ for some representative T in another orbit. This led to a list of 76 orbits, comprising 42706 ideals $\mathcal{F}_S$ in total. For the representative $\mathcal{F}_S$, we computed the set $\mathrm{Ass}(\mathcal{F}_S)$ of associated primes P. By construction, the sets $\mathrm{Ass}(\mathcal{F}_S)$ partition the set of relevant primes of $\mathcal{F}$. The block sizes $|\mathrm{A}ss(\mathcal{F}_S)|$ range from 1 to 7. Up to symmetry, each prime is uniquely determined by its attributes in Table 2.2. These are its set S, its degree and codimension, and the ranks rA = rank(A), rB = rank(B), rR = rank(R), rP = rank(P) at a generic point. Our list starts with the four primes from coming from $S = \emptyset$. See Example 2.1.29. In each case, the primality of the ideal was verified using a linear elimination sequence as in [73, Proposition 23 (b)]. Proofs in `Macaulay2` code are posted on our website. □

Below is the complete list of all 108 classes of prime ideals in Theorem 2.1.19. Three components are marked with stars. After the table, we discuss these components in Examples 2.1.20, 2.1.21 and 2.1.22.

Table 2.2: Minimal primes of the EM fixed ideal $\mathcal{F}$ for 4×4-matrices of rank 3

set S	$\|S\|$	a's	b's	deg	codim	rA	rB	rR	rP	$\|orbit\|$
$\emptyset$	0	0	0	1	24	0	0	4	0	1
	0	0	0	1630	19	1	1	3	1	1
	0	0	0	3491	16	2	2	2	2	1
	0	0	0	245	15	3	3	1	3	1
$\{a_{11}\}$	1	1	0	245	16	3	3	1	3	24
	1	1	0	3491	17	2	2	2	2	24
$\{a_{11}, a_{21}\}$	2	2	0	20	17	3	3	1	3	36
	2	2	0	245	17	3	3	1	3	36
	2	2	0	1460	17	2	3	2	2	36
$\{a_{11}, a_{21}, a_{31}\}$	3	3	0	53	17	3	3	1	3	24
	3	3	0	188	17	2	3	2	2	24
$*\{a_{11}, a_{21}, b_{11}, b_{12}\}*$	4	2	2	245	19	3	3	1	3	108
	4	2	2	20	19	3	3	1	3	108 x 2
	4	2	2	1460	19	2	3	2	2	108 x 2
	4	2	2	2370	20	2	2	3	2	108
	4	2	2	240	19	3	3	2	3	108
$\{a_{11}, a_{21}, b_{21}, b_{22}\}$	4	2	2	825	18	3	3	2	3	216
$\{a_{11}, a_{21}, a_{31}, a_{41}\}$	4	4	0	689	16	2	3	2	2	6
	4	4	0	474	17	1	2	3	1	6
$\{a_{11}, a_{21}, a_{12}, a_{22}\}$	4	4	0	592	17	2	3	2	2	36
	4	4	0	9	17	3	3	1	3	36
$\{a_{11}, a_{21}, a_{32}, a_{42}\}$	4	4	0	20	19	3	3	1	3	36 x 2
	4	4	0	245	19	3	3	1	3	36
	4	4	0	400	18	2	3	2	2	36
$\{a_{11}, a_{21}, a_{31}, b_{11}, b_{12}\}$	5	3	2	474	20	2	2	3	2	144
	5	3	2	188	19	2	3	2	2	144
	5	3	2	448	19	3	3	2	3	144
	5	3	2	53	19	3	3	1	3	144
$\{a_{11}, a_{21}, a_{31}, b_{21}, b_{22}\}$	5	3	2	125	18	3	3	2	3	288
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{31}\}$	5	4	1	723	19	3	3	2	3	144
$\{a_{11}, a_{21}, a_{31}, b_{11}, b_{12}, b_{13}\}$	6	3	3	689	19	3	3	2	3	48
	6	3	3	474	20	2	2	3	2	48
$\{a_{11}, a_{21}, a_{31}, b_{21}, b_{22}, b_{23}\}$	6	3	3	21	18	3	3	2	3	96
$\{a_{11}, a_{21}, a_{32}, b_{11}, b_{12}, b_{33}\}$	6	3	3	2785	20	3	3	3	3	864
$*\{a_{11}, a_{22}, a_{33}, b_{11}, b_{22}, b_{33}\}*$	6	3	3	9016	21	3	3	4	3	576
	6	3	3	245	21	3	3	1	3	576
$\{a_{11}, a_{21}, a_{31}, a_{41}, b_{21}, b_{22}\}$	6	4	2	265	17	2	3	2	2	72
$\{a_{11}, a_{21}, a_{12}, a_{22}, b_{11}, b_{12}\}$	6	4	2	592	19	2	3	2	2	432
	6	4	2	9	19	3	3	1	3	432
	6	4	2	104	19	3	3	2	3	432
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{11}, b_{12}\}$	6	4	2	825	20	3	3	2	3	432
	6	4	2	100	20	3	3	2	3	432
	6	4	2	400	20	2	3	2	2	432
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{31}, b_{32}\}$	6	4	2	301	19	3	3	2	3	216
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}\}$	6	6	0	265	17	2	3	2	2	72
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}\}$	6	6	0	35	16	2	3	2	2	24
$\{a_{11}, a_{21}, a_{12}, a_{22}, a_{33}, a_{43}\}$	6	6	0	180	18	2	3	2	2	36
	6	6	0	9	19	3	3	1	3	36
$\{a_{11}, a_{21}, a_{31}, a_{41}, b_{21}, b_{22}, b_{23}\}$	7	4	3	35	17	2	3	2	2	48
$\{a_{11}, a_{21}, a_{31}, a_{42}, b_{11}, b_{12}, b_{33}\}$	7	4	3	557	20	3	3	3	3	576

set S	$\|S\|$	a's	b's	deg	codim	rA	rB	rR	rP	$\|orbit\|$
$\{a_{11}, a_{21}, a_{12}, a_{22}, b_{11}, b_{12}, b_{13}\}$	7	4	3	191	19	3	3	2	3	288
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{11}, b_{12}, b_{13}\}$	7	4	3	140	20	3	3	2	3	288
	7	4	3	125	20	3	3	2	3	288
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{11}, b_{12}, b_{33}\}$	7	4	3	835	20	3	3	3	3	864
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{31}, b_{32}, b_{33}\}$	7	4	3	49	19	3	3	2	3	144
$*\{a_{11}, a_{21}, a_{32}, a_{43}, b_{11}, b_{22}, b_{33}\}*$	7	4	3	3087	21	3	3	4	3	1728
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, b_{21}, b_{22}\}$	7	5	2	31	19	3	3	2	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}\}$	7	5	2	225	20	3	3	2	3	864
$\{a_{11}, a_{21}, a_{12}, a_{32}, a_{43}, b_{11}, b_{22}\}$	7	5	2	1193	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{31}, a_{41}, b_{21}, b_{22}, b_{23}, b_{24}\}$	8	4	4	85	15	2	2	3	1	6
$\{a_{11}, a_{21}, a_{31}, a_{41}, b_{21}, b_{22}, b_{33}, b_{34}\}$	8	4	4	81	18	2	3	2	2	36
$\{a_{11}, a_{21}, a_{31}, a_{42}, b_{11}, b_{12}, b_{13}, b_{34}\}$	8	4	4	557	20	3	3	3	3	96
$\{a_{11}, a_{21}, a_{31}, a_{42}, b_{11}, b_{12}, b_{33}, b_{34}\}$	8	4	4	167	20	3	3	3	3	288
$\{a_{11}, a_{21}, a_{12}, a_{22}, b_{11}, b_{12}, b_{21}, b_{22}\}$	8	4	4	850	20	2	2	3	2	108
	8	4	4	45	19	3	3	2	3	108
$\{a_{11}, a_{21}, a_{12}, a_{22}, b_{11}, b_{12}, b_{23}, b_{24}\}$	8	4	4	9	21	3	3	1	3	216
	8	4	4	1024	21	3	2	3	2	216
	8	4	4	104	21	3	3	2	3	216 x 2
	8	4	4	592	21	2	3	2	2	216
$\{a_{11}, a_{21}, a_{12}, a_{32}, b_{11}, b_{12}, b_{21}, b_{23}\}$	8	4	4	2121	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{12}, a_{32}, b_{11}, b_{12}, b_{23}, b_{24}\}$	8	4	4	2125	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{11}, b_{12}, b_{23}, b_{24}\}$	8	4	4	2125	21	3	3	3	3	108
$\{a_{11}, a_{21}, a_{32}, a_{42}, b_{11}, b_{12}, b_{33}, b_{34}\}$	8	4	4	265	20	3	3	3	3	216
$\{a_{11}, a_{21}, a_{32}, a_{43}, b_{11}, b_{12}, b_{23}, b_{34}\}$	8	4	4	2205	21	3	3	4	3	432
$\{a_{11}, a_{21}, a_{32}, a_{43}, b_{11}, b_{22}, b_{23}, b_{34}\}$	8	4	4	1029	21	3	3	4	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, b_{21}, b_{22}, b_{23}\}$	8	5	3	35	19	3	3	2	3	576
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}, b_{13}\}$	8	5	3	265	20	3	3	2	3	576
$\{a_{11}, a_{21}, a_{12}, a_{32}, a_{43}, b_{11}, b_{12}, b_{23}\}$	8	5	3	1185	21	3	3	3	3	3456
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, b_{21}, b_{22}\}$	8	6	2	425	18	2	3	3	2	432
$\{a_{11}, a_{21}, a_{12}, a_{22}, a_{33}, a_{43}, b_{11}, b_{12}\}$	8	6	2	180	20	2	3	2	2	432
	8	6	2	45	20	3	3	2	3	432
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, a_{32}, a_{42}\}$	8	8	0	85	15	1	3	3	1	6
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, a_{33}, a_{43}\}$	8	8	0	81	18	2	3	2	2	36
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, b_{11}, b_{12}, b_{23}, b_{24}\}$	9	5	4	296	21	3	3	3	3	864
	9	5	4	31	21	3	3	2	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}, b_{21}, b_{23}\}$	9	5	4	425	21	3	3	3	3	3456
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}, b_{23}, b_{24}\}$	9	5	4	425	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{12}, a_{22}, a_{33}, b_{11}, b_{12}, b_{23}, b_{24}\}$	9	5	4	839	21	3	3	3	3	432
$\{a_{11}, a_{21}, a_{12}, a_{32}, a_{43}, b_{11}, b_{12}, b_{13}, b_{24}\}$	9	5	4	237	21	3	3	3	3	1152
$\{a_{11}, a_{21}, a_{12}, a_{32}, a_{43}, b_{11}, b_{12}, b_{23}, b_{24}\}$	9	5	4	875	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, b_{21}, b_{22}, b_{23}\}$	9	6	3	85	18	2	3	3	2	288
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{43}, b_{11}, b_{12}, b_{23}\}$	9	6	3	163	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{12}, a_{22}, a_{33}, a_{43}, b_{11}, b_{12}, b_{13}\}$	9	6	3	63	20	3	3	2	3	288
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, a_{32}, b_{21}, b_{22}\}$	9	7	2	85	18	2	3	3	2	288
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, b_{11}, b_{12}, b_{13}, b_{21}, b_{24}\}$	10	5	5	425	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, b_{11}, b_{12}, b_{21}, b_{22}, b_{23}\}$	10	5	5	85	20	3	3	3	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}, b_{13}, b_{21}, b_{24}\}$	10	5	5	425	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{42}, b_{11}, b_{12}, b_{21}, b_{23}, b_{24}\}$	10	5	5	85	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}, b_{11}, b_{12}, b_{21}, b_{22}\}$	10	6	4	85	19	2	3	3	2	144
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{42}, b_{11}, b_{12}, b_{21}, b_{23}\}$	10	6	4	85	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{42}, b_{11}, b_{12}, b_{23}, b_{24}\}$	10	6	4	85	21	3	3	3	3	432
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{43}, b_{11}, b_{12}, b_{13}, b_{24}\}$	10	6	4	237	21	3	3	3	3	576
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{43}, b_{11}, b_{12}, b_{23}, b_{24}\}$	10	6	4	175	21	3	3	3	3	864
$\{a_{11}, a_{21}, a_{12}, a_{22}, a_{33}, a_{43}, b_{11}, b_{12}, b_{23}, b_{24}\}$	10	6	4	225	21	3	3	3	3	216
$\{a_{11}, a_{21}, a_{31}, a_{41}, a_{12}, a_{22}, a_{32}, b_{21}, b_{22}, b_{23}\}$	10	7	3	85	18	2	3	3	2	192
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{42}, b_{11}, b_{12}, b_{13}, b_{21}, b_{24}\}$	11	6	5	85	21	3	3	3	3	1728
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}, b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{23}\}$	12	6	6	85	20	2	2	3	2	48
$\{a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{42}, b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{24}\}$	12	6	6	85	21	3	3	3	3	432

We illustrate our census of relevant primes for three sets S that are especially interesting.

Example 2.1.20. *Let $S = \{a_{11}, a_{21}, b_{11}, b_{12}\}$. The cellular component $\mathcal{F}_S$ is the ideal generated by S, $det(R^{34}_{34})$, $det(R)$, and the entries of the matrices $B^{23}R^T, B^1(R^T)_{34}, R^T A_{23}, (R^T)^{34}A_1$. In specifying submatrices, upper indices refer to rows and lower indices refer to columns. The ideal $\mathcal{F}_S$ is radical with 7 associated primes, to be discussed in order of their appearance in Table 2.2. For instance, the prime (1) below has degree 245. The phrase "Generated by" is meant modulo $\mathcal{F}_S$:*

(1) Generated by entries of $BR^T, A^T R$, and 2×2-minors of R. This gives 60 quadrics.

(2) Generated by entries of $A^T R, R^{34}$, and 2×2-minors of R, A^{12}_{23}. This gives 19 quadrics.

(2') Mirror image of (2) under swapping A and B^T.

(3) Generated by entries of $A^T R$, 2×2-minors of A^{12}_{23}, R^{34}, and 3×3-minors of A, R^{123}, R^{124}. This gives 29 quadrics and 10 cubics.

(3') Mirror image of (3) under swapping A and B^T.

(4) Generated by 2×2-minors of A_{23} and B^{23}. This gives 33 quadrics and one quartic.

(5) Generated by entries of R^{34}_{34}, 2×2-minors of $R^{12}_{34}, R^{34}_{12}, A^{12}_{23}, B^{23}_{12}$, and 3×3-minors of R. This gives 20 quadrics and 4 cubics.

These primes have the following meaning for the EM algorithm.

(1) The fixed points $P = \phi(A, R, B)$ given by this prime ideal are those critical points for the likelihood function ℓ_U for which the parameters $a_{11}, a_{21}, b_{11}, b_{21}$ happen to be 0.

(2) The fixed points $P = \phi(A, R, B)$ given by this prime ideal have the last two rows of P fixed and equal to the last two rows of the data matrix U (divided by the sample size u_{++}). Therefore, the points coming from this ideal are the maximum likelihood estimates with these eight entries fixed and which factor so that $a_{11}, a_{21}, b_{11}, b_{21}$ are 0.

(3) Since the 3×3 minors of A lie in this ideal, we have rank$(P) \leq 2$. Therefore, these fixed points give an MLE of rank 2. This component is the restriction to $V(\mathcal{F}_S)$ of the generic behavior on the singular locus of $\mathcal{V}$.

(4) On this component, the duality relation in (2.1.26) fails since rank$(P) = 2$ but rank$(R) = 3$.

(5) The fixed points $P = \phi(A, R, B)$ given by this ideal have the four entries in the last 2 rows and last 2 columns of P fixed and equal to the corresponding entries in U (divided by u_{++}). Therefore, the points coming from this ideal are maximum likelihood estimates with those four entries fixed, and parameters $a_{11}, a_{21}, b_{11}, b_{21}$ being 0. ♢

Example 2.1.21. *Let* $S = \{a_{11}, a_{21}, a_{32}, a_{43}, b_{11}, b_{22}, b_{33}\}$. *The ideal* $\mathcal{F}_S$ *has codimension* 21, *degree* 3087, *and is generated modulo* $\langle \mathcal{S} \rangle$ *by* 20 *quadrics and two cubics. To show that* $\mathcal{F}_S$ *is prime, we use the elimination method of [73, Proposition 23 (b)], with the variable* x_1 *taken successively to be* $r_{44}, r_{43}, r_{34}, a_{13}, r_{21}, r_{12}, r_{14}, r_{33}, b_{21}, a_{31}, r_{41}, a_{21}, a_{32}$. *The last elimination ideal is generated by an irreducible polynomial of degree* 9, *thus proving primality of* $\mathcal{F}_S$.

If we add the relation $P = AB$ *to* $\mathcal{F}_{\mathcal{S}}$ *and thereafter eliminate* $\{A, B, R\}$, *then we obtain a prime ideal in* $\mathbb{Q}[P]$. *That prime ideal has height one over the determinantal ideal* $\langle det(P) \rangle$. *Any such prime gives a candidate for a component in the boundary of our model* $\mathcal{M}$. *By matching the set* S *with the combinatorial analysis in subsection 2.1.4, we see that Figure 2.5 (b) corresponds to* $V(S)$. *Hence, by Corollary 2.1.12, this component does in fact contribute to the boundary* $\partial\mathcal{M}$. *This is a special case of Theorem 2.1.23 below; see equation (2.1.21) in Example 2.1.24.*

This component is the most important one for EM. It represents the typical behavior when the output of the EM algorithm is not critical. In particular, the duality relation (2.1.26) fails in the most dramatic form because rank$(R) = 4$. *As seen in Table 2.1, this failure is still rare* (4.4%) *for* $m = n = 4$. *For larger matrix sizes, however, the non-critical behavior occurs with overwhelming probability.* ♢

Example 2.1.22. *Let* $S = \{a_{11}, a_{22}, a_{33}, b_{11}, b_{22}, b_{33}\}$. *The computation for the ideal* $\mathcal{F}_S$ *was the hardest among all cellular components. It was found to be radical, with two associated primes of codimension* 21. *The first prime has the largest degree, namely* 9016, *among all entries in Table 2.2. In contrast to Example 2.1.21, the set* S *cannot contribute to* $\partial\mathcal{M}$. *Indeed, for both primes, the elimination ideal in* $\mathbb{Q}[P]$ *is* $\langle det(P) \rangle$. *The degree* 9016 *ideal is the only prime in Table 2.2 that has rank*$(R) = 4$ *but does not map to the boundary of the model* $\mathcal{M}$. *Starting the EM algorithm with zero parameters in* S *generally leads to the correct MLE.* ♢

2.1.6 Algebraic Boundaries

In Subsection 2.1.4 we studied the real algebraic geometry of the mixture model $\mathcal{M}$ for rank three. In this subsection we also fix $r = 3$ and focus on the algebraic boundary of our model. Our main result in this subsection is the characterization of its irreducible components.

Theorem 2.1.23. *The algebraic boundary* $\overline{\partial\mathcal{M}}$ *is a pure-dimensional reducible variety in* $\mathbb{P}^{mn-1}$. *All irreducible components have dimension* $3m + 3n - 11$ *and their number equals*

$$mn + \frac{m(m-1)(m-2)(m+n-6)n(n-1)(n-2)}{4}.$$

Besides the mn *components* $\{p_{ij} = 0\}$ *that come from* $\partial\Delta_{mn-1}$ *there are:*

(a) $36\binom{m}{3}\binom{n}{4}$ *components parametrized by* $P = AB$, *where* A *has three zeros in distinct rows and columns, and* B *has four zeros in three rows and distinct columns.*

(b) $36\binom{m}{4}\binom{n}{3}$ *components parametrized by* $P = AB$*, where* A *has four zeros in three columns and distinct rows, and* B *has three zeros in distinct rows and columns.*

This result takes the following specific form in the first non-trivial case:

Example 2.1.24. *For* $m = n = 4$*, the algebraic boundary of our model* $\mathcal{M}$ *has* 16 *irreducible components* $\{p_{ij} = 0\}$*,* 144 *irreducible components corresponding to factorizations like*

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & b_{13} & b_{14} \\ b_{21} & b_{22} & 0 & b_{24} \\ b_{31} & b_{32} & b_{33} & 0 \end{bmatrix}, \tag{2.1.20}$$

and 144 *irreducible components that are transpose to those in (2.1.20), i.e.*

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \\ a_{41} & a_{42} & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & b_{12} & b_{13} & b_{14} \\ b_{21} & 0 & b_{23} & b_{24} \\ b_{31} & b_{32} & 0 & b_{34} \end{bmatrix}. \tag{2.1.21}$$

The prime ideal of each component is generated by the determinant and four polynomials of degree six. These are the maximal minors of a 4×5*-matrix. For the component (2.1.21), this can be chosen as*

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & 0 \\ p_{21} & p_{22} & p_{23} & p_{24} & 0 \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{33}(p_{11}p_{22} - p_{12}p_{21}) \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{41}(p_{12}p_{23} - p_{13}p_{22}) + p_{43}(p_{11}p_{22} - p_{12}p_{21}) \end{bmatrix}. \tag{2.1.22}$$

This matrix representation was suggested to us by Aldo Conca and Matteo Varbaro. ♢

We begin by resolving a problem that was stated in [85, §5] and [93, Example 2.13]:

Proposition 2.1.25. *The ML degree of each variety (2.1.20) in the algebraic boundary* $\overline{\partial\mathcal{M}}$ *is* 633*.*

Proposition 2.1.25 is a first step towards deriving an exact representation of the MLE function $U \mapsto \widehat{P}$ for our model $\mathcal{M} =$ ●—○—●. As highlighted in Table 2.1, the MLE $\widehat{P}$ typically lies on the boundary $\partial\mathcal{M}$. We now know that this boundary has $304 = 16{+}144{+}144$ strata $X_1, X_2, \ldots, X_{304}$. If $\widehat{P}$ lies on exactly one of the strata (2.1.20) or (2.1.21), then we can expect the coordinates of $\widehat{P}$ to be algebraic numbers of degree 633 over the rationals $\mathbb{Q}$. This is the content of Proposition 2.1.25. By [85, Theorem 1.1] the degree of $\widehat{P}$ over $\mathbb{Q}$ is only 191 if $\widehat{P}$ happens to lie in the interior of $\mathcal{M}$.

In order to complete the exact analysis of MLE for the 4×4-model, we also need to determine which intersections $X_{i_1} \cap \cdots \cap X_{i_s}$ are non-empty on $\partial\mathcal{M}$. For each such non-empty stratum, we would then need to compute its ML degree. This is a challenge left for a future project.

Proof of Theorem 2.1.23. By Corollary 2.1.12, an $m \times n$ matrix P of rank 3 without zero entries lies on $\partial\mathcal{M}_3^{m\times n}$ if and only if all triangles Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ contain an edge of $\mathcal{B}$ on one of its edges and a vertex of $\mathcal{B}$ on all other edges, or one of its vertices coincides with a vertex of $\mathcal{A}$ and all other edges contain a vertex of $\mathcal{B}$. We will write down these conditions algebraically.

The columns of A correspond to the vertices of Δ, and the columns of B correspond to the convex combinations of the vertices of Δ that give the columns of $P = AB$. If a vertex of Δ and a vertex of $\mathcal{A}$ coincide, then the corresponding column of A has two 0's. Otherwise the corresponding column of A has one 0. If a vertex of $\mathcal{B}$ lies on an edge of Δ, then one entry of B is zero.

We can freely permute the columns of the left $m \times 3$ matrix A of a factorization – this corresponds to permuting the rows of the corresponding right $3 \times n$ matrix B. Thus we can assume that the first column contains two 0's and/or the rest of the 0's appear in the increasing order.

In the first case, there are $\binom{m}{3}$ possibilities for choosing the three rows of A containing 0's, there are 3 choices for the row of B with two 0's, $\binom{n}{2}$ possibilities for choosing the positions for the two 0's, and $(n-2)(n-3)$ possibilities for choosing the positions of the 0's in the other two rows of B. In the second case, there are $\binom{m}{2}$ possibilities for choosing the 0's in the first column of A and $\binom{m-2}{2}$ choices for the positions of the 0's in other columns. There are $\binom{n}{3}$ choices for the columns of B containing 0's and 3! choices for the positions of the 0's in these columns. □

The prime ideal in (2.1.22) can be found and verified by direct computation, e.g. by using the software `Macaulay2` [81]. For general values of m and n, the prime ideal of an irreducible boundary component is generated by quartics and sextics that generalize those in Example 2.1.24. The following theorem was stated as a conjecture in the original December 2013 version of this section. That conjecture was proved in April 2014 by Eggermont, Horobeţ and Kubjas [59].

Theorem 2.1.26 (Eggermont, Horobeţ and Kubjas)**.** *Let $m \geq 4, n \geq 3$ and consider the irreducible component of $\overline{\partial\mathcal{M}}$ in Theorem 2.1.23 (b). The prime ideal of this component is minimally generated by $\binom{m}{4}\binom{n}{4}$ quartics, namely the 4×4-minors of P, and by $\binom{n}{3}$ sextics that are indexed by subsets $\{i,j,k\}$ of $\{1,2,\ldots,n\}$. These form a Gröbner basis with respect to the graded reverse lexicographic order. The sextic indexed by $\{i,j,k\}$ is homogeneous of degree $e_1{+}e_2{+}e_3{+}e_i{+}e_j{+}e_k$ in the column grading by $\mathbb{Z}^n$ and homogeneous of degree $2e_1{+}2e_2{+}e_3{+}e_4$ in the row grading by $\mathbb{Z}^m$.*

The row and column gradings of the polynomial ring $\mathbb{Q}[P]$ are given by $deg(p_{ij}) = e_i$ and $deg(p_{ij}) = e_j$ where e_i and e_j are unit vectors in $\mathbb{Z}^m$ and $\mathbb{Z}^n$ respectively.

Example 2.1.27. *If $m = 5$ and $n = 6$ then our component is given by the parametrization*

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \\ a_{41} & a_{42} & 0 \\ a_{51} & a_{52} & a_{53} \end{bmatrix} \cdot \begin{bmatrix} 0 & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} \\ b_{21} & 0 & b_{23} & b_{24} & b_{25} & b_{26} \\ b_{31} & b_{32} & 0 & b_{34} & b_{35} & b_{36} \end{bmatrix}.$$

This parametrized variety has codimension 7 and degree 735 in $\mathbb{P}^{29}$. Its prime ideal is generated by 75 quartics and 20 sextics of the desired row and column degrees. ♢

The base case for Theorem 2.1.26 is the case of 4×3-matrices, even though $\partial\mathcal{M} = \mathcal{M}\cap\Delta_{11}$ is trivial in this case. The corresponding ideal is principal, and it is generated by the determinant of the 4×4-matrix that is obtained by deleting the fourth column of (2.1.22).

The sextics in Theorem 2.1.26 can be constructed as follows. Start with the polynomial

$$\big(((a_1 \wedge a_2) \vee b_1) \wedge a_3\big) \vee \big(((a_1 \wedge a_2) \vee b_2) \wedge a_4\big) \vee b_3$$

that is given in (2.1.16). Now multiply this with the 3×3-minor $b_i \vee b_j \vee b_k$ of B. The result has bidegree $(6, 6)$ in the parameters (A, B) and can be written as a sextic in $P = AB$. By construction, it vanishes on our component of $\overline{\partial\mathcal{M}}$, and it has the asserted degrees in the row and column gradings on $\mathbb{Q}[P]$. This is the generator of the prime ideal referred to in Theorem 2.1.26.

Theorem 2.1.23 characterizes the probability distributions in the algebraic boundary of our model, but not those in the topological boundary, since the following inclusion is strict:

$$\partial\mathcal{M} \;\subset\; \overline{\partial\mathcal{M}} \cap \Delta_{mn-1} \tag{2.1.23}$$

In fact, the left hand side is much smaller than the right hand side.

To quantify the discrepancy between the two semialgebraic sets in (2.1.23), we conducted the following experiment in the smallest interesting case $m = n = 4$. We sampled from the component (2.1.20) of $\overline{\partial\mathcal{M}} \cap \Delta_{15}$ by generating random rational numbers for the nine parameters a_{ij} and the eight parameters b_{ij}. This was done using the built-in `Macaulay2` function `random(QQ)`. The resulting matrix in $\overline{\partial\mathcal{M}} \cap \Delta_{15}$ was obtained by dividing by the sum of the entries. For each matrix we tested whether it lies in $\partial\mathcal{M}$. This was done using the criterion in Corollary 2.1.28. The answer was affirmative only in 257 cases out of 5000 samples. This suggests that $\partial\mathcal{M}$ occupies only a tiny part of the set $\overline{\partial\mathcal{M}} \cap \Delta_{15}$. One of those rare points in the topological boundary is the matrix

$$\begin{bmatrix} 6 & 13 & 3 & 1 \\ 4 & 16 & 6 & 2 \\ 12 & 4 & 8 & 12 \\ 5 & 9 & 10 & 9 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 3 \\ 1 & 0 & 4 \\ 4 & 4 & 0 \\ 4 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 2 & 2 \\ 3 & 1 & 0 & 1 \\ 1 & 4 & 1 & 0 \end{bmatrix}. \tag{2.1.24}$$

To construct this particular example, the parameters a_{ij} and b_{ij} were selected uniformly at random among the integers between 1 and 4. Only 1 out of 1000 samples gave a matrix lying in $\partial\mathcal{M}$. In fact, this matrix lies on precisely one of the 304 strata in the topological boundary $\partial\mathcal{M}$.

We close with a quantifier-free semialgebraic formula for the topological boundary.

Corollary 2.1.28. *An $m \times n$-matrix P lies on the topological boundary $\partial\mathcal{M}$ if and only if*

- *the conditions of Theorem 2.1.9 are satisfied, and*
- *P contains a zero, or $\operatorname{rank}(P) = 3$ and for each i, j, i', j' for which the conditions of Theorem 2.1.9 are satisfied there exist k, l such that (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] = 0$.*

This corollary will be derived (in Appendix 2.1.7.1) from our results in Subsection 2.1.4.

2.1.7 Appendix to Section 2.1

2.1.7.1 Proofs

This appendix furnishes the proofs for all lemmas, propositions and corollaries in this section.

Proof of Lemma 2.1.5. (3) $\Rightarrow$ (2): If (A, Λ, B) remains fixed after one completion of the E-step and the M-step, then it will remain fixed after any number of rounds of the E-step and the M-step.
(2) $\Rightarrow$ (3): By the proof of [122, Theorem 1.15], the log-likelihood function ℓ_U grows strictly after the completion of an E-step and an M-step unless the parameters (A, Λ, B) stay fixed, in which case ℓ_U also stays fixed. Thus, the only way to start with (A, Λ, B) and to end with it is for (A, Λ, B) to stay fixed after every completion of an E-step and an M-step.
(2) $\Rightarrow$ (1): If (A, Λ, B) is the limit point of EM when we start with it, then it is in the set of all limit points. This argument is reversible, and so we also get (1) $\Rightarrow$ (2), (3). $\square$

Proof of Lemma 2.1.11. The if-direction of the first sentence follows from the following two observations:

1. The function that takes $P \in \mathbb{R}^{m\times n}_{\geq 0}$ to the vertices of $\mathcal{B}$ is continuous on all $m \times n$ nonnegative matrices without zero columns, since the vertices of $\mathcal{B}$ are of the form P^j/P_{+j}, where P_{+j} denotes the j-th column sum of P.

2. The function that takes $P \in \mathbb{R}^{m\times n}_{\geq 0}$ to the vertices of $\mathcal{A}$ is continuous on all $m \times n$ nonnegative matrices of rank r, since the vertices of $\mathcal{A}$ are solutions to a system of linear equations in the entries of P.

For the only-if-direction of the first sentence assume that P lies in the interior of $\mathcal{M}_r$. Each P' of rank r in a small neighborhood of P has nonnegative rank r. We can choose P' in this neighborhood such that the columns of P' are in $\operatorname{span}(P)$ and $\operatorname{cone}(P') = t \cdot \operatorname{cone}(P)$ for some $t > 1$. Since P' has nonnegative rank r, there exists an $(r-1)$-simplex Δ such that

$\mathcal{B}' \subseteq \Delta' \subseteq \mathcal{A}$. Hence $\mathcal{B}$ is contained in the interior of Δ'. Finally, the second sentence is the contrapositive of the first sentence. □

Proof of Corollary 2.1.12. The if-direction follows from the second sentence of Lemma 2.1.11. For the only-if-direction, assume that $P \in \partial\mathcal{M}_3$ and it contains no zeros. We first consider the case $\text{rank}(P) = 3$. By Lemma 2.1.11, every triangle Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ contains a vertex of $\mathcal{B}$ on its boundary. Moreover, by the discussion above, every edge of Δ contains a vertex of $\mathcal{B}$, and (a) or (b) must hold. It remains to be seen that $\text{rank}(P) \leq 2$ is impossible on the strictly positive part of the boundary of $\mathcal{M}_3$. Indeed, for every rank 3 matrix P' in a neighborhood of P, the polygons $\mathcal{A}', \mathcal{B}'$ have the property that $\mathcal{B}'$ is very close to a line segment strictly contained in the interior of $\mathcal{A}'$. Hence, $t\mathcal{B}' \subseteq \Delta \subseteq \mathcal{A}'$ for some triangle Δ. Thus $P' \notin \partial\mathcal{M}_3$, and therefore $P \notin \partial\mathcal{M}_3$. □

Proof of Corollary 2.1.13. The if-direction is immediate. For the only-if direction, consider any $P \in \mathcal{M}_3$. If $P \in \partial\mathcal{M}_3$, then the only-if-direction follows from Corollary 2.1.12. If P lies in the interior of $\mathcal{M}_3$, then let t be maximal such that $t\mathcal{B} \subseteq \Delta' \subseteq \mathcal{A}$ for some triangle Δ'. Then either a vertex of Δ' coincides with a vertex of $\mathcal{A}$ or an edge of Δ' contains an edge of $t\mathcal{B}$. In the first case, we take $\Delta = \Delta'$. In the second case, we take $\Delta = \frac{1}{t}\Delta'$. In the first case, a vertex of Δ coincides with a vertex of $\mathcal{A}$, and in the second case, an edge of Δ contains an edge of $\mathcal{B}$. □

Proof of Corollary 2.1.14. If P has a nonnegative factorization of size 3, then it has one that corresponds to a geometric condition in Corollary 2.1.13. The left matrix in the factorization can be taken to be equal to the vertices of the nested triangle, which can be expressed as rational functions in the entries of P. Finally, the right matrix is obtained from solving a system of linear equations with rational coefficients, hence its entries are again rational functions in the entries of P. □

Proof of Proposition 2.1.15. Consider the sequence of linear maps

$$\mathbb{R}^r \xrightarrow{B^T} \mathbb{R}^n \xrightarrow{R} \mathbb{R}^m \xrightarrow{A^T} \mathbb{R}^r. \tag{2.1.25}$$

The ideal $\mathcal{C}$ says that the two compositions are zero. It defines a *variety of complexes* [114, Example 17.8]. The irreducible components of that variety correspond to *irreducible rank arrays* [114, §17.1] that fit inside the format (2.1.25) and are maximal with this property. By [114, Theorem 17.23], the quiver loci for these rank arrays are irreducible and their prime ideals are the ones we listed. These can also be described by *lacing diagrams* [114, Prop. 17.9].

The proof that $\mathcal{C}$ is radical was suggested to us by Allen Knutson. Consider the Zelevinski map [114, §17.2] that sends the triple (A^T, R, B^T) to the $(r+m+n+r) \times (r+m+n+r)$ matrix

$$\begin{bmatrix} 0 & 0 & B^T & 1 \\ 0 & R & 1 & 0 \\ A^T & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Next apply the map that takes this matrix to the big cell (the open Borel orbit) in the flag variety $GL(2r+m+n)/parabolic(r,m,n,r)$ corresponding to the given block structure.

Our scheme is identified with the intersection of two Borel invariant Schubert varieties. The first Schubert variety encodes the fact that there are 0's in the North West block, and the $(r+n+m) \times (r+m)$ North West rectangle has rank $\leq m$. The second Schubert variety corresponds to the $(r+n)\times(r+m+n)$ NorthWest rectangle having rank $\leq n$. The intersection of Schubert varieties is reduced by [28, §2.3.3, p.74]. Hence the original scheme is reduced, and we conclude that $\mathcal{C}$ is the radical ideal defining the variety of complexes (2.1.25). □

The following relations hold for $P = AB$ and R on the variety of critical points $V(\mathcal{C})$:

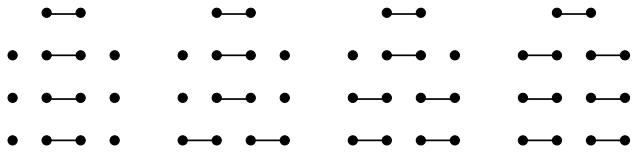
$$P^T \cdot R \,=\, 0 \quad \text{and} \quad R \cdot P^T \,=\, 0. \tag{2.1.26}$$

These bilinear equations characterize the *conormal variety* associated to a pair of determinantal varieties. Suppose P is fixed and has rank r. Then P is a nonsingular point in $\mathcal{V}$, and (2.1.26) is the system of linear equations that characterizes normal vectors R to $\mathcal{V}$ at P.

Example 2.1.29. *Let $m = n = 4$ and $r = 3$. Then $\mathcal{C}$ has four minimal primes, corresponding to the four columns in the table below. These are the ranks for generic points on that prime:*

$rank(A) = 0$	$rank(A) = 1$	$rank(A) = 2$	$rank(A) = 3$
$rank(R) = 4$	$rank(R) = 3$	$rank(R) = 2$	$rank(R) = 1$
$rank(B) = 0$	$rank(B) = 1$	$rank(B) = 2$	$rank(B) = 3$

The lacing diagrams that describe these four irreducible components are as follows:



For instance, the second minimal prime is $\mathcal{C} + \langle 2{\times}2\text{-minors of } A \text{ and } B \rangle + \langle det(R) \rangle$.

Note that the ranks of $P = AB$ and R are complementary on each irreducible component. They add up to 4. The last component gives the behavior of EM for random data: the MLE P has rank 3, it is a nonsingular point on the determinantal hypersurface $\mathcal{V}$, and the normal space at P is spanned by the rank 1 matrix R. This is the duality (2.1.26). The third component expresses the behavior on the singular locus of $\mathcal{V}$. Here the typical rank of both P and R is 2. ♢

Proof of Proposition 2.1.25. Let f, g_1, g_2, g_3, g_4 denote the $4{\times}4$ minors of the matrix (2.1.22), where $deg(f)$=4 and $deg(g_i) = 6$. Fix $i \in \{1,2,3,4\}$, select $u_{11}, \ldots, u_{44} \in \mathbb{N}$ randomly, and set

$$L \;=\; \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{44} \\ p_{11} & p_{12} & \cdots & p_{44} \\ p_{11}\,\partial f/\partial p_{11} & p_{12}\,\partial f/\partial p_{12} & \cdots & p_{44}\,\partial f/\partial p_{44} \\ p_{11}\,\partial g_i/\partial p_{11} & p_{12}\,\partial g_i/\partial p_{12} & \cdots & p_{44}\,\partial g_i/\partial p_{44} \end{bmatrix}. \tag{2.1.27}$$

This is a 4×16 matrix. Let λ_1 and λ_2 be new unknowns and consider the row vector

$$\begin{bmatrix} 1 & -u_+ & \lambda_1 & \lambda_2 \end{bmatrix} \cdot L. \tag{2.1.28}$$

Inside the polynomial ring $\mathbb{Q}[p_{ij}, \lambda_k]$ with 20 unknowns, let I denote the ideal generated by $\{f, g_1, g_2, g_3, g_4\}$, the 16 entries of (2.1.28), and the linear polynomial $p_{11}+p_{12}+\cdots+p_{44}-1$. Thus I is the ideal of *Lagrange likelihood equations* introduced in [82, Definition 2]. Gross and Rodriguez [82, Proposition 3] showed that I is a 0-dimensional radical ideal, and its number of roots is the ML degree of the variety $V(f, g_1, g_2, g_3, g_4)$. We computed a Gröbner bases for I using the computer algebra software `Magma` [158]. This computation reveals that $V(I)$ consists of 633 points over $\mathbb{C}$. □

Proof of Corollary 2.1.28. A matrix P has nonnegative rank 3 if and only if the conditions of Theorem 2.1.9 are satisfied. Assume $\text{rank}(P) = 3$. By Corollary 2.1.12, a matrix $P \in \mathcal{M}$ lies on the boundary of $\mathcal{M}$ if and only if it contains a zero or for any triangle Δ with $\mathcal{B} \subseteq \Delta \subseteq \mathcal{A}$ every edge of Δ contains a vertex of $\mathcal{B}$ and (a) or (b) holds. By proof of Theorem 2.1.9, the latter implies that for each i, j, i', j' for which the conditions of Theorem 2.1.9 are satisfied there exist k, l such that (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] = 0$. On the other hand, if P lies in the interior of $\mathcal{M}_3^{m\times n}$, then by the proof of Corollary 2.1.13, the following holds: there exists a triangle Δ with a vertex coinciding with a vertex of $\mathcal{A}$ or with an edge containing an edge of $\mathcal{B}$, and such that the inequality (2.1.16) $\cdot$ (2.1.16)$[k \leftrightarrow l] > 0$ holds for all k, l in the corresponding semialgebraic condition. □

2.1.7.2 Basic Concepts in Algebraic Geometry

This appendix gives a synopsis of basic concepts from algebraic geometry that are used in this section. It furnishes the language to speak about solutions to polynomial equations in many variables.

2.1.7.3 Ideals and Varieties

Let $R = K[x_1, \ldots, x_n]$ be the ring of polynomials in n variables with coefficients in a subfield K of the real numbers $\mathbb{R}$, usually the rational numbers $K = \mathbb{Q}$. The concept of an ideal I in the ring R is similar to the concept of a normal subgroup in a group.

Definition 2.1.30. *A subset $I \subseteq R$ is an* ideal *in R if I is an subgroup of R under addition, and for every $f \in I$ and every $g \in R$ we have $fg \in I$. Equivalently, an ideal I is closed under taking linear combinations with coefficients in the ring R.*

Let T be any set of polynomials in R. Their set of zeros is called the *variety* of T. It is denoted

$$V(T) \;=\; \bigl\{P \in \mathbb{C}^n : f(P) = 0 \text{ for all } f \in T\bigr\}.$$

Here we allow zeros with complex coordinates. This greatly simplifies the study of $V(T)$ because $\mathbb{C}$ is algebraically closed, *i.e.* every non-constant polynomial has a zero.

The *ideal generated by* T, denoted by $\langle T \rangle$, is the smallest ideal in R containing T. Note that

$$V(T) \;=\; V(\langle T \rangle).$$

In computational algebra, it is often desirable to replace the given set T by a *Gröbner basis* of $\langle T \rangle$. This allows us to test ideal membership and to determine geometric properties of the variety $V(T)$.

Definition 2.1.31. *A subset $X \subseteq \mathbb{C}^n$ is a* variety *if $X = V(T)$ for some subset $T \subseteq R$.*

Hilbert's Basis Theorem ensures that here T can always be chosen to be a finite set of polynomials. The concept of variety allows us to define a new topology on $\mathbb{C}^n$. It is coarser than the usual topology.

Definition 2.1.32. *We define the* Zariski topology *on $\mathbb{C}^n$ by taking closed sets to be the varieties and open sets to be the complements of varieties. This topology depends on the choice of K.*

If $K = \mathbb{Q}$ then $X = \{+\sqrt{2}, -\sqrt{2}\}$ is a variety (for $n = 1$) but $Y = \{+\sqrt{2}\}$ is not a variety. Indeed, $X = \overline{Y}$ is the *Zariski closure* of Y, *i.e.* it is the smallest variety containing Y, because the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$. Likewise, the set of 1618 points in Example 2.1.2 is a variety in $\mathbb{C}^2$. It is the Zariski closure over $\mathbb{Q}$ of the four points on the topological boundary on the left in Figure 2.2. The following proposition justifies the fact that the Zariski topology is a topology.

Proposition 2.1.33. *Varieties satisfy the following properties:*

1. *The empty set $\emptyset = V(R)$ and the whole space $\mathbb{C}^n = V(\langle 0 \rangle)$ are varieties.*
2. *The union of two varieties is a variety:*

$$V(I) \cup V(J) \;=\; V(I \cdot J) \;=\; V(I \cap J).$$

3. *The intersection of any family of varieties is a variety:*

$$\bigcap_{i \in \mathcal{I}} V(I_i) \;=\; V(\langle I_i : i \in \mathcal{I} \rangle).$$

Given any subset $X \subseteq \mathbb{C}^n$ (not necessarily a variety), we define the *ideal* of X by

$$I(X) \;=\; \{f \in R : f(P) = 0 \text{ for all } P \in X\}.$$

Thus, $I(X)$ consists of all polynomials in R that vanish on X. The *Zariski closure* $\overline{X}$ of X equals

$$\overline{X} \;=\; V(I(X)).$$

2.1.7.4 Irreducible Decomposition

A variety $X \subseteq \mathbb{C}^n$ is *irreducible* if we cannot write $X = X_1 \cup X_2$, where $X_1, X_2 \subsetneq X$ are strictly smaller varieties. An ideal $I \subseteq R$ is *prime* if $fg \in I$ implies $f \in I$ or $g \in I$. For instance, $I(\{\pm\sqrt{2}\}) = \langle x^2 - 2 \rangle$ is a prime ideal in $\mathbb{Q}[x]$.

Proposition 2.1.34. *The variety X is irreducible if and only if $I(X)$ is a prime ideal.*

An ideal is *radical* if it is an intersection of prime ideals. The assignment $X \mapsto I(X)$ is a bijection between varieties in $\mathbb{C}^n$ and radical ideals in R. Indeed, every variety X satisfies $V(I(X)) = X$.

Proposition 2.1.35. *Every variety X can be written uniquely as $X = X_1 \cup X_2 \cup \cdots \cup X_m$, where $X_1, X_2, \ldots, X_m$ are irreducible and none of these m components contains any other. Moreover,*

$$I(X) \;=\; I(X_1) \cap I(X_2) \cap \cdots \cap I(X_m)$$

is the unique decomposition of the radical ideal $I(X)$ as an intersection of prime ideals.

For an explicit example, with $m = 11$, we consider the ideal (2.1.13) with the last intersectant removed. In that example, the EM fixed variety X is decomposed into 11 irreducible components.

All ideals I in R can be written as intersections of *primary ideals*. Primary ideals are more general than prime ideals, but they still define irreducible varieties. A *minimal prime* of an ideal I is a prime ideal J such that $V(J)$ is an irreducible component of $V(I)$. See [147, Chapter 5] for the basics on *primary decomposition*.

Definition 2.1.36. *Let $I \subseteq R$ be an ideal and $f \in R$ a polynomial. The saturation of I with respect to f is the ideal*

$$(I : f^\infty) = \langle g \in R : gf^k \in I \text{ for some } k > 0 \rangle.$$

Saturating an ideal I by a polynomial f geometrically means that we obtain a new ideal $J = (I : f^\infty)$ whose variety $V(J)$ contains all components of the variety $V(I)$ except for the ones on which f vanishes. For the more on these concepts from algebraic geometry we recommend the text [46].

2.1.7.5 Semialgebraic Sets

The discussion above also applies if we consider the varieties $V(T)$ as subsets of $\mathbb{R}^n$ instead of $\mathbb{C}^n$. This brings us to the world of *real algebraic geometry*. The field $\mathbb{R}$ of real numbers is not algebraically closed, it comes with a natural order, and it is fundamental for applications. These features explain why real algebraic geometry is a subject in its own right. In addition to the polynomial equations we discussed so far, we can now also introduce inequalities:

Definition 2.1.37. *A* basic semialgebraic set $X \subseteq \mathbb{R}^n$ *is a subset of the form*

$$X \;=\; \big\{P \in \mathbb{R}^n :\; f(P) = 0 \text{ for all } f \in T \text{ and } g(P) \geq 0 \text{ for all } g \in S\big\},$$

where S *and* T *are finite subsets of* R*. A* semialgebraic set *is a subset* $X \subseteq \mathbb{R}^n$ *that is obtained by a finite sequence of unions, intersections, and complements of basic semialgebraic sets.*

In other words, semialgebraic sets are described by finite Boolean combinations of polynomial equalities and polynomial inequalities. For basic semialgebraic sets, only conjunctions are allowed. For example, the following two simple subsets of the plane are both semialgebraic:

$$X = \big\{(x,y) \in \mathbb{R}^2 : x \geq 0 \text{ and } y \geq 0\big\} \quad \text{and} \quad Y = \big\{(x,y) \in \mathbb{R}^2 : x \geq 0 \text{ or } y \geq 0\big\}.$$

The set X is basic semialgebraic, but Y is not. All convex polyhedra are semialgebraic. A fundamental theorem due to Tarski states that the image of a semialgebraic set under a polynomial map is semialgebraic. Applying this to the map (2.1.2), we see that the model $\mathcal{M}$ is semialgebraic. The boundary of any semialgebraic set is again semialgebraic. The formulas in Theorem 2.1.9 and Corollary 2.1.28 make this explicit. For more on semialgebraic sets and real algebraic geometry see [13].

Acknowledgements

My coauthors and I would like to thank Aldo Conca, Allen Knutson, Pierre-Jean Spaenlehauer, and Matteo Varbaro for helping us with this project. Mathias Drton, Sonja Petrović, John Rhodes, Caroline Uhler, and Piotr Zwiernik provided comments on various drafts of the paper. We thank Christopher Miller for pointing out an inaccuracy in Example 2.1.2.

2.2 Positive Semidefinite Rank

The set of matrices of given positive semidefinite rank is semialgebraic. In this section we study the geometry of this set, and in small cases we describe its boundary. Furthermore, for general values of the positive semidefinite rank, we give a conjecture for the description of this boundary. Our proof techniques are geometric in nature. As in the previous section, we think of nonnegative matrices as slack matrices of pairs of nested polyhedra, and we interpret positive semidefinite rank via the existence of nested spectrahedral shadows between these polyhedra. This section is based on joint work with Kaie Kubjas and Richard Robinson titled *Positive Semidefinite Rank and Nested Spectrahedra* [104].

2.2.1 Introduction

Standard matrix factorization is used in a wide range of applications in statistics, optimization, machine learning, and others. Given a $p \times q$ real matrix $M \in \mathbb{R}^{p\times q}$ of rank r, the goal is to find vectors $a_1, ..., a_p, b_1, ..., b_q \in \mathbb{R}^r$ such that the i,j-th entry of M is $M_{ij} = \langle a_i, b_j\rangle$.

Often times, however, the matrix at hand as well as the elements in the factorization are imposed certain positivity structure [64, 76, 77]. In statistical mixture models, for instance, we need to find a nonnegative factorization of a matrix M with nonnegative entries [43, 75, 105, 156]. In other words, the vectors a_i and b_j need to be nonnegative. Another type of factorization of a matrix with nonnegative entries, which has applications in convex optimization and quantum information theory, is positive semidefinite factorization. The vectors a_i and b_j are now replaced by $k \times k$ symmetric positive semidefinite matrices $A_i, B_j \in \mathcal{S}^k_+$. Here the space of symmetric $k \times k$ matrices is denoted by $\mathcal{S}^k$, the cone of $k \times k$ positive semidefinite matrices by $\mathcal{S}^k_+$, and the inner product on $\mathcal{S}^k$ is given by

$$\langle A, B\rangle = \text{trace}(AB).$$

Definition 2.2.1. *Given a matrix $M \in \mathbb{R}^{p\times q}_{\geq 0}$ with nonnegative entries, a* positive semidefinite (psd) factorization *of size k is a collection of matrices $A_1, ..., A_p, B_1, ..., B_q \in \mathcal{S}^k_+$ such that $M_{ij} = \langle A_i, B_j\rangle$. The* positive semidefinite rank (*or* psd rank) *of the matrix M is the smallest number k for which such a factorization exists. It is denoted by* $\text{rank}_{psd}(M)$.

We remark that, given two psd matrices $A, B \in \mathcal{S}^k_+$, it is always the case that $\langle A, B\rangle \geq 0$, which is why the entries of the matrix M need to be nonnegative.

The geometric aspects as well as many of the properties of positive semidefinite rank have been studied in a number of recent articles [65, 76, 77, 78, 79, 80].

In this section we study the space $\mathcal{P}^{p\times q}_{r,k}$ of $p \times q$ nonnegative real matrices of rank at most r and psd rank at most k. If p and q are understood from the context, we write $\mathcal{P}_{r,k}$ for short. By Tarski-Seidenberg's Theorem [13, Theorem 2.76] this set is semialgebraic, i.e. it is defined by finitely many polynomial equations and inequalities, or is a finite union of such sets. It lies inside the variety of $p \times q$ matrices of rank at most r, denoted by $\mathcal{V}^{p\times q}_r$ (for

short $\mathcal{V}_r$). In this section, we study the geometry of $\mathcal{P}_{r,k}$. In particular, we investigate the *boundary* of $\mathcal{P}_{r,k}$ as a subset of $\mathcal{V}_r$.

Definition 2.2.2. *The* topological boundary *of $\mathcal{P}_{r,k}$, denoted by $\partial\mathcal{P}_{r,k}$, is its boundary as a subset of $\mathcal{V}_r$. In other words, it consists of all matrices $M \in \mathcal{V}_r$ such that for every $\epsilon > 0$, the ball with radius ϵ and center M, denoted by $\mathcal{B}_\epsilon(M)$, satisfies the condition that $\mathcal{B}_\epsilon(M) \cap \mathcal{V}_r$ intersects both $\mathcal{P}_{r,k}$ and its complement $\mathcal{V}_r \setminus \mathcal{P}_{r,k}$. The* algebraic boundary *of $\mathcal{P}_{r,k}$, denoted by $\overline{\partial\mathcal{P}_{r,k}}$ is the Zariski closure of $\partial\mathcal{P}_{r,k}$ over $\mathbb{R}$.*

In this section we focus on studying $\partial\mathcal{P}^{p\times q}_{k+1,k}$. In other words, we restrict to the case when the rank of our matrix is 1 more than the psd rank. In Subsection 2.2.3, we completely describe $\partial\mathcal{P}^{p\times q}_{3,2}$, as well as $\overline{\partial\mathcal{P}^{p\times q}_{3,2}}$. More precisely, Corollary 2.2.13 shows that a matrix M lies on the boundary $\partial\mathcal{P}^{p\times q}_{3,2}$ if and only if in every psd factorization $M_{ij} = \langle A_i, B_j\rangle$, at least three of the matrices $A_1, \ldots, A_p$ have rank 1 and at least three of the matrices $B_1, \ldots, B_q$ have rank 1.

In Subsections 2.2.4 and 2.2.5 we study the general case $\mathcal{P}^{p\times q}_{k+1,k}$, and we attempt to extend our results from the $k = 2$ case. We restrict ourselves to the simplest situation where $p = q = k+1$. Conjecture 2.2.16 is an analogue to Corollary 2.2.13. It states that a matrix M lies on the boundary $\partial\mathcal{P}^{(k+1)\times(k+1)}_{k+1,k}$ if and only if for every psd factorization $M_{ij} = \langle A_i, B_j\rangle$, all of the matrices $A_1, \ldots, A_{k+1}$ have rank 1 and all of the matrices $B_1, \ldots, B_{k+1}$ have rank 1. In Subsection 2.2.5 we give theoretical and computational evidence supporting this conjecture. The code for our computations is available at

`https://github.com/kaiekubjas/psd-rank` .

Our results are based on the geometric interpretation of psd rank explained in Subsection 1.1. We review this interpretation once again in Subsection 2.2.2. Given a nonnegative matrix M of rank $n + 1$, we can associate to it two nested polytopes $P \subseteq Q \subset \mathbb{R}^n$. Theorem 2.2.4, proven in [79], shows that M has psd rank at most k if and only if we can fit a projection of a slice of the cone of $k \times k$ positive semidefinite matrices $\mathcal{S}^k_+$ between P and Q. When we restrict to the case when the rank of M is 3, this seemingly sophisticated result states that M has psd rank 2 if and only if we can nest an ellipse between the two nested polygons P and Q associated to M. In Theorem 2.2.12 we show that M lies on the boundary $\partial\mathcal{P}^{p\times q}_{3,2}$ if and only if every ellipse that nests between the two polygons P and Q has to touch at least three of the vertices of P and at least three of the edges of Q. In Conjecture 2.2.18 we give an analogue to Theorem 2.2.12 for the general case $\partial\mathcal{P}^{(k+1)\times(k+1)}_{k+1,k}$.

2.2.2 Preliminaries

Many of the basic properties of positive semidefinite rank have been studied in [65]. We give a brief overview of the results used in the present section.

2.2.2.1 Bounds

The positive semidefinite rank of a matrix is bounded below by the inequality

$$\text{rank}(M) \leq \binom{\text{rank}_{\text{psd}}(M)+1}{2}.$$

This holds because we can vectorize the symmetric matrices in a given psd factorization and consider the trace inner product as a dot product. On the other hand, the psd rank is upper bounded by the nonnegative rank

$$\text{rank}_{psd}(M) \leq \text{rank}_{+}(M)$$

since we can obtain a psd factorization from a nonnegative factorization by using diagonal matrices. The psd rank of M can be any integer satisfying these inequalities.

2.2.2.2 Geometric description

We now describe the geometric interpretation of psd rank. Let $P \subseteq \mathbb{R}^n$ be a polytope and $Q \subseteq \mathbb{R}^n$ be a polyhedron such that $P \subseteq Q$. Assume that P is the convex hull of p points: $P = \text{conv}\{v_1, ..., v_p\}$ and Q has the following inequality representation: $Q = \{x \in \mathbb{R}^n | h_j^T x \leq z_j, j = 1, ..., q\}$, where $v_1, ..., , v_p, h_1, ..., h_q \in \mathbb{R}^n$ and $z_1, \ldots, z_q \in \mathbb{R}$. Then, the *generalized slack matrix* of the pair P, Q, denoted $S_{P,Q}$ is the $p \times q$ matrix whose i, j-th entry is $z_j - h_j^T v_i$.

Remark 2.2.3. *The generalized slack matrix depends on the representations of P and Q as the convex hull of finitely many points and as the intersection of finitely many halfspaces whereas the slack matrix depends only on P and Q. We will abuse the notation and write $S_{P,Q}$ for the generalized slack matrix as by the next result the* $\text{rank}_{psd}(S_{P,Q})$ *is independent of the representations of P and Q.*

Theorem 2.2.4 (Proposition 3.6 in [79])**.** *Let $P \subset \mathbb{R}^n$ be a polytope and $Q \subseteq \mathbb{R}^n$ a polyhedron such that $P \subseteq Q$. Then,* $\text{rank}_{psd}(S_{P,Q})$ *is the smallest integer k for which there exists an affine subspace $\mathcal{L}$ of $\mathcal{S}^k$ and a linear map π such that $P \subseteq \pi(\mathcal{L} \cap \mathcal{S}_+^k) \subseteq Q$.*

A *spectrahedron* of size k is a slice of the cone of $k \times k$ positive semidefinite matrices $\mathcal{S}_+^k$. A *spectrahedral shadow* of size k is a projection of a spectrahedron of size k. Therefore, Theorem 2.2.4 states that the matrix $S_{P,Q}$ has psd rank at most k if and only if we can fit a spectrahedral shadow of size k between P and Q.

Remark 2.2.5. *Given M the polytopes P and Q are not unique, but the statement of Theorem 2.2.4 still holds regardless of which pair P, Q, such that $M = S_{P,Q}$, is chosen.*

Conversely, given a matrix M, after rescaling the rows of M (which doesn't change its psd rank), we can find polytopes P and Q such that M is their generalized slack matrix.

Lemma 2.2.6 (Lemma 4.1 in [65]). *Let $M \in \mathbb{R}^{p\times q}_{\geq 0}$ be a nonnegative matrix and assume that $M\mathbf{1} = \mathbf{1}$. Let $\operatorname{rank}(M) = n + 1$. Then, there exist polytopes $P, Q \subseteq \mathbb{R}^n$ (where P and Q are bounded) such that $P \subseteq Q$ and M is the generalized slack matrix of the pair P, Q.*

We define the *interior* of $\mathcal{P}^{p\times q}_{d,k}$ to be the set of matrices $M \in \mathcal{P}^{p\times q}_{d,k}$ for which there exists $\epsilon > 0$ such that $\mathcal{V}^{p\times q}_d \cap \mathcal{B}_\epsilon(M) \subset \mathcal{P}^{p\times q}_{d,k}$, where $\mathcal{B}_\epsilon(M)$ is the ball of radius ϵ centered M. We make the following observation.

Lemma 2.2.7. *Let $M \in \mathbb{R}^{p\times q}_{>0}$ be a matrix with positive entries. The following are equivalent*

1. *M lies in the interior of $\mathcal{P}_{r,k}$;*
2. *When we rescale the rows of M so as to obtain a matrix N that satisfies $N\mathbf{1} = \mathbf{1}$, the matrix N lies in the interior of $\mathcal{P}_{r,k} \cap \{P : P\mathbf{1} = \mathbf{1}\}$ (in other words, there exists $\epsilon > 0$ such that $\mathcal{B}_\epsilon(N) \cap \mathcal{V}_r \cap \{P : P\mathbf{1} = \mathbf{1}\} \subseteq \mathcal{P}_{r,k} \cap \{P : P\mathbf{1} = \mathbf{1}\}$).*

Lemma 2.2.7, whose proof can be found in Subsection 2.2.6.1, implies that if we want to study the boundary of $\mathcal{P}_{r,k}$ as a subset of $\mathcal{V}_r$, we can restrict ourselves to the boundary of the space $\mathcal{P}_{r,k} \cap \{P : P\mathbf{1} = \mathbf{1}\}$ as a subset of $\mathcal{V}_r \cap \{P : P\mathbf{1} = \mathbf{1}\}$, and Lemma 2.2.6 gives us a recipe for thinking of the elements of this space geometrically.

2.2.2.3 Comparison with nonnegative rank

Three different versions of nonnegative matrix factorization appear in the literature: In [156] Vavasis considered the exact nonnegative factorization which asks whether a nonnegative matrix M has nonnegative factorization of size equal to the rank of M. The geometric version of this question asks whether we can nest a simplex between the polytopes P and Q.

In [75] Gillis and Glineur defined restricted nonnegative rank as the minimum value r such that there exist $A \in \mathbb{R}^{p\times r}_+$ and $B \in \mathbb{R}^{r\times q}_+$ with $M = AB$ and $\operatorname{rank}(A) = \operatorname{rank}(M)$. The geometric interpretation of the restricted nonnegative rank asks for the minimal r such that there exist r points whose convex hull can be nested between P and Q.

The geometric version of the nonnegative rank factorization asks for the minimal r such that there exist r points whose convex hull can be nested between an $(r-1)$-dimensional polytope inside a q-simplex. These polytopes are not P and Q as defined in this section. See [43, Theorem 3.1] for details.

In the positive semidefinite rank case there is no distinction between the psd rank and the restricted psd rank, because taking an intersection with a subspace does not change the size of a spectrahedral shadow while intersecting a polytope with a subspace can change the number of vertices. Conjecture 2.2.25 also suggests that there is no distinction between the spectrahedron and the spectrahedral shadow case. This is not the case with simplices and polytopes in the nonnegative rank case, or equivalently the exact nonnegative matrix factorization and restricted nonnegative factorization.

2.2.3 Matrices of rank 3 and psd rank 2

In this subsection we study the set $\mathcal{P}_{3,2}$ of matrices of rank at most 3 and psd rank at most 2. Rather than providing a semialgebraic description of $\mathcal{P}_{3,2}$, we completely characterize its topological and algebraic boundaries $\partial\mathcal{P}_{3,2}$ and $\overline{\partial\mathcal{P}_{3,2}}$.

Consider a matrix $M \in \mathbb{R}^{p\times q}_{\geq 0}$ of rank 3. We get two nested polygons $P \subseteq Q \subset \mathbb{R}^2$. Theorem 2.2.4 now has the following simpler form.

Corollary 2.2.8 (Proposition 4.1 in [79])**.** *Let M be a nonnegative matrix of rank three such that $M\mathbf{1} = \mathbf{1}$. Let $P \subseteq Q \subseteq \mathbb{R}^2$ be two nested polygons for which $M = S_{P,Q}$. Then $\text{rank}_{psd}(M) = 2$ if and only if we can fit an ellipse between P and Q.*

Using this geometric interpretation of psd rank 2, we give a condition on when a matrix M lies in the interior of $\mathcal{P}_{3,2}$.

Lemma 2.2.9. *A matrix $M \in \mathbb{R}^{p\times q}_{\geq 0}$ of rank 3 lies in the interior of $\mathcal{P}_{3,2}$ if and only if there exist polygons $P \subset Q \subseteq \mathbb{R}^2$ and an ellipse E such that M is the generalized slack matrix of P and Q, $P \subset E \subset Q$, and the boundary of E does not contain any of the vertices of P.*

The proof of this lemma can be found in Subsection 2.2.6.2. We can now show how $\mathcal{P}_{3,2}$ relates to the variety $\mathcal{V}_3$.

Proposition 2.2.10. *The Zariski closure of $\mathcal{P}^{p\times q}_{3,2}$ over the real numbers is the rank-3 variety $\mathcal{V}^{p\times q}_3$.*

Proof. Suppose there exists a ball $\mathcal{B} \subseteq \mathbb{R}^{p\times q}$ such that $\mathcal{B} \cap \mathcal{V}_3 \subseteq \mathcal{P}_{3,2}$. This implies that the dimension of $\mathcal{P}^{p\times q}_{3,2}$ is equal to that of $\mathcal{V}^{p\times q}_3$, and since $\mathcal{P}^{p\times q}_{3,2} \subset \mathcal{V}^{p\times q}_3$ and $\mathcal{V}^{p\times q}_3$ is irreducible, the Zariski closure of $\mathcal{P}_{3,2}$ over the real numbers equals $\mathcal{V}_3$.

We show how to find such a ball $\mathcal{B}$. It suffices to find a matrix M in the interior of $\mathcal{P}^{p\times q}_{3,2}$. By Lemma 2.2.9, it would suffice to find nested polygons $P \subseteq Q \subseteq \mathbb{R}^2$ such that P has p vertices, Q has q sides and there exists an ellipse nested between them that does not touch the vertices of P. Such a configuration certainly exists, for example, we can consider a regular p-gon P centered at the origin with length 1 from the origin to any of its vertices, and a regular q-gon Q centered at the origin with length 5 from the origin to any of its sides. Then, we can fit a circle of radius 2 and center the origin between P and Q so that it doesn't touch the vertices of P. □

Remark 2.2.11. *The set of $p \times q$ matrices of psd rank at most k is connected as it is the image under the parametrization map of the connected set $(\mathcal{S}^k_+)^p \times (\mathcal{S}^k_+)^q$.*

The following theorem is the main result of this section.

Theorem 2.2.12. *We describe the topological and algebraic boundaries of $\mathcal{P}^{p\times q}_{3,2}$.*

a. *A matrix $M \in \mathcal{P}_{3,2}^{p\times q}$ lies on the topological boundary $\partial\mathcal{P}_{3,2}^{p\times q}$ if and only if $M_{ij} = 0$ for some i, j, or each ellipse that fits between the two polygons P and Q contains at least 3 vertices of the inner polygon P and is tangent to at least 3 edges of the outer polygon Q.*

b. *A matrix $M \in \overline{\mathcal{P}_{3,2}^{p\times q}} = \mathcal{V}_3$ lies on the algebraic boundary $\overline{\partial\mathcal{P}_{3,2}^{p\times q}}$ if and only if $M_{ij} = 0$ for some i, j or there exists an ellipse that contains at least three vertices of P and is tangent to at least three edges of Q.*

c. *The algebraic boundary of $\mathcal{P}_{3,2}^{p\times q}$ is the union of $\binom{p}{3}\binom{q}{3} + pq$ irreducible components. Besides the pq components $M_{ij} = 0$, there are $\binom{p}{3}\binom{q}{3}$ components each of which is defined by the 4×4 minors of M and one additional polynomial equation with 1035 terms homogeneous of degree 24 in the entries of M and homogeneous of degree 8 in each row and each column of a 3×3 submatrix of M.*

Proof.
(a) Only if: We will show the contrapositive of the statement: If all entries of M are positive and there is an ellipse between P and Q whose boundary contains at most two vertices of P or is tangent to at most two edges of Q, then M lies in the interior of $M_{3,2}^{p\times q}$.

First, if there is an ellipse E between P and Q whose boundary touches neither of the polytopes, then M is in the interior of $\mathcal{P}_{3,2}^{p\times q}$ by Lemma 2.2.9. If at most two edges of Q are tangent to the boundary of the ellipse E, then $P \subset E \subset Q$ can be transformed by a projective transformation such that the two tangent facets are $x = 0$ and $y = 0$ and that the points of tangency are $(0, 1)$ and $(1, 0)$. Now, the equation of the ellipse E has the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$. We know that the only point that lies on the ellipse E with $x = 0$ is the point $(0, 1)$ since E touches the line $x = 0$ at $(0, 1)$. If we plug in $x = 0$, we get

$$cy^2 + ey + f = 0.$$

Since $c > 0$, we must have $cy^2 + ey + f = (y - 1)^2$. Therefore, $c = 1, e = -2, f = 1$. Similarly, since E touches the line $y = 0$ at $(1, 0)$, when we plug in $y = 0$, we get that $ax^2 + dx + f = (x - 1)^2$, so, $a = 1, d = -2, f = 1$. Thus, the ellipse E has the form

$$\{(x, y) : x^2 + bxy + y^2 - 2x - 2y + 1 = 0\},$$

for some b. The values of b for which this is an ellipse are $-2 < b < 2$. Moreover, if we choose a slightly smaller value of b in this family, we would obtain a slightly larger ellipse E' that contains E and touches E only at the points $(1, 0)$ and $(0, 1)$. Thus, we would have $P \subseteq E \subset E' \subseteq Q$ and the ellipse E' does not touch P. Thus, by Lemma 2.2.9, M lies in the interior of $\mathcal{P}_{3,2}^{p\times q}$. The case when E goes through at most two vertices of P follows by duality.

If: By Lemma 2.2.9, if $M \in \mathcal{P}_{3,2}$ lies in the interior, then there is an ellipse between P and Q that does not touch P. Thus, if every ellipse nested between P and Q contains at least three of the vertices of P and touches at least three of the facets of Q, then M lies on the boundary $\partial\mathcal{P}_{3,2}$

$(b), (c)$ If $M \in \mathbb{R}^{p\times q}$, then one can define polytopes P and Q as in the nonnegative case. The difference is that $P \subseteq Q$ does not hold anymore. Hence given three points a, b, c in $\mathbb{P}^2$ and three lines d, e, f in $\mathbb{P}^2$, each given by three homogeneous coordinates, we seek the condition that there exists an conic X such that a, b, c lie on X and d, e, f are tangent to X.

Let $X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{12} & x_{22} & x_{23} \\ x_{13} & x_{23} & x_{33} \end{bmatrix}$ be the matrix of a conic. Then the corresponding ellipse goes through the points a, b, c if and only if

$$a^T X a = b^T X b = c^T X c = 0. \tag{2.2.1}$$

Similarly, the lines d, e, f are tangent to the ellipse if and only

$$d^T Y d = e^T Y e = f^T Y f = 0, \tag{2.2.2}$$

where $XY = I_3$. We seek to eliminate the variables X and Y from (2.2.1) and (2.2.2).

Let $[a, b, c]$ denote the matrix whose columns are a, b, c. First we assume that $[a, b, c]$ is the 3×3-identity matrix. Then we proceed in two steps:

1) The equations (2.2.1) imply that x_{11}, x_{22}, x_{33} are zero. We make the corresponding replacements in equations (2.2.2).

2) We use [147, formula (4.5) on page 48] for the resultant of three ternary quadrics to get a single polynomial in the entries of d, e, f.

Now we use invariant theory to obtain the desired polynomial in the general case. Let $g \in \mathrm{GL}_3(\mathbb{R})$. The ellipse X goes through the points a, b, c and touches the lines d, e, f if and only if the ellipse $g^{-T} X g^{-1}$ goes through the points ga, gb, gc and touches the lines $g^{-T}d, g^{-T}e, g^{-T}f$. Thus our desired polynomial belongs to the ring of invariants $\mathbb{R}[V^3 \oplus V^{*3}]^{\mathrm{GL}_3(\mathbb{R})}$ where $V = \mathbb{R}^3$ and the action of $\mathrm{GL}_3(\mathbb{R})$ on $V^3 \oplus V^{*3}$ is given by

$$g \cdot (a, b, c, d, e, f) := (ga, gb, gc, g^{-T}d, g^{-T}e, g^{-T}f).$$

The First Fundamental Theorem states that $\mathbb{R}[V^3 \oplus V^{*3}]^{\mathrm{GL}_3(\mathbb{R})}$ is generated by the bilinear functions $(i|j)$ on $V^3 \oplus V^{*3}$ defined by

$$(i|j) : (a, b, c, d, e, f) \mapsto ([a, b, c]^T [d, e, f])_{ij}.$$

For the FFT see for example [102, Chapter 2.1]. In the special case when $[a, b, c]$ is the 3×3 identity matrix, $(i|j)$ maps to the (i, j)-th entry of $[d, e, f]$. Hence to obtain the desired polynomial in the general case, we replace in the resultant obtained in the special case the entries of the matrix $[d, e, f]$ by the entries of the matrix $[a, b, c]^T [d, e, f]$.

`Maple` code for doing the steps in the previous paragraphs can be found at our website. This program outputs one polynomial of degree 24 with 1035 terms. More precisely, this polynomial is homogeneous of degree 8 in each of the rows and the columns of the matrix $\begin{bmatrix} - & a & - \\ - & b & - \\ - & c & - \end{bmatrix} \begin{bmatrix} | & | & | \\ d & e & f \\ | & | & | \end{bmatrix}$. By construction, if this homogeneous polynomial vanishes and the

convex hull of a, b, c lies inside the triangle with edges d, e, f and a, b, c, d, e, f are real, then there exists an ellipse nested between the polytopes touching d, e, f and containing a, b, c. Therefore, the Zariski closure of the condition that the only possible ellipses that can fit between the two polygons touch at least 3 edges of the outer polygon and at least 3 vertices of the inner polygon is exactly that there exists an ellipse that touches at least 3 edges of Q and at least 3 vertices of P. This proves (b).

To prove (c), let $M \in \mathcal{V}_3^{p\times q}$ be such that $M = AB$ and a, b, c are three of the rows of A and d, e, f are three of the columns of B. Then, the above-computed polynomial contains variables only from the entries of a 3×3 submatrix of M corresponding to these rows and columns. For each three rows of and three columns of M we have one such polynomial, so the algebraic boundary is given by the union over each 3 rows and 3 columns of M of the variety defined by the 4×4 minors of M. The corresponding polynomial has degree 24 and 1035 terms. □

Here is an algebraic version of Theorem 2.2.12.

Corollary 2.2.13. *A matrix $M \in \mathbb{R}^{p\times q}_{\geq 0}$ lies on the boundary $\partial\mathcal{P}_{3,2}$ if and only if for every size 2 psd factorization $M_{ij} = \langle A_i, B_j\rangle$, at least three of the matrices $A_1, \ldots, A_p \in \mathcal{S}^k_+$ have rank one and at least three of the matrices $B_1, \ldots, B_q \in \mathcal{S}^k_+$ have rank one.*

We now investigate the topological boundary more thoroughly.

Proposition 2.2.14. *Suppose $M \in \mathcal{M}^{p\times q}_{3,2}$ is strictly positive. Then M lies on the topological boundary if and only if there exists a unique ellipse that nests between P and Q.*

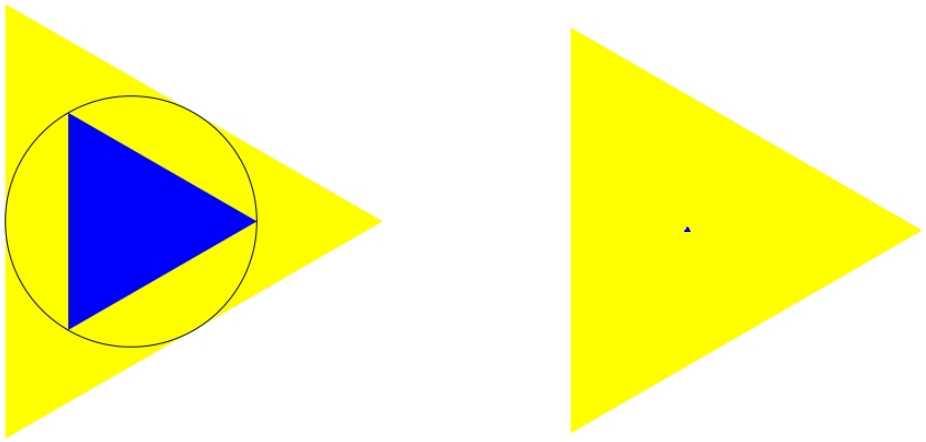
Proof. A matrix in the relative interior of $\mathcal{M}^{p\times q}_{3,2}$ will have multiple ellipses nested between P and Q: By the only if direction of the proof of Theorem 2.2.12 part (a), there exists an ellipse that is contained in Q and strictly contains P. We can just take slight scalings of this ellipse to get multiple ellipses. This proves the "if" direction.

For the "only if" direction, suppose M lies on the topological boundary and E_0 and E_1 are two ellipses nested between P and Q. Let $E_{1/2}$ be the ellipse determined by averaging the quadratics defining E_0 and E_1, i.e.

$$E_{1/2} = \{x \mid q_0(x) + q_1(x) \geq 0\} \text{ where } E_i = \{x \mid q_i(x) \geq 0\}.$$

It is straightforward to see that $E_{1/2}$ is nested between P and Q. Furthermore, if v is a vertex of P, then $E_{1/2}$ passes through v if and only if both E_0 and E_1 pass through v. Similarly, if f is a facet of Q, then $E_{1/2}$ is incident to f if and only if E_0 and E_1 are tangent to f at the same point. By Theorem 2.2.12, the ellipse $E_{1/2}$ must pass through three vertices of P and three facets of Q. Hence, there must exist six distinct points that both E_0 and E_1 pass through. No three of the six points are collinear, since ellipses E_0 and E_1 pass through them. Since five distinct points in general position determine a unique conic, we must have that $E_0 = E_1$. □

Example 2.2.15. *In the previous result, we examined the geometric configurations on the boundary of the semialgebraic set coming from strictly positive matrices. The simplest idea for such a matrix is to take two equilateral triangles and expand the inner one until we are on a boundary configuration as in Figure 2.6a.*



(a) Boundary configuration

(b) Interior configuration which also lies on the algebraic boundary $\overline{\partial \mathcal{P}_{3,2}}$

Figure 2.6: Geometric configurations of matrices in $\mathcal{P}_{3,2}^{3\times 3}$

This configuration has the slack matrix

$$\begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix}.$$

The 1035 term boundary polynomial from Theorem 2.2.12 vanishes on this matrix, as we expect.

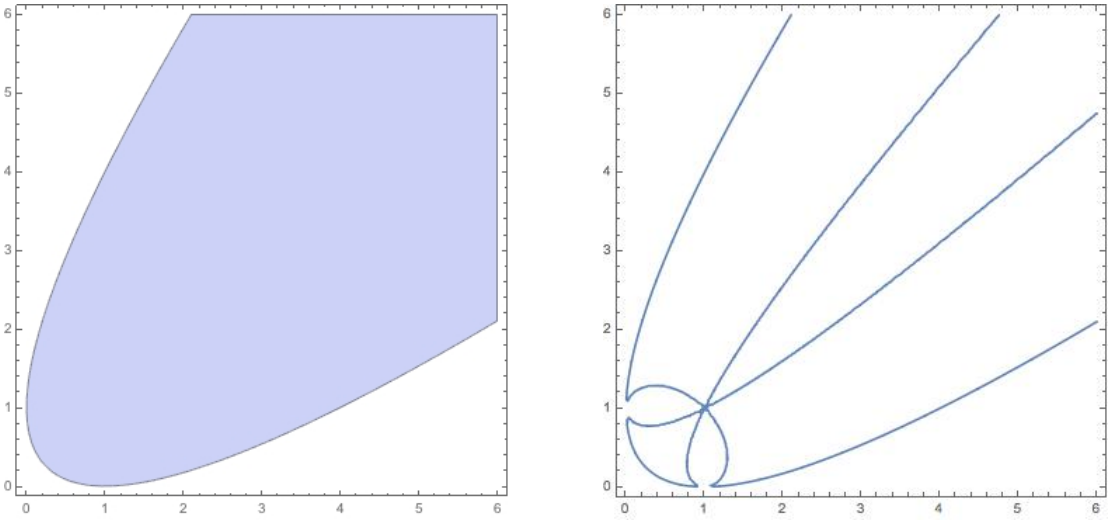
This matrix lies in the set of 3×3 *circulant matrices which have the form*

$$\begin{bmatrix} a & b & c \\ c & a & b \\ b & c & a \end{bmatrix}.$$

It was shown in [65, Example 2.7] that these matrices have psd rank at most 2 precisely when $a^2 + b^2 + c^2 - 2(ab + ac + bc) \leq 0$. *As expected, whenever this polynomial vanishes, the 1035 term boundary polynomial vanishes as well. Figure 2.6b shows an instance of the parameters* a, b, c *such that the matrix is on the algebraic boundary but not on the topological boundary – the polynomial vanishes, but the matrix lies in the interior of* $\mathcal{P}_{3,2}$.

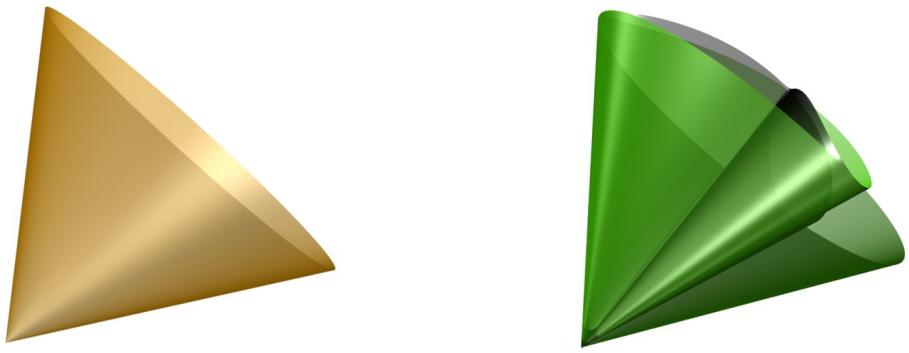
We were interested in finding out if the boundary polynomial could be used in an inequality to classify circulant matrices of psd rank at most 2. The family of circulant matrices which

have $c = 1$ and whose psd rank is at most 2 is depicted in Figure 2.7a. The boundary polynomial, shown in Figure 2.7b, takes both positive and negative values on the interior of the space. Figures 2.8a and 2.8b show the semialgebraic set and the boundary polynomial in the 3-dimensional space.



(a) Circulant matrices of psd rank at most 2 (b) The boundary polynomial

Figure 2.7: 3×3 circulant matrices in $\mathbb{R}^2$



(a) Circulant matrices of psd rank at most 2 (b) The boundary polynomial

Figure 2.8: 3×3 circulant matrices in $\mathbb{R}^3$

The phenomenon that the algebraic boundary of a semialgebraic set is relatively simple, e.g. consists of coordinate hyperplanes and one additional polynomial, but a semialgebraic description involves other polynomials also happens in the case of matrices of nonnegative rank at most three studied in Subsection 2.1.4 and partial matrices that can be completed to a rank one matrix in the standard simplex [106, Section 3].

2.2.4 Matrices of higher psd rank

In this subsection we focus on the space of $(k+1)\times(k+1)$ nonnegative matrices of psd rank at most k, and we study what it means for a matrix to lie on its boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$. In analogy with Corollary 2.2.13, we conjecture that a matrix lies on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ if and only if in every psd factorization, the matrices $A_1,\dots,A_{k+1}$ and $B_1,\dots,B_{k+1}$ all have to have rank 1.

Conjecture 2.2.16. *A matrix $M\in\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ lies on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ if and only if for every psd factorization $M_{ij}=\langle A_i,B_j\rangle$ with $A_i,B_j\in\mathcal{S}_+^k$,*

$$\operatorname{rank}(A_1)=\cdots=\operatorname{rank}(A_{k+1})=\operatorname{rank}(B_1)=\cdots=\operatorname{rank}(B_{k+1})=1.$$

Note that, according to Theorem 2.2.4, a matrix $M\in\mathbb{R}^{(k+1)\times(k+1)}$ has psd rank at most k if and only if we can nest a spectrahedral shadow of size k between the polytopes $P\subseteq Q\subseteq\mathbb{R}^k$ for which $M=S_{P,Q}$.

Recall that a spectrahedral shadow of size k is a linear projection of a spectrahedron of size k, which in turn is a slice of the cone of positive semidefinite matrices $\mathcal{S}_+^k$. Suppose we are given a spectrahedral shadow C of size k, and suppose that C is a linear projection of the spectrahedron $\tilde{C}=\mathcal{L}\cap\mathcal{S}_+^k$. A vector $v\in C$ lies in the *rank s locus* of C if there exists a $k\times k$ psd matrix in $\tilde{C}$ of rank s that projects onto v.

Let $P=\operatorname{conv}(v_1,\dots,v_p)$ with the origin in its interior, and let $Q=\{x:\langle h_i,x\rangle\le 1,i=1,\dots,q\}$. Denote the matrix with rows $v_1,\dots,v_p$ by V. Assume that $P\subseteq Q$. Let $\operatorname{rank}_{psd}(S_{P,Q})=k$ and let $A_1,\dots,A_p,B_1,\dots,B_q\in\mathcal{S}_+^k$ give a size k psd factorization of $S_{P,Q}$. We define two spectrahedral shadows of size k that are nested between P and Q. We follow [76, Section 4.1]:

$$C_A=\left\{x\in\mathbb{R}^n:\exists y\in\mathcal{S}_+^k\text{ s.t. }1-\langle h_j,x\rangle=\langle B_j,y\rangle\text{ for }j=1,\dots,q\right\},$$
$$C_B=\left\{Vz:\mathbf{1}^Tz=1,A_iz\in\mathcal{S}_+^k\text{ for }i=1,\dots,p\right\}.$$

By [76, Proposition 4], we have that

$$P\subseteq C_B\subseteq C_A\subseteq Q.$$

Lemma 2.2.17. *If $\operatorname{rank}(A_i)=1$, then v_i lies in the rank one locus of C_A, and if $\operatorname{rank}(B_j)=1$, then, C_B touches the facet of Q defined by $\langle h_j,x\rangle\le 1$ at a point $u\in Q$ from its rank $(k-1)$ locus.*

We prove this lemma in Subsection 2.2.6.3. It leads us to the following the geometric version of Conjecture 2.2.16.

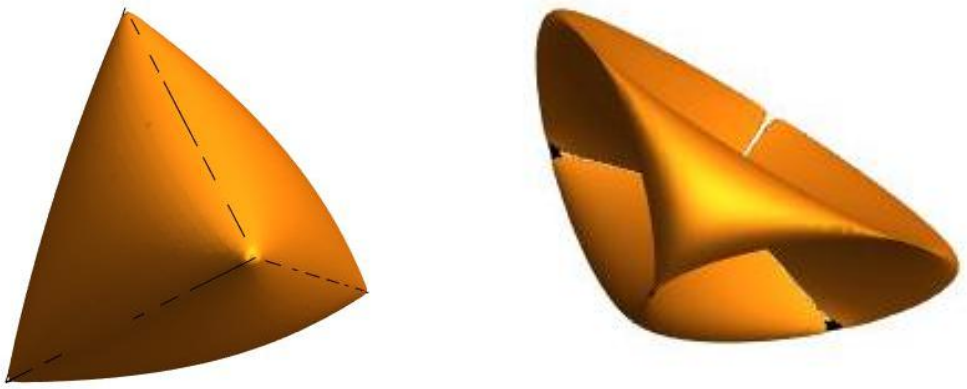
Conjecture 2.2.18. *A matrix $S_{P,Q}$ lies on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ if and only if all vertices of P lie on the rank one locus of the spectrahedral shadow C_A, and every facet of Q touches the spectrahedral shadow C_B at points lying on its rank $k-1$ locus.*

Since $C_B \subseteq C_A$, the boundaries of C_A and C_B intersect at the vertices of P and at the tangency points with Q. This motivates us to state the following stronger conjecture:

Conjecture 2.2.19. *A matrix $S_{P,Q}$ is on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ if and only if for all spectrahedral shadows C of $k \times k$ matrices such that $P \subseteq C \subseteq Q$, $k+1$ of the vertices of P lie on the rank one locus of C and $k+1$ of the facets of Q touch C at points on its rank $k-1$ locus.*

The psd rank 3 and rank 4 setting corresponds to the geometric configuration where a 3-dimensional spectrahedral shadow size 3 is nested between 3-dimensional polytopes. A detailed study of generic spectrahedral shadows can be found in [140].

Example 2.2.20. *We now give an example of a geometric configuration as in Conjecture 2.2.19. It is depicted in Figure 2.9a. We stipulate that the vertices of the interior polytope coincide with the nodes of the spectrahedron and the facets of the outer polytope touch the boundary of this spectrahedron at rank 2 loci. In the dual picture, the vertices of the inner polytope lie on the rank 1 locus depicted in Figure 2.9b and the facets of the outer polytope contain the rank 2 locus of this spectrahedral shadow.*



(a) Spectrahedron (b) Rank 1 locus of a spectrahedral shadow

Figure 2.9: 3-dimensional spectrahedral shadows

We end this section with a restatement of the conjecture using Hadamard square roots.

Definition 2.2.21. *Given a nonnegative matrix M, let $\sqrt{M}$ denote a Hadamard square root of M obtained by replacing each entry in M by one of its two possible square roots. The square root rank of a nonnegative matrix M, denoted as $\text{rank}_{\surd}(M)$, is the minimum rank of a Hadamard square root of M.*

Lemma 2.2.22 ([78], Lemma 2.4)**.** *The smallest k for which a nonnegative real matrix M admits a $\mathcal{S}_+^k$-factorization in which all factors are matrices on rank one is $k = \text{rank}_{\surd}(M)$.*

Hence Conjecture 2.2.16 is equivalent to the statement that a matrix $M \in \mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ lies on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ if and only if its square root rank is at most k. We conclude this section with a conjecture which would lead to a semialgebraic description of $\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$.

Conjecture 2.2.23. *Every matrix $M \in \mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$ has a psd factorization with at least $2k+1$ of the matrices in the factorization being rank 1.*

2.2.5 Evidence towards Conjecture 2.2.16

In this section, we present partial evidence towards proving Conjecture 2.2.16. Section 2.2.5.1 is theoretical in nature, while Section 2.2.5.2 is computational.

2.2.5.1 Nested spectrahedra

We know from Theorem 2.2.4 that a matrix M such that $M\mathbf{1} = \mathbf{1}$ has psd rank k if and only if we can fit a spectrahedral shadow of size k in between the two polytopes P and Q corresponding to M. In the following lemma, we show that a $(k+1)\times(k+1)$ matrix M has psd rank k if and only if we can fit a spectrahedron of size k in between P and Q. We show that if there is a spectrahedral shadow C nested between P and Q, then we can find a spectrahedron C' of the same size such that $P \subseteq C' \subseteq C \subseteq Q$.

Lemma 2.2.24. *Let $M \in \mathbb{R}_{\geq 0}^{(k+1)\times(k+1)}$ be a full-rank matrix such that $M\mathbf{1} = \mathbf{1}$. Then, M has psd rank at most k if and only if we can nest a spectrahedron of size k between the two polytopes P and Q corresponding to M.*

The proof of this lemma can be found in Subsection 2.2.6.4. We believe that its statement also holds for matrices of any size.

Conjecture 2.2.25. *Let $M \in \mathbb{R}_{\geq 0}^{p\times}$ have rank $k+1$ and assume that $M\mathbf{1} = \mathbf{1}$. Then, M has psd rank at most k if and only if we can nest a spectrahedron of size k between the two polytopes P and Q corresponding to M.*

We now show that given a spectrahedron C of size k such that $P \subseteq C \subseteq Q$, where P is a simplex and k of the vertices of P are also vertices of C, one can find a new spectrahedron C' such that $P \subseteq C' \subseteq C \subseteq Q$ such that all $k+1$ of the vertices of P are also vertices of C' (in other words, they correspond to rank 1 matrices in C').

Lemma 2.2.26. *Let $P \subseteq \mathbb{R}^k$ be the simplex $P = conv(e_1, \ldots, e_k, 0)$. Let C be a slice of $\mathcal{S}_+^k$ such that $P \subseteq C$ and the vertices $e_1, \ldots, e_k$ lie in the rank one locus of C. Then, we can find another spectrahedral shadow C' of size k such that $P \subseteq C' \subseteq C$ with all $k+1$ vertices of P corresponding to rank 1 matrices in C'.*

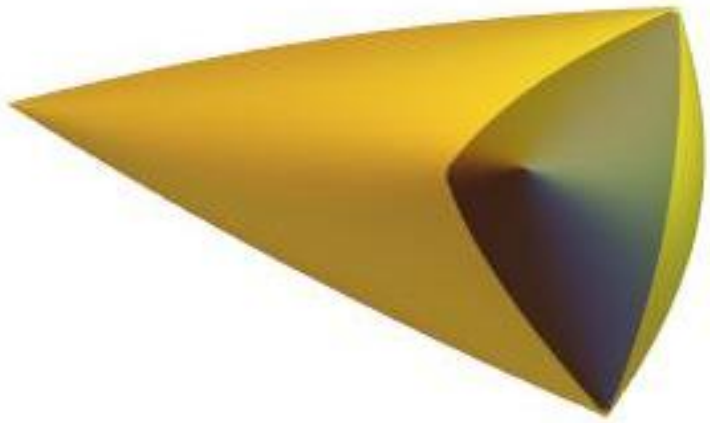


Figure 2.10: The spectrahedra C (in yellow) and C' (in blue) as in Lemma 2.2.26
.

The proof of this lemma can be found in Subsection 2.2.6.5. Consider the slack matrix $S_{P,Q}$ of the polytopes defined in Lemma 2.2.26. The statement of the Lemma implies that $S_{P,Q}$ does not lie on the boundary $\partial\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$, because once we find the new spectrahedron C', we see that it does not touch Q. As we saw in Section 2.2.3, in order for a matrix to lie on the boundary, the configuration $P \subseteq C \subseteq Q$ has to be very tight, and Lemma 2.2.26 shows that having k of the vertices of P lie in the rank one locus of C is not tight enough. Similarly, having k of the facets of Q touch Q at rank $k-1$ loci won't be enough. This is why we believe that all $k+1$ vertices of P have to be in the rank one locus of C, and all $k+1$ of the facets of Q have to touch C at its rank $k-1$ locus, which is the statement of Conjecture 2.2.18.

2.2.5.2 Computational results

In this section we provide computational evidence for Conjecture 2.2.16 when $k > 2$.

Example 2.2.27. *We consider the 2-dimensional family of* 4×4 *circulant matrices*

$$\begin{bmatrix} a & b & 1 & b \\ b & a & b & 1 \\ 1 & b & a & b \\ b & 1 & b & a \end{bmatrix} \tag{2.2.3}$$

which is parametrized by a *and* b.

In Figure 2.11, the 4126 green dots correspond to randomly chosen matrices of the form (2.2.3) that have psd rank at most three. The psd rank is computed using the code provided by the authors of [153] adapted to the computation of the semidefinite rank [97, Section 5.6]. The red curves correspond to matrices of the form (2.2.3) that have a psd factorization by 3×3 *rank one matrices. These curves are obtained by an elimination procedure in* `Macaulay2`.

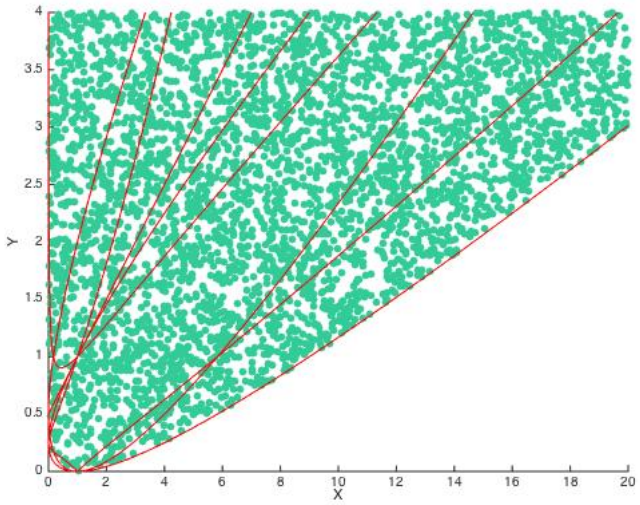


Figure 2.11: A family of 4×4 circulant matrices of psd rank at most 3

If the condition that all of the matrices $A_1, \ldots, A_{k+1}, B_1, \ldots, B_{k+1}$ have rank one is equivalent to the matrix M being on the boundary $\partial \mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$, then the set of matrices M in whose psd factorization all A_i's and B_j's have rank one should have codimension 1 inside $\mathcal{P}_{k+1,k}^{(k+1)\times(k+1)}$. In other words, it should have codimension 1 inside $\mathcal{V}_{k+1}^{(k+1)\times(k+1)} = \mathbb{R}^{(k+1)\times(k+1)}$. Consider the map that takes rank one matrices A_i and B_j and gets M such that $M_{ij} = \langle A_i, B_j \rangle$. The rank of its Jacobian should be $(k+1)^2 - 1$ if Conjecture 2.2.16 is true. In the following example, we test several different assignments of ranks to each of the matrices A_i, B_j, and we check those for which the Jacobian has dimension $(k+1)^2 - 1$.

Example 2.2.28. *We construct $k \times k$ positive semidefinite matrices $A_1, \ldots, A_{k+1}, B_1, \ldots, B_{k+1}$ of ranks $r_1, \ldots, r_{2k+2}$. We construct a matrix M such that $M_{ij} = \langle A_i, B_j \rangle$. We vectorize the matrix M and compute its Jacobian J with respect to the entries of $A_1, \ldots, A_{k+1}, B_1, \ldots, B_{k+1}$. Finally we substitute the entries of $A_1, \ldots, A_{k+1}, B_1, \ldots, B_{k+1}$ by random nonnegative integers and compute the rank of J. If $\operatorname{rank}(J) = (k+1)^2 - 1$, then the matrices that have psd factorization by $\{r_1, \ldots, r_p\}, \{r_{p+1}, \ldots, r_{2k+2}\}$ rank matrices give a candidate for a boundary component (assuming that the boundary components are only dependent on the ranks of the A_i's and the B_j's).*

psd rank	p	q	ranks
3	4	4	{{1,1,1,1},{1,1,1,1}}
3	4	5	{{1,1,1,1},{1,1,1,1,2/3}}
3	4	6	{{1,1,1,1},{1,1,1,1,2/3,2/3}},{{1,1,1,2},{1,1,1,1,1,1}}
3	5	5	{{1,1,1,1,2/3},{1,1,1,1,2/3}}
3	5	6	{{1,1,1,1,2/3},{1,1,1,1,2/3,2/3}},{{1,1,1,2,3},{1,1,1,1,1,1}}
3	6	6	{{1,1,1,1,2/3,2/3},{1,1,1,1,2/3,2/3}},{{1,1,1,1,1,1},{1,1,1,2,3,3}}, {{1,1,1,1,1,1},{1,1,2,2,2,2}},{{1,1,1,1,1,2},{1,1,1,2,2,2}}

Table 2.3: Ranks of matrices in the psd factorization of a psd rank three matrix that can potentially give boundary components

The possible candidates for $k = 3$ are summarized in Table 2.3. For all p, q the case where four matrices A_i and four matrices B_j have rank 1 and all other matrices have any rank greater than 1 are represented. For $k = 4$ the analogous statement is not true. If $M \in \mathbb{R}^{10\times 10}$, exactly five A_i and five B_j matrices have rank one and the rest of the matrices have rank two, then the Jacobian has rank 94. If the rest of the matrices in the psd factorization have rank three or four, then the Jacobian has rank 99 as expected. Hence without further constraints on the ranks of the rest of the matrices Conjecture 2.2.16 does not hold for general r and k.

Example 2.2.29. *Using the same strategy as in Example 2.2.28, we have checked that the Jacobian has the expected rank for $r = k + 1$ and $k < 10$.*

2.2.6 Proofs

2.2.6.1 Proof of Lemma 2.2.7

The fact that the first statement implies the second follows from the definition of interior of $\mathcal{P}_{r,k}$. For the other direction, assume that for the rescaled matrix N there exists $\epsilon > 0$ such that $\mathcal{B}_\epsilon(N) \cap \mathcal{V}_r \cap R \subseteq \mathcal{P}_{r,k} \cap R$. Let $\alpha_1, \ldots, \alpha_p$ be the row sums of M, i.e. $M\mathbf{1} = \alpha$. Without loss of generality, assume that $0 < \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_p$. Then, consider the ball $\mathcal{B}_{\epsilon\alpha_1}(M)$. If a matrix $M' = M + A \in \mathcal{B}_{\epsilon\alpha_1}(M) \cap \mathcal{V}_r$, then, after dividing the rows of M' by $\alpha_1, \ldots, \alpha_p$ respectively, we obtain the matrix $N + B$, where B is the rescaled version of A. Since $\alpha_1 \leq \cdots \leq \alpha_p$, then $\|B\| \leq \frac{1}{\alpha_1}\|A\|$. Thus, $N + B \in \mathcal{B}_\epsilon(N)$, so, the matrix $N + B \in \mathcal{B}_\epsilon(N) \cap \mathcal{V}_r \cap R \subseteq \mathcal{P}_{r,k} \cap R$. Thus, $M' \in \mathcal{P}_{r,k}$, so, $\mathcal{B}_{\epsilon\alpha_1}(M) \cap \mathcal{V}_r \subseteq \mathcal{P}_{r,k}$, i.e. M is in the interior of $\mathcal{P}_{r,k}$.

2.2.6.2 Proof of Lemma 2.2.9

Note that for a matrix to lie in the interior of $\mathcal{P}_{3,2}$ all of its entries need to be strictly positive.

Assume first that M lies in the interior of $\mathcal{P}_{3,2}$. Since it lies in$\mathcal{P}_{3,2}$, then there exist polygons $P, Q \subseteq \mathbb{R}^2$ and an ellipse E such that $P \subseteq E \subseteq Q$, and $M = S_{P,Q}$. If the boundary of E does not contain any of the vertices of P, then we are done. Suppose that the boundary

of E contains some of the vertices of P. We are going to find another ellipse E' such that $P \subset E \subset E' \subset Q$ and the boundary of E' doesn't contain any of the vertices of P.

Since M is in the interior of $\mathcal{P}_{3,2}$, none of the entries of M are 0, so the boundary of the polygon Q does not contain any of the vertices of P. Moreover, there exists $\epsilon > 0$ such that $\mathcal{V}_3 \cap \mathcal{B}_\epsilon(M) \subset \mathcal{P}_{3,2}$. Pick a point in the interior of the polygon P and consider the polygon tP obtained by a homotety centered at the selected point with some $t > 1$. Then, $P \subset tP \subseteq Q$ for a small enough $t > 1$, and P is strictly contained in tP. Now consider the generalized slack matrix of tP and Q and call it M_t. We can choose t close enough to 1 so that $M_t \in \mathcal{B}_\epsilon(M) \subseteq \mathcal{P}_{3,2}$. Thus, M_t has psd rank at most 2 and there exists an ellipse E' such that $tP \subseteq E' \subseteq Q$. Therefore, $P \subset tP \subseteq E' \subseteq Q$ and the boundary of the ellipse E' does not contain the vertices of P.

Now, suppose there exists an ellipse E and polygons P and Q obtained from a factorization $M = AB$ as before such that $P \subset E \subseteq Q$ and the ellipse E does not contain any of the vertices of P. Therefore, it is possible to shrink the ellipse E slightly so that it also doesn't touch any of the sides of Q. So, now we have an ellipse E that does not touch any of the vertices of P and does not touch any of the sides of Q. Let $\epsilon > 0$. By perturbing A and B, we can express any matrix $N \in \mathcal{B}_\epsilon(M) \cap \mathcal{V}_3$ as $N = A_\epsilon B_\epsilon$. But perturbing A and B results in a perturbation of P and Q, which are defined linearly according to A and B. Therefore, we can choose ϵ small enough so that any matrix $N \in \mathcal{B}_\epsilon(M) \cap \mathcal{V}_3$ can be expressed as $N = A'B'$ where A' and B' are perturbations of A and B such that the corresponding P' and Q' are perturbations of P and Q that still satisfy $P' \subseteq E \subseteq Q'$. Therefore, $N \in \mathcal{P}_{3,2}$ so that $\mathcal{B}_\epsilon(M) \cap \mathcal{V}_3 \subseteq \mathcal{P}_{3,2}$.

2.2.6.3 Proof of Lemma 2.2.17

Since $1 - \langle h_i, v_j \rangle = \langle A_i, B_j \rangle$, the matrix B_j in the $\mathcal{S}_+^k$-lift of C_A projects to $v_j \in C_A$. If $\operatorname{rank}(B_j) = 1$, then v_j lies in the rank one locus of the spectrahedral shadow C_A.

In the dual picture, the inner polytope P becomes the outer polytope P° and the outer polytope Q becomes the inner polytope Q°. Then Q° is the convex hull of $h_1, \ldots, h_q$ and P° is defined by $\langle v_j, x \rangle \leq 1$ for $j = 1, \ldots, p$.

Lemma 2.2.30. *The dual of the convex body C_A is the convex body $\left\{w^T H : w^T \mathbf{1} \leq 1, w^T A \in \mathcal{S}_+^k\right\}$.*

Proof. The proof we will give here is analogous to the proof of [76, Theorem 3]. By definition

$$(C_A)^\circ = \left\{z \in \mathbb{R}^n : z^T x \leq 1 \ \forall x \in C_A\right\}.$$

Consider the problem

$$\max\left\{z^T x : x \in C_A\right\} = \max\left\{z^T x : 1 - \langle h_i, x \rangle = \langle B_i, y \rangle \text{ for } i = 1, \ldots, q, y \in \mathcal{S}_+^k\right\}.$$

Strong duality holds since $\max\left\{z^T x : x \in C_A\right\}$ is a convex optimization problem and C_A has an interior point because it contains P. The dual program is given by

$$\min\left\{w^T \mathbf{1} : z = w^T H, w^T A \in \mathcal{S}_+^k\right\}.$$

This gives

$$(C_A)^\circ = \left\{ w^T H : w^T \mathbf{1} \leq 1, w^T A \in \mathcal{S}_+^k \right\}. \tag{2.2.4}$$

□

Remark 2.2.31. *We can replace the inequality $w^T\mathbf{1} \leq 1$ in (2.2.4) by the equality $w^T\mathbf{1} = 1$.*

Proof. The proof we will give here is analogous to the proof of [76, Remark 3]. There exists $s \geq 0$ such that $w^T\mathbf{1} + s = 1$. Since the polytopes P and Q contain 0 in their interiors, also the dual polytopes P° and Q° contain 0 in their interiors. Hence there exist $\lambda_1, \ldots, \lambda_q \geq 0$ such that $\sum \lambda_i = 1$ and $\sum \lambda_i h_i = 0$. Define $\widetilde{w} = w + s\lambda$ where $\lambda = (\lambda_i)$. Then

$$\begin{aligned} \widetilde{w}^T\mathbf{1} &= w^T\mathbf{1} + s\lambda^T\mathbf{1} = w^T\mathbf{1} + s = 1, \\ \widetilde{w}^T A &= w^T A + s\lambda^T A \in \mathcal{S}_+^k \end{aligned}$$

because $\lambda \geq 0$ and each component is in $\mathcal{S}_+^k$ and

$$\widetilde{w}^T H = w^T H + s\lambda^T H = w^T H.$$

□

Hence the dual bodies of C_A and C_B are

$$\begin{aligned} (C_A)^\circ &= \left\{ z^T H : z^T\mathbf{1} = 1, z^T B_i \in \mathcal{S}_+^k \text{ for } i = 1, \ldots, q \right\}, \\ (C_B)^\circ &= \left\{ x \in \mathbb{R}^n : \exists y \in \mathcal{S}_+^k \text{ s.t. } 1 - \langle x, v_j \rangle = \langle y, A_j \rangle \text{ for } j = 1, \ldots, p \right\}. \end{aligned}$$

As before, if $\operatorname{rank}(A_i) = 1$ then h_i lies in the rank one locus of the spectrahedral shadow $(C_B)^\circ$. In the primal picture this means that the spectrahedral shadow C_B touches the polytope Q at a generic point (i.e. a matrix of rank $k-1$) on the boundary.

2.2.6.4 Proof of Lemma 2.2.24

If we can fit a spectrahedron of size k between P and Q, then M has psd rank at most k.

Now, suppose that M has psd rank at most k. Since M is full rank, we can factor it as $M = AB$, where $A, B \in \mathbb{R}^{(k+1)\times(k+1)}$ and

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = A^{-1}M.$$

Then, the inner polytope P comes from a slice of the cone over the convex hull of the rows of A. Let the slice be given by last coordinate equal to 1. Then, P is the standard simplex in $\mathbb{R}^k$, i.e.

$$P = \operatorname{conv}\{e_1, \ldots, e_k, 0\}.$$

Since M has psd rank k, there exists a slice of L of $\mathcal{S}_+^k$ and a linear map π such that $C = \pi(L \cap \mathcal{S}_+^k)$ lies between P and Q:

$$P \subseteq C \subseteq Q.$$

If π is a $1:1$ linear map, then the image C is just a linear transformation of a slice of $\mathcal{S}_+^k$, which is considered to be a slice. So, assume that π is not $1:1$, i.e. it has a non-trivial kernel.

We can write

$$L \cap \mathcal{S}_+^k = \{(x_1, \ldots, x_s) | \sum_{i=1}^{s} x_i A_i + (1 - \sum_i x_i)) A_{s+1} \succeq 0\}$$

for some symmetric matrices $A_1, \ldots, A_{s+1}$. Now, let $u_1, \ldots, u_s$ be an orthonormal basis of $\mathbb{R}^s$ such that $\ker(\pi) = \text{span}(u_{k+1}, \ldots, u_s)$. Let U be the orthogonal matrix with columns $u_1, \ldots, u_s$. Consider new coordinates y such that $x = Uy$. Then, we can rewrite (after a linear transformation)

$$L \cap \mathcal{S}_+^k = \{(y_1, \ldots, y_s) | \sum_i y_i B_i + (1 - (\sum_i y_i)) B_{s+1} \succeq 0\},$$

where $B_1, \ldots, B_{s+1}$ are linear combinations of the A_i's. Then,

$$C = \{(y_1, \ldots, y_k) | \exists y_{k+1}, \ldots, y_s \text{ s.t.} \sum_i y_i B_i + (1 - (\sum_i y_i)) B_{s+1}) \succeq 0\}.$$

We know that $P \subseteq C$ and $P = \text{conv}(e_1, \ldots, e_k, 0)$. Since $e_i \in P \subseteq C$, then there exist $y_{k+1}^{(i)}, \ldots, y_s^{(i)} \in \mathbb{R}$ such that

$$D_i := B_i + \sum_{j=k+1}^{s} [y_j^{(i)} (B_j - B_{s+1})] \succeq 0.$$

Since $0 \in P \subseteq C$, then, there exist $y_{k+1}^{(0)}, \ldots, y_s^{(0)} \in \mathbb{R}$ such that

$$D_{k+1} := B_{s+1} + \sum_{j=k+1}^{s} [y_j^{(0)} (B_j - B_{s+1})] \succeq 0.$$

Consider the spectrahedron

$$C' := \{(y_1, \ldots, y_k) | \sum_{i=1}^{k} y_i D_i + (1 - \sum_i y_i) D_{k+1} \succeq 0\}.$$

Note that $e_i \in C'$ for every $i = 1, \ldots, k$ since $D_i \succeq 0$. Moreover, $0 \in C'$ since $D_{k+1} \succeq 0$. Thus, $P \subseteq C'$.

Moreover, if $(y_1, \ldots, y_k) \in C'$, then

$$\begin{aligned} 0 \preceq \sum_{i=1}^{k} y_i D_i + (1 - \sum_i y_i) D_{k+1} &= \sum_{i=1}^{k} y_i \left(B_i + \sum_{j=k+1}^{s} [y_j^{(i)} (B_j - B_{s+1})] \right) \\ &+ (1 - \sum_i y_i) \left(B_{s+1} + \sum_{j=k+1}^{s} [y_j^{(0)} (B_j - B_{s+1})] \right) \\ &= \sum_{i=1}^{k} y_i B_i + \sum_{j=k+1}^{s} \left(\sum_{i=1}^{k} y_i y_j^{(i)} - (1 - \sum_{i=1}^{k} y_i) y_j^{(0)} \right) B_j \\ &+ \left(1 - \sum_{i=1}^{k} y_i - \sum_{j=k+1}^{s} \left(\sum_{i=1}^{k} y_i y_j^{(i)} - (1 - \sum_{i=1}^{k} y_i) y_j^{(0)} \right) \right) B_{s+1}. \end{aligned}$$

Therefore, $(y_1, \ldots, y_k) \in C$ and so $P \subseteq C' \subseteq C \subseteq Q$. Therefore, we can nest the spectrahedron C' in between P and Q.

2.2.6.5 Proof of Lemma 2.2.26

This Lemma is trivial when $k = 1$. We proceed by induction on k.

By the conditions in the statement of the lemma, we can assume that

$$C = \{(x_1, \ldots, x_k) | x_1 a_1 a_1^T + x_2 a_2 a_2^T + \cdots + x_k a_k a_k^T + (1 - \sum_i x_i) B \succeq 0\},$$

where $B \succeq 0$ since $0 \in C$ and $a_1, \ldots, a_k \in \mathbb{R}^k$ are vectors.

Suppose first that $\dim \text{span}\{a_1, \ldots, a_k\} = \ell < k$. Let U be a change of coordinates that transforms $\text{span}\{a_1, \ldots, a_k\}$ into $\text{span}\{e_1, \ldots, e_l\}$. Then, if $a_i' = U a_i$, we have that

$$C = \{(x_1, \ldots, x_k) | x_1 a_1' (a_1')^T + x_2 a_2' (a_2')^T + \cdots + x_k a_k' (a_k')^T + (1 - \sum_i x_i) U B U^T \succeq 0\},$$

where $B' := UBU^T$ is still positive semidefinite. If $B'_{i,j} = 0$ for <u>all</u> $i, j \geq \ell + 1$, then, the statement reduces to the case of ℓ, which is true by induction. So, suppose that, say, (since $B' \succeq 0$) $B'_{\ell+1,\ell+1} > 0$. Then, choose a vector $d \in \mathbb{R}^k$ such that $d_{\ell+1} \neq 0$ and $dd^T \preceq B'$. Consider the spectrahedron

$$C' := \{(x_1, \ldots, x_k) | x_1 a_1' (a_1')^T + x_2 a_2' (a_2')^T + \cdots + x_k a_k' (a_k')^T + (1 - \sum_i x_i) dd^T \succeq 0\}.$$

First note that clearly $e_1, \ldots, e_k, 0 \in C'$. We will show that $C' \subseteq C$. Indeed, let $(x_1, \ldots, x_k) \in C'$. Since $(a_i')_{\ell+1} = 0$ for all i, $d_{\ell+1} \neq 0$ and

$$x_1 a_1' (a_1')^T + x_2 a_2' (a_2')^T + \cdots + x_k a_k' (a_k')^T + (1 - \sum_i x_i) dd^T \succeq 0,$$

we have $(1 - \sum_i x_i) \geq 0$. But then

$$0 \preceq x_1 a_1'(a_1')^T + x_2 a_2'(a_2')^T + \cdots + x_k a_k'(a_k')^T + (1 - \sum_i x_i) dd^T$$

$$\preceq x_1 a_1'(a_1')^T + x_2 a_2'(a_2')^T + \cdots + x_k a_k'(a_k')^T + (1 - \sum_i x_i) B'$$

and, therefore, $C' \subseteq C$.

Now, assume that dim span$\{a_1, \ldots, a_k\} = k$. Then, let U be an invertible transformation such that $Ua_i = e_i$. Then,

$$C = \{(x_1, \ldots, x_k) | x_1 e_1 e_1^T + x_2 e_2 e_2^T + \cdots + x_k e_k e_k^T + (1 - \sum_i x_i) UBU^T \succeq 0\},$$

where $B' := UBU^T \succeq 0$. Let $d \in \mathbb{R}^k$ be such that $d_i = \sqrt{B'_{i,i}}$ and let $S \in \mathbb{R}^{k \times k}$ be such that

$$S_{i,j} = \begin{cases} \frac{B'_{i,j}}{\sqrt{B'_{i,i} B'_{j,j}}} & \text{if } B'_{i,i} B'_{j,j} \neq 0, \\ 1 & \text{if } B'_{i,i} B'_{j,j} = 0 \text{ and } i = j, \\ 0 & \text{if } B'_{i,i} B'_{j,j} = 0 \text{ and } i \neq j. \end{cases}$$

Since $B' \succeq 0$, it is clear that $S \succeq 0$ as well since it is obtained from B' by rescaling some rows and columns and by adding 1 on the diagonal in places that are 0 in B'. Let

$$C' = \{(x_1, \ldots, x_k) | x_1 e_1 e_1^T + x_2 e_2 e_2^T + \cdots + x_k e_k e_k^T + (1 - \sum_i x_i) dd^T \succeq 0\}.$$

Then, clearly $e_1, \ldots, e_k, 0 \in C'$. We will show that $C' \subseteq C$. Let $(x_1, \ldots, x_k) \in C'$. Then,

$$x_1 e_1 e_1^T + x_2 e_2 e_2^T + \cdots + x_k e_k e_k^T + (1 - \sum_i x_i) dd^T \succeq 0. \tag{2.2.5}$$

By the Schur Product Theorem, we know that the Hadamard product of two positive semidefinite matrices is positive semidefinite. Therefore, when we take the Hadamard product of the matrix (2.2.5) with S, we get a positive semidefinite matrix. But that Hadamard product equals

$$x_1 e_1 e_1^T + x_2 e_2 e_2^T + \cdots + x_k e_k e_k^T + (1 - \sum_i x_i) B' \succeq 0,$$

therefore, $C' \subseteq C$.

Acknowledgements

Part of this work was done while Kaie and I were visiting the Simons Institute for the Theory of Computing, UC Berkeley. We thank Kristian Ranestad and Bernd Sturmfels for very helpful discussions, Rekha Thomas for reading the first draft of the article and Sophia Sage Elia for making Figure 2.8.

2.3 Conclusion

In this chapter we explored two different types of matrix factorizations: nonnegative and positive semidefinite. We studied the set $\mathcal{M}_{r,k}$ of matrices of rank at most r and nonnegative rank at most k, and the set $\mathcal{P}_{r,k}$ of matrices of rank at most r and positive semidefinite rank at most k. Both $\mathcal{M}_{r,k}$ and $\mathcal{P}_{r,k}$ are full-dimensional semialgebraic subsets of the determinantal variety $\mathcal{V}_r$. Moreover, both nonnegative and positive semidefinite factorizations have beautiful geometric interpretations via nested polyhedra. Using these, we were able to describe the boundaries of $\mathcal{M}_{r,k}$ and $\mathcal{P}_{r,k}$ for small values of r and k, and to obtain a conjecture for general r and k.

Chapter 3

Orthogonally Decomposable Tensors

Orthogonally decomposable tensors possess many appealing properties. In this chapter we focus mainly on their spectral properties. In Section 3.1 we study the eigenvectors of symmetric orthogonally decomposable tensors, while in Section 3.2 we study the singular vector tuples of ordinary orthogonally decomposable tensors.

3.1 Symmetric Odeco Tensors

A real symmetric tensor is orthogonally decomposable (or odeco) if it can be written as a linear combination of symmetric powers of n vectors which form an orthonormal basis of $\mathbb{R}^n$. Motivated by the spectral theorem for real symmetric matrices, we study the properties of odeco tensors. We give a formula for all of the eigenvectors of an odeco tensor. Moreover, we formulate a set of polynomial equations that vanish on the odeco variety and we conjecture that these polynomials generate its prime ideal. We prove this conjecture in some cases and give strong evidence for its overall correctness. This section is based on my paper *Orthogonal Decomposition of Symmetric Tensors* [131]. In the last Subsection we present a conjecture which has been resolved in subsequent work [23], and is presented in Section 2.1.2.

3.1.1 Introduction

The spectral theorem states that every $n \times n$ real symmetric matrix M possesses n real eigenvectors $v_1, \ldots, v_n$ which form an orthonormal basis of $\mathbb{R}^n$. Moreover, one can express M as $M = \sum_{i=1}^{n} \lambda_i v_i v_i^T$, where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are the corresponding eigenvalues. In this section we investigate when such a decomposition is possible for real symmetric tensors. We address the following two questions.

Question 1. *Which real symmetric tensors T can be decomposed as $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_n v_n^{\otimes d}$, form some orthonormal basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ and some $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$? More precisely, can we find equations in the entries of T that cut out the set of tensors for which such a decomposition exists?*

Question 2. *Given that a tensor T can be decomposed as $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_n v_n^{\otimes d}$, where $v_1, \ldots, v_n \in \mathbb{R}^n$ are orthonormal, can we express the eigenvectors of T (to be defined) in terms of $v_1, \ldots, v_n$?*

Let $S^d(\mathbb{R}^n)$ denote the space of $n \times n \times \cdots \times n$ (d times) symmetric tensors, i.e. tensors whose entries are real numbers $T_{i_1 \ldots i_d}$ invariant under permuting the indices: $T_{i_1 \ldots i_d} = T_{i_{\sigma(1)} \ldots i_{\sigma(d)}}$ for all permutations σ of the set $\{1, 2, \ldots, d\}$. For example, when $d = 2$, the space $S^2(\mathbb{R}^n)$ consists of all $n \times n$ real symmetric matrices. We study the elements $T \in S^d(\mathbb{R}^n)$ which can be written as $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_n v_n^{\otimes d}$, where $v_1, \ldots, v_n \in \mathbb{R}^n$ form an orthonormal basis of $\mathbb{R}^n$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. We call such tensors T *orthogonally decomposable* or, for short, *odeco*.

The notion of eigenvectors of matrices was extended to symmetric tensors by Lim [112] and by Qi [126] independently in 2005. A vector $w \in \mathbb{C}^n$ is an *eigenvector* of $T \in S^d(\mathbb{C}^n)$ if there exists $\lambda \in \mathbb{C}$, the corresponding *eigenvalue*, such that

$$Tw^{d-1} := \Big[\sum_{i_2, \ldots, i_d = 1}^{n} T_{i, i_2, \ldots, i_d} w_{i_2} \ldots w_{i_d} \Big]_i = \lambda w.$$

Two *eigenpairs* (w, λ) and (w', λ') are equivalent if there exists $t \neq 0$ such that $w = tw'$ and $\lambda = t^{d-2}\lambda'$. When $d = 2$, these definitions agree with the usual definitions of eigenvectors, eigenvalues, and equivalence of eigenpairs for matrices.

The spectral theorem answers both Question 1 and Question 2 in the case $d = 2$: every symmetric matrix $M \in S^2(\mathbb{R}^n)$ can be written as $M = \sum_{i=1}^n \lambda_i v_i v_i^T = \sum_{i=1}^n \lambda_i v_i^{\otimes 2}$, where $v_1, \ldots, v_n$ are orthonormal. Moreover, if M is generic (in the sense that its eigenvalues are distinct), then $v_1, \ldots, v_n$ are *all* of the eigenvectors of M up to scaling.

In Subsection 3.2.2 we give an explicit algebraic formula of all of the eigenvectors of an odeco tensor $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_n v_n^{\otimes d}$ in terms of $v_1, \ldots, v_n$, answering Question 2 above. It easily follows from the definition of eigenvectors that $v_1, \ldots, v_n$ are eigenvectors of T. These are not all of the eigenvectors of T, but it turns out that one can explicitly express the rest of them in terms of $v_1, \ldots, v_n$.

For general d, not all tensors $T \in S^d(\mathbb{R}^n)$ are odeco. Section 4.1 is dedicated to finding the equations defining the variety of odeco tensors. In Subsection 3.1.3, we give partial results towards what is done in Section 4.1. We study the set of all odeco tensors and find equations that vanish on this set. In Conjecture 3.1.16 we claim that these define the prime ideal of the odeco variety, which is the Zariski closure of the set of odeco tensors inside $S^d(\mathbb{C}^n)$. In Theorem 3.1.20 we prove Conjecture 3.1.16 for the special case $n = 2$. In Subsection 3.1.3.1 we conclude the section by giving evidence for the correctness of this conjecture. This conjecture is later proved set-theoretically in Section 4.1.

In the remainder of this subsection we review symmetric tensor decomposition as well as the equivalent characterization of symmetric tensors as homogeneous polynomials. We conclude by describing an algorithm, called the tensor power method, which finds the orthogonal decomposition of an odeco tensor.

3.1.1.1 Symmetric tensor decomposition

Orthogonal decomposition is a special type of *symmetric tensor decomposition* which has been of much interest in the recent years; references include [26, 109, 120], and many others. Given a tensor $T \in S^d(\mathbb{C}^n)$, the aim is to decompose it as

$$T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes d},$$

where $v_1, \ldots, v_r \in \mathbb{C}^n$ are any vectors and $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$. The smallest r for which such a decomposition exists is called the *(symmetric) rank* of T. Finding the symmetric decomposition of a given tensor T is an NP hard problem [92] and algorithms for it have been proposed by several authors, for example [26, 120].

According to the Alexander-Hirschowitz Theorem, when $d \geq 3$ the rank of a generic tensor T is $\left\lceil \frac{\binom{n+d-1}{d}}{n} \right\rceil$ except in a finite number of cases in which it is one more than this number [2]. However, the rank of an odeco tensor $T \in S^d(\mathbb{R}^n)$ is at most n. This means that the set of odeco tensors is a low-dimensional subvariety in the space of all tensors. We explore this further in Section 3.1.3.

Remark 3.1.1. *Orthogonal tensor decomposition has also been studied in the non-symmetric case [98, 99]. An odeco tensor is also orthogonally decomposable according to the definition in the non-symmetric case. We shall return to the non-symmetric case in Section 3.2.*

3.1.1.2 Symmetric tensors as homogeneous polynomials

An equivalent way to think about a symmetric matrix $M \in S^2(\mathbb{C}^n)$ is via its corresponding quadratic form $f_M \in \mathbb{C}[x_1, \ldots, x_n]$ given by

$$f_M(x_1, \ldots, x_n) = x^T M x = \sum_{i,j} M_{ij} x_i x_j.$$

More generally, a tensor $T \in S^d(\mathbb{C}^n)$ can equivalently be represented by a homogeneous polynomial $f_T \in \mathbb{C}[x_1, \ldots, x_n]$ of degree d given by

$$f_T(x_1, \ldots, x_n) = T \cdot x^d := \sum_{i_1, \ldots, i_d = 1}^{n} T_{i_1, \ldots, i_d} x_{i_1} x_{i_2} \ldots x_{i_d}.$$

Given $T \in S^d(\mathbb{C}^n)$, we can describe the notions of eigenvectors, eigenvalues, and symmetric decomposition in terms of the corresponding polynomial $f_T \in \mathbb{C}[x_1, \ldots, x_n]$ as follows.

A vector $x \in \mathbb{C}^n$ is an *eigenvector* of T with eigenvalue λ if and only if

$$\nabla f_T(x) = \lambda d x.$$

The tensor T can be decomposed as $T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes d}$ if and only if the corresponding polynomial f_T can be decomposed as

$$f_T(x_1, \ldots, x_n) = \sum_{i=1}^{r} \lambda_i (v_{i1}x_1 + \cdots + v_{in}x_n)^d.$$

Similarly, a real tensor $T \in S^d(\mathbb{R}^n)$ is orthogonally decomposable with $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_r v_r^{\otimes d}$, where $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ and $v_1, \ldots, v_r \in \mathbb{R}^n$ are orthonormal, if and only if $f_T(x_1, \ldots, x_n) = \lambda_1 (v_1 \cdot x)^d + \cdots + \lambda_r (v_r \cdot x)^d$.

This equivalent characterization of symmetric tensors as homogeneous polynomials proves to be quite useful in the sequel.

3.1.1.3 Finding an orthogonal decomposition

Finding the symmetric decomposition of a general $T \in S^d(\mathbb{C}^n)$ is NP-hard [92]. However, there are efficient algorithms that recover the orthogonal decomposition of an odeco tensor $T \in S^d(\mathbb{R}^n)$ [8, 100]. One such algorithm is the *tensor power method*.

Let $T \in S^d(\mathbb{R}^n)$. If T is orthogonally decomposable, i.e. $T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes d}$ and $v_1, \ldots, v_k \in \mathbb{R}^n$ orthonormal, then

$$T \cdot v_j^{d-1} = \sum_{i=1}^{k} \lambda_i (v_i \cdot v_j)^{d-1} v_i = \lambda_j v_j,$$

for all $j = 1, 2, \ldots, k$. Thus, $v_1, \ldots, v_k$ are eigenvectors of T with corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$. Note that requiring T and $v_1, \ldots, v_k$ to be real forces $\lambda_1, \ldots, \lambda_k$ to be real as well.

Definition 3.1.2. *A unit vector $u \in \mathbb{R}^n$ is a* robust eigenvector *of $T \in S^d(\mathbb{R}^n)$ if there exists $\epsilon > 0$ such that for all $\theta \in \{u' \in \mathbb{R}^n : \|u - u'\| < \epsilon\}$, repeated iteration of the map*

$$\overline{\theta} \mapsto \frac{T\overline{\theta}^{d-1}}{\|T\overline{\theta}^{d-1}\|}, \tag{3.1.1}$$

starting from θ converges to u.

The following theorem shows that if T has an orthogonal decomposition $T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes d}$, then the set of robust eigenvectors of T is precisely the set $\{v_1, v_2, \ldots, v_k\}$, implying that the orthogonal decomposition is unique up to the obvious reordering.

Theorem 3.1.3 (Theorem 4.1, [8])**.** *Let $T \in S^d(\mathbb{R}^n)$, where $d \geq 3$, have an orthogonal decomposition $T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes d}$, where $v_1, \ldots, v_k \in \mathbb{R}^n$ are orthonormal, and $\lambda_1, \ldots, \lambda_k > 0$.*

1. *The set of $\theta \in \mathbb{R}^n$ which do not converge to some v_i under repeated iteration of* (3.1.1) *has measure* 0*.*

2. *The set of robust eigenvectors of T is equal to $\{v_1, v_2, \ldots, v_k\}$.*

Remark 3.1.4. *In fact, the set of $\theta \in \mathbb{R}^n$ which do not converge to some v_i under repeated iteration of* (3.1.1) *is a hyperplane arrangement. This is the set of those* eigenvectors *of the tensor T that are not equal to one of $v_1, \ldots, v_k$, and are described in detail in Theorem 3.1.8.*

Theorem 3.1.3 says that to recover the orthogonal decomposition of T, one needs to find the robust eigenvectors. The definition of robust eigenvectors suggests an algorithm to compute them, using repeated iteration of the map (3.1.1) starting with random vectors $u \in \mathbb{R}^n$.

Algorithm 2 The Tensor Power Method

1: **Input:** an orthogonally decomposable tensor T.
2: Set $i = 1$.
3: **Repeat** until $T = 0$.
4: Choose random $u \in \mathbb{R}^m$.
5: Let v_i be the result of repeated iteration of (3.1.1) starting with u.
6: Compute the eigenvalue λ_i corresponding to v_i, from the equation $Tv_i^{d-1} = \lambda_i v_i$.
7: Set $T = T - \lambda_i v_i^{\otimes d}$.
8: $i \leftarrow i + 1$.
9: **Output** $v_1, \ldots, v_k$ and $\lambda_1, \ldots, \lambda_k$.

In certain cases, this algorithm can be used to find the symmetric decomposition of a given tensor. For example, the authors of [8] consider a class of statistical models, such as the exchangeable single topic model, in which one observes tensors T_2 and T_3, where $T_d = \sum_{i=1}^k \omega_i \mu_i^{\otimes d}$ for $d = 2, 3$ and the aim is to recover the unknown parameters $\omega = (\omega_1, \ldots, \omega_k) \in \mathbb{R}^k$ and $\mu_1, \ldots, \mu_k \in \mathbb{R}^n$. (Note that T_2 and T_3 have decompositions using the same vectors and observing both of them gives more information than observing only T_3). This is done by transforming T_2 and T_3 (in an invertible way) into orthogonally decomposable tensors $\tilde{T}_2$ and $\tilde{T}_3$, where $\tilde{T}_d = \sum_{i=1}^k \tilde{\omega}_i \tilde{\mu}_i^{\otimes d}$ and $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ are orthonormal, $d = 2, 3$. Then, they use the tensor power method to find $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ and $\tilde{\omega}_1, \ldots, \tilde{\omega}_k$ and use the inverse transformation to recover the original $\mu_1, \ldots, \mu_k$ and $\omega_1, \ldots, \omega_k$.

Remark 3.1.5. *As mentioned above, Theorem 3.1.3 also implies that an odeco tensor T has a unique orthogonal decomposition. That is because the elements in the orthogonal decomposition are uniquely determined as the robust eigenvectors $v_1, \ldots, v_k$ and the corresponding constants $\lambda_1, \ldots, \lambda_k$ are uniquely determined by $\lambda_i = T \cdot v_i^d$.*

Another method, described in [26], can also be used to efficiently compute the decomposition of a symmetric tensor T of rank at most n. It involves computing generalized eigenvectors of sub-matrices of the Hankel matrices associated to T.

3.1.2 The Variety of Eigenvectors of a Tensor

In this subsection, we are going to study the set of all eigenvectors of a given orthogonally decomposable tensor.

As we mentioned in the introduction, a symmetric tensor $T \in S^d(\mathbb{R}^n)$ can equivalently be represented by a homogeneous polynomial $f_T \in \mathbb{R}[x_1, \ldots, x_n]$ of degree d. Indeed, given T, we obtain f_T by

$$f_T(x_1, \ldots, x_n) = \sum_{i_1, \ldots, i_d} T_{i_1, \ldots, i_d} x_{i_1} \cdots x_{i_d}.$$

Then, for $x \in \mathbb{C}^n$, $Tx^{d-1} = \lambda x$ is equivalent to $\nabla f_T(x) = d\lambda x$, i.e. $\nabla f_T(x)$ and x are parallel to each other. This is equivalent to the vanishing of the 2×2 minors of the $n \times 2$ matrix $\left[\nabla f_T(x) \,\middle|\, x\right]$.

Definition 3.1.6. *The* variety of eigenvectors $\mathcal{V}_T$ *of a given symmetric tensor T with corresponding polynomial f_T is the zero set of the 2×2 minors of the matrix $\left[\nabla f_T(x) \,\middle|\, x\right]$.*

Remark 3.1.7. *Consider the gradient map as a map on projective spaces:*

$$\nabla f_T : \mathbb{CP}^{n-1} \to \mathbb{CP}^{n-1}$$

$$[x] \mapsto [\nabla f_T(x)].$$

Then, the eigenvectors of f_T are precisely the fixed points of ∇f_T. This map is well-defined provided the hypersurface $\{f_T = 0\}$ has no singular points.

The aim of this subsection is to prove the following theorem.

Theorem 3.1.8. *Let $T \in S^d(\mathbb{R}^n)$ be odeco with $f_T(x) = \sum_{i=1}^{l} \lambda_i (v_i \cdot x)^d$, where $v_1, \ldots, v_l \in \mathbb{R}^n$ are orthonormal. Assume that $1 \leq l \leq n$ and $\lambda_1, \ldots, \lambda_l \neq 0$. Then, T has $\frac{(d-1)^l - 1}{d-2}$ eigenvectors in $\mathbb{C}^n$, given explicitly in terms of $v_1, \ldots, v_l$ and the $(d-2)$-nd roots of $\lambda_1, \ldots, \lambda_l$ as follows. Let $V = \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_l & - \end{bmatrix} \in \mathbb{R}^{l \times n}$. Then, for any $1 \leq k \leq l$, any $\mathcal{I} = \{i_1, i_2, \ldots, i_k\} \subseteq [l]$ and any $(k-1)$-tuple $\eta_1, \ldots, \eta_{k-1}$ of $(d-2)$-nd roots of unity, there is one eigenvector w, up to scaling, where $w = V^T (y_1, \ldots, y_l)^T$ and*

$$y_i = \begin{cases} \eta_j \lambda_{i_j}^{-\frac{1}{d-2}} & \text{if } i = i_j \text{ and } j \in \{1, \ldots, k-1\} \\ \lambda_{i_k}^{-\frac{1}{d-2}} & \text{if } i = i_k \\ 0 & \text{if } i \notin \mathcal{I}. \end{cases}$$

The rest of the eigenvectors are all the elements in the nullspace of V.

Remark 3.1.9. *It is known by [37] that if a tensor $T \in S^d(\mathbb{R}^n)$ has finitely many equivalence classes of eigenpairs (x, λ) over $\mathbb{C}$, then their number, counted with multiplicity, is equal to $\frac{(d-1)^n-1}{d-2}$. If the entries of T are sufficiently generic, then all multiplicities are equal to 1, so there are exactly $\frac{(d-1)^n-1}{d-2}$ equivalence classes of eigenpairs.*

In the proof of Theorem 3.1.8 we independently show that an odeco tensor T with orthogonal decomposition $T = \lambda_1 v_1^{\otimes d} + \cdots + \lambda_n v_n^{\otimes d}$, such that $\lambda_1, \ldots, \lambda_n \neq 0$ has finitely many equivalence classes of eigenvectors and their number is exactly $\frac{(d-1)^n-1}{d-2}$.

Remark 3.1.10. *The explicit formulation of the eingevectors of an odeco tensor given in Theorem 3.1.8 can be used to find the eigenvectors of any tensor $T \in S^d(\mathbb{C}^n)$. This can be done via a homotopy continuation computation with numerical software such as* `Bertini` *[15].*

We illustrate Theorem 3.1.8 by two simple concrete examples.

Example 3.1.11. *Let $d = n = 3$ and consider the odeco tensor T with polynomial form*

$$f_T(x, y, z) = \lambda_1 x^3 + \lambda_2 y^3 + \lambda_3 z^3.$$

This type of polynomial is called a Fermat polynomial. *In this case $v_1 = (1, 0, 0), v_2 = (0, 1, 0), v_3 = (0, 0, 1)$ and the matrix $V = Id_3$. Since $d - 2 = 1$, taking the $(d-2)$-nd root is the identity map. Thus, the eigenvectors of T are as follows.*

When $k = 1$, $\mathcal{I} = \{1\}, \{2\}$, or $\{3\}$. The corresponding three eigenvectors are

$$\left(\frac{1}{\lambda_1}, 0, 0\right)^T, \left(0, \frac{1}{\lambda_2}, 0\right)^T, \left(0, 0, \frac{1}{\lambda_3}\right)^T.$$

When $k = 2$, $\mathcal{I} = \{1, 2\}, \{1, 3\}$, or $\{2, 3\}$. The corresponding eigenvectors are

$$\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, 0\right)^T, \left(\frac{1}{\lambda_1}, 0, \frac{1}{\lambda_3}\right)^T, \left(0, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}\right)^T.$$

When $k = 3$, $\mathcal{I} = \{1, 2, 3\}$ and the corresponding eigenvector is

$$\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}\right)^T.$$

Figure 3.1 shows what these eigenvectors look like geometrically.

Example 3.1.12. *Let $d = 4, n = 4$ and consider $T \in S^4(\mathbb{R}^4)$ with corresponding polynomial*

$$f_T(x_1, \ldots, x_4) = x_1^4 + 2x_2^4.$$

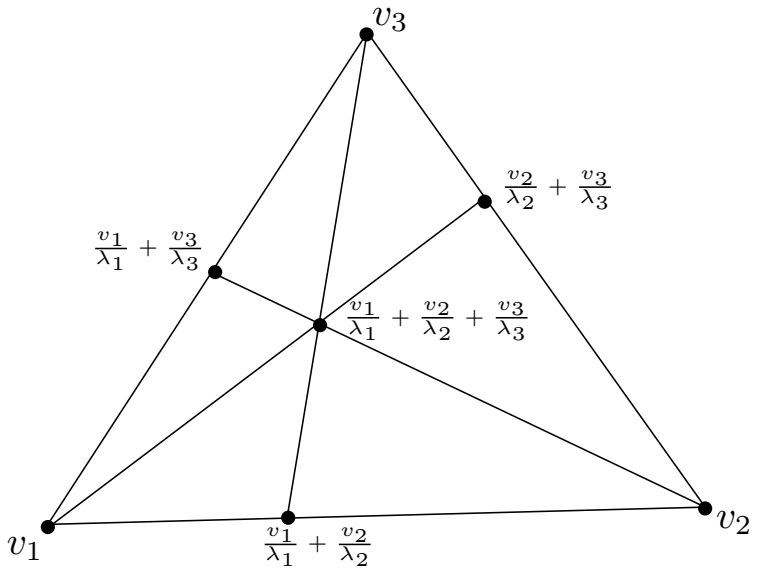


Figure 3.1: This figure shows the structure of the eigenvectors inside $\mathbb{CP}^2$ of an odeco tensor $T \in S^3(\mathbb{R}^3)$ such that $T = \lambda_1 v_1^{\otimes 3} + \lambda_2 v_2^{\otimes 3} + \lambda_3 v_3^{\otimes 3}$ with $\lambda_1, \lambda_2, \lambda_3 \neq 0$.

In the notation of Theorem 3.1.8, the number of nonzero coefficients is $l = 2 < n$. We have that $v_1 = e_1, v_2 = e_2$ and $\lambda_1 = 1, \lambda_2 = 2$. Since $d - 2 = 2$, the roots η_i can be ± 1. Thus, the eigenvectors of T are as follows.

When $k = 1$, $\mathcal{I} = \{1\}, \{2\}$. The corresponding eigenvectors are

$$(1, 0, 0, 0)^T, (0, \frac{1}{\sqrt{2}}, 0, 0)^T.$$

When $k = 2$, $\mathcal{I} = \{1, 2\}$. The corresponding eigenvectors are

$$(1, \frac{1}{\sqrt{2}}, 0, 0)^T, (-1, \frac{1}{\sqrt{2}}, 0, 0)^T.$$

The rest of the eigenvectors are all vectors perpendicular to e_1 and e_2, i.e.

$$(0, 0, a, b)^T$$

for any $a, b \in \mathbb{C}$ not both zero.

In the rest of this subsection we prove Theorem 3.1.8. We proceed as follows. First we show that the theorem is valid when $f_T = \lambda_1 x_1^d + \cdots + \lambda_n v_n^d$, where $\lambda_1, \ldots, \lambda_n \neq 0$. This is done in Lemma 3.1.14. For the general case, $f_T = \lambda_1 (v_1 \cdot x)^d + \cdots + \lambda_l (v_l \cdot x)^d$, where $\lambda_1, \ldots, \lambda_l \neq 0$ and $v_1, \ldots, v_l$ are orthonormal, we observe that setting $y_i = v_i \cdot x$ the eigenvectors of the Fermat polynomial tensor $\lambda_1 y_1^d + \cdots + \lambda_l y_l^d$ are in a 1-to-1 correspondence with some of the eigenvectors of T via the transformation given by the matrix V with rows $v_1, \ldots, v_l$. This is how we recover the formula in Theorem 3.1.8.

Definition 3.1.13. *Given* $f(x_1,\dots,x_n)=\lambda_1x_1^d+\cdots+\lambda_nx_n^d$, $\mathcal{I}=\{i_1,\dots,i_k\}\subseteq\{1,2,\dots,n\}$, *and* $\eta=\{\eta_1,\dots,\eta_{k-1}\}$ *such that* $\eta_1,\dots,\eta_{k-1}$ *are* $(d-2)$*-nd roots of unity, we define the ideal*

$$I_{\mathcal{I},\eta}=\langle\lambda_{i_1}^{\frac{1}{d-2}}x_{i_1}-\eta_1\lambda_{i_k}^{\frac{1}{d-2}},\dots,\lambda_{i_{k-1}}^{\frac{1}{d-2}}x_{i_{k-1}}-\eta_{k-1}\lambda_{i_k}^{\frac{1}{d-2}}x_{i_k}\rangle+\langle x_j|j\notin\mathcal{I}\rangle$$

in the polynomial ring $\mathbb{C}[x_1,\dots,x_n]$.

Lemma 3.1.14. *Theorem 3.1.8 is true in the case* $f_T(x_1,\dots,x_n)=\lambda_1x_1^d+\lambda_2x_2^d+\cdots+\lambda_nx_n^d$, *where* $\lambda_1,\dots,\lambda_n\neq0$. *In particular, the radical of the ideal* I *of* 2×2 *minors of* $\left[\nabla f(x)|x\right]$ *can be decomposed as follows.*

$$\sqrt{I}=\bigcap_{\mathcal{I}\subseteq[n],\eta=\{\eta_1,\dots,\eta_{|\mathcal{I}|-1}\}}I_{\mathcal{I},\eta},\tag{3.1.2}$$

where $\eta_1,\dots,\eta_{k-1}$ *are* $(d-2)$*-nd roots of unity. For every* $k\in\{1,\dots,n\}$, *there are* $\binom{n}{k}(d-2)^{k-1}$ *homogeneous prime ideals* $I_{\mathcal{I},\eta}$ *with* $|\mathcal{I}|=k$. *Each ideal* $I_{\mathcal{I},\eta}$ *has exactly one solution in* $\mathbb{CP}^{n-1}$, *representing one eigenvector, namely* $w=(w_1:\cdots:w_n)$ *such that*

$$w_i=\begin{cases}\eta_l\frac{1}{\lambda_{i_l}}^{-\frac{1}{d-2}} & \text{if } i=i_l \text{ and } l\leq k-1,\\ \lambda_{i_k}^{-\frac{1}{d-2}} & \text{if } i=i_k,\\ 0 & \text{if } i\notin\mathcal{I}.\end{cases}$$

The total number of such solutions is $\frac{(d-1)^n-1}{d-2}$.

Proof. Note that in this case, up to a factor of d in the first row, we have that

$$\left[\nabla f(x)\,\middle|x\right]=\begin{bmatrix}\lambda_1x_1^{d-1} & x_1\\ \lambda_2x_2^{d-1} & x_2\\ \vdots & \vdots\\ \lambda_nx_n^{d-1} & x_n\end{bmatrix}$$

Therefore, the ideal of 2×2 minors is given by

$$I=\langle x_ix_j\left(\lambda_ix_i^{d-2}-\lambda_jx_j^{d-2}\right):i\neq j\rangle.$$

We would like to decompose the variety of this ideal. Note that for any primary ideal $P\supseteq I$ its associated prime $\sqrt{P}$ would either contain x_ix_j or $\lambda_ix_i^{d-2}-\lambda_jx_j^{d-2}$ for all $i\neq j$. Suppose that for a given $P\supseteq I$, $\sqrt{P}$ contains exactly $n-k$ of the variables $x_1,\dots,x_n$. Let $\mathcal{I}=\{i_1,\dots,i_k\}\subseteq[n]$ and assume that $\sqrt{P}$ contains exactly those x_i for which $i\notin\mathcal{I}$. Thus, $\sqrt{P}$ also contains $\lambda_ix_i^{d-2}-\lambda_jx_j^{d-2}$ for $i\neq j,i,j\in\mathcal{I}$. Moreover, we can write $\sqrt{P}$ as $\sqrt{P}=\langle x_i:i\notin\mathcal{I}\rangle+\sqrt{P}\cap\mathbb{C}[x_i:i\in\mathcal{I}]$. Then, the ideal $\sqrt{P}\cap\mathbb{C}[x_i:i\in\mathcal{I}]$ is prime, it doesn't contain x_i for $i\in\mathcal{I}$ and contains $I_{\mathcal{I}}\subseteq\mathbb{C}[x_i:i\in\mathcal{I}]$, where

$$I_{\mathcal{I}}:=\langle\lambda_ix_i^{d-2}-\lambda_jx_j^{d-2}:i\neq j,i,j\in\mathcal{I}\rangle=\langle\lambda_{i_j}x_{i_j}^{d-2}-\lambda_{i_{j+1}}x_{i_{j+1}}^{d-2}:j=1,\dots,k-1\rangle.$$

Therefore, $\sqrt{P} \cap \mathbb{C}[x_i : i \in \mathcal{I}]$ is a prime ideal containing $(I_{\mathcal{I}} : \langle x_i : i \in \mathcal{I} \rangle^{\infty})$.

We now describe the decomposition of the ideal $(I_{\mathcal{I}} : \langle x_i : i \in \mathcal{I} \rangle^{\infty})$ following Theorem 2.1 and Corollary 2.5 in [60]. Recall that $\mathcal{I} = \{i_1, \ldots, i_k\} \subseteq [n]$. Let $L_\rho := \langle (d-2)\left(e_{i_j} - e_{i_k}\right) : j = 1, \ldots, k-1 \rangle$ be a lattice with *partial character* $\rho : L_\rho \to \mathbb{C}^*$ given by

$$\rho\left((d-2)\left(e_{i_j} - e_{i_k}\right)\right) = \frac{\lambda_{i_k}}{\lambda_{i_j}}.$$

For any partial character $\sigma : L_\sigma \to \mathbb{C}^*$, define the ideal $I_+(\sigma) := \langle x^{m_+} - \sigma(m)\, x^{m_-} : m \in L_\sigma \rangle$, where $m = m_+ - m_-$ and m_+, m_- have nonnegative entries. From this definition, we see that

$$I_+(\rho) = (I_{\mathcal{I}} : \langle x_i : i \in \mathcal{I} \rangle^{\infty}).$$

Then, by Corollary 2.5 in [60], the decomposition of $(I_{\mathcal{I}} : \langle x_i : i \in \mathcal{I} \rangle^{\infty}) = I_+(\rho)$ is

$$(I_{\mathcal{I}} : \langle x_i : i \in \mathcal{I} \rangle^{\infty}) = \bigcap_{\rho' \text{ extends } \rho \text{ to } L} I_+(\rho'),$$

where L is a sublattice of $\mathbb{Z}^n$ such that $L_\rho \subseteq L \subseteq \mathbb{Z}^n$ and $|L/L_\rho|$ is finite. In this case, we can choose

$$L = \langle e_{i_j} - e_{i_k} : j = 1, \ldots, k-1 \rangle.$$

Then, $|L/L_\rho| = (d-2)^{k-1}$. Moreover, by the same theorem, the number of ρ' extending ρ is exactly $|L/L_\rho| = (d-2)^{k-1}$. Also, note that each such $\rho' : L \to \mathbb{C}^*$ is uniquely defined by the values

$$\eta_j \left(\frac{\lambda_{i_k}}{\lambda_{i_j}}\right)^{\frac{1}{d-2}} := \rho'\left(e_{i_j} - e_{i_k}\right)$$

for some $(d-2)$-nd root of unity η_j. Therefore,

$$I_+(\rho') = \left\langle x_{i_j} - \eta_j \left(\frac{\lambda_{i_k}}{\lambda_{i_j}}\right)^{\frac{1}{d-2}} x_{i_k} : j = 1, 2, \ldots, k-1 \right\rangle$$

and each such ideal is maximal inside $\mathbb{C}[x_i : i \in \mathcal{I}]$. Thus, the prime $\sqrt{P} \cap \mathbb{C}[x_i : i \in \mathcal{I}]$ must contain one of the ideals $I_+(\rho')$. Therefore, $\sqrt{P}$ contains $\langle x_i : i \notin \mathcal{I} \rangle + I_+(\rho')$ for some ρ'. But this ideal is maximal in $\mathbb{C}[x_1, \ldots, x_n]$, therefore, $\sqrt{P} = \langle x_i : i \notin \mathcal{I} \rangle + I_+(\rho')$.

Therefore, (3.1.2) holds and the minimal associated primes of the ideal I are

$$I_{\mathcal{I},\eta} = \langle x_i : i \notin \mathcal{I} \rangle + \left\langle x_{i_j} - \eta_j \left(\frac{\lambda_{i_k}}{\lambda_{i_j}}\right)^{\frac{1}{d-2}} x_{i_k} : j = 1, 2, \ldots, k-1 \right\rangle,$$

where $\mathcal{I} = \{i_1, \ldots, i_k\} \subseteq [n]$ and $\eta_1, \ldots, \eta_{k-1}$ are $(d-2)$-nd roots of unity. Each ideal $I_{\mathcal{I},\eta}$ is zero-dimensional and corresponds to one eigenvector $w = (w_1 : \cdots : w_n)$, where

$$w_i = \begin{cases} \eta_l \frac{1}{\lambda_{i_l}}^{-\frac{1}{d-2}} & \text{if } i = i_l \text{ and } l \leq k-1, \\ \lambda_{i_k}^{-\frac{1}{d-2}} & \text{if } i = i_k, \\ 0 & \text{if } i \notin \mathcal{I}. \end{cases}$$

Moreover, since there are $\binom{n}{k}$ options for choosing $\mathcal{I} \subseteq [n]$ with $|\mathcal{I}| = k$ and $(d-2)^{k-1}$ options for choosing $\eta = (\eta_1, \ldots, \eta_{k-1})$, the total number of eigenvectors of f is

$$\begin{aligned}\sum_{k=1}^{n} \binom{n}{k} (d-2)^{k-1} &= \frac{1}{d-2} \sum_{k=1}^{n} \binom{n}{k} (d-2)^{k} \\ &= \frac{1}{d-2} \left((d-2+1)^n - 1 \right) = \frac{(d-1)^n - 1}{d-2},\end{aligned}$$

recovering the formula expected by [37]. □

Now, we proceed with the proof of Theorem 3.1.8.

Proof of Theorem 3.1.8. Let $T = \sum_{i=1}^{l} \lambda_i v_i^{\otimes d}$ be odeco with $\lambda_1, \ldots, \lambda_l \neq 0$. Then,

$$f_T(x) = \sum_{i=1}^{l} \lambda_i (v_i \cdot x)^d$$

and

$$\frac{1}{d} \nabla f_T(x) = \sum_{i=1}^{l} \lambda_i (v_i \cdot x)^{d-1} v_i.$$

If $x \in \mathbb{C}^n$ is an eigenvector, then

$$\frac{1}{d} \nabla f_T(x) = \sum_{i=1}^{l} \lambda_i (v_i \cdot x)^{d-1} v_i = \lambda x.$$

Define the vectors $v_{l+1}, \ldots, v_n \in \mathbb{R}^n$ to complete the set of vectors $\{v_1, \ldots, v_l\}$ to an orthonormal basis of $\mathbb{R}^n$. Then, they are also a basis of $\mathbb{C}^n$ and $x = \sum_{i=1}^{n} (v_i \cdot x) v_i$ for any $x \in \mathbb{C}^n$, where $v_i \cdot x = \sum_j v_{ij} x_j$ is still the usual dot product on $\mathbb{R}^n$. Since the v_i form a basis of $\mathbb{C}^n$ and

$$\sum_{i=1}^{l} \lambda_i (v_i \cdot x)^{d-1} v_i = \lambda \sum_{i=1}^{n} (v_i \cdot x) v_i,$$

then x is an eigenvector if and only if the vectors $\left(\lambda_1 (v_1 \cdot x)^{d-1}, \ldots, \lambda_l (v_n \cdot x)^{d-1}, 0, \ldots, 0 \right)$ and $(v_1 \cdot x, \ldots, v_n \cdot x)$ are parallel. Let $\tilde{V} = \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_n & - \end{bmatrix} \in \mathbb{R}^{n \times n}$ be the orthogonal matrix whose rows are $v_1, \ldots, v_n$. Let

$$y_i = (v_i \cdot x), \text{ i.e. } y = \tilde{V} x.$$

Then, an equivalent description of x being an eigenvector is that $\left(\lambda_1 y_1^{d-1}, \ldots, \lambda_l y_l^{d-1}, 0, \ldots, 0\right)$ and y are parallel. In other words, the matrix

$$\begin{bmatrix} \lambda_1 y_1^{d-1} & \cdots & \lambda_l y_l^{d-1} & 0 & \cdots & 0 \\ y_1 & \cdots & y_l & y_{l+1} & \cdots & y_n \end{bmatrix}$$

has rank at most one. There are two cases.

Case 1: One of the numbers $y_{l+1}, \ldots, y_n$ is nonzero. This forces $y_1 = \cdots = y_l = 0$ and any choice of $y_{l+1}, \ldots, y_n$ gives a solution. This means that any vector $x \in \operatorname{span}\{v_1, \ldots, v_l\}^\perp$ is an eigenvector of the original tensor T.

Case 2: The other case is that $y_{l+1} = \cdots = y_n = 0$. Then the above matrix having rank at most one is equivalent to the smaller matrix

$$\begin{bmatrix} \lambda_1 y_1^{d-1} & \cdots & \lambda_l y_l^{d-1} \\ y_1 & \cdots & y_l \end{bmatrix}$$

having rank at most one. The ideal of the 2×2 minors of this matrix is

$$I = \langle \lambda_i y_i^{d-1} y_j - \lambda_j y_j^{d-1} y_i : i < j \leq l \rangle.$$

By Lemma 3.1.14, the radical of this ideal decomposes as

$$\sqrt{I} = \bigcap_{\mathcal{I} \subseteq [l], \eta} I_{\mathcal{I},\eta}$$

and each ideal $I_{\mathcal{I},\eta}$ with $\mathcal{I} = \{i_1, \ldots, i_k\} \subseteq [l]$ has the form

$$I_{\mathcal{I},\eta} = \langle \lambda_{i_1}^{\frac{1}{d-2}} y_{i_1} - \eta_1 \lambda_{i_k}^{\frac{1}{d-2}} y_{i_k}, \ldots, \lambda_{i_{k-1}}^{\frac{1}{d-2}} y_{i_{k-1}} - \eta_{k-1} \lambda_{i_k}^{\frac{1}{d-2}}, y_{i_k} \rangle + \langle y_i : i \notin \mathcal{I} \rangle, \tag{3.1.3}$$

where $\eta_1, \ldots, \eta_{k-1}$ are $(d-2)$-nd roots of unity. By the Nullstellensatz, all elements in $\mathcal{V}(I)$ are the same as those in $\mathcal{V}(\sqrt{I})$, which are in turn the elements in $\bigcup \mathcal{V}(I_{\mathcal{I},\eta})$. Each ideal $I_{\mathcal{I},\eta}$ gives exactly one solution in $\mathbb{P}^n_{\mathbb{C}}$, representing one eigenvector $(y_1, \ldots, y_n)$ such that

$$y_i = \begin{cases} \eta_s \frac{1}{\lambda_{i_s}}^{-\frac{1}{d-2}} & \text{if } i = i_s \text{ and } s \leq k-1, \\ \lambda_{i_k}^{-\frac{1}{d-2}} & \text{if } i = i_k, \\ 0 & \text{if } i \in [n] \setminus \mathcal{I}. \end{cases} \tag{3.1.4}$$

Note that $y = \tilde{V}x$ and $\tilde{V}$ is an orthogonal matrix. Therefore,

$$x = \tilde{V}^T y.$$

By Lemma 3.1.14, we know that for each k there are $\binom{l}{k}(d-2)^{k-1}$ eigenvectors with k nonzero entries, which makes for a total of

$$\sum_{k=1}^{l} \binom{l}{k} (d-2)^{k-1} = \frac{1}{d-2} \left(\sum_{k=1}^{l} \binom{n}{k} (d-2)^k \right)$$

$$= \frac{1}{d-2}\left(\sum_{k=0}^{l}\binom{n}{k}(d-2)^k - 1\right) = \frac{(d-1)^l - 1}{d-2}$$

eigenvectors of T in this case. □

3.1.3 The Odeco Variety

The *odeco variety* is the Zariski closure in $S^d(\mathbb{C}^n)$ of the set of all tensors $T \in S^d(\mathbb{R}^n)$ which are orthogonally decomposable. If a symmetric tensor is odeco, then, in particular, its corresponding polynomial f_T is decomposable as a sum of n d-th powers of linear forms, i.e. it lies in the n-th secant variety of the d-th Veronese variety, denoted by $\sigma_n(v_d(\mathbb{C}^n))$.

When $d = n = 3$, there is one equation defining $\sigma_3(v_3(\mathbb{C}^3))$, called the Aronhold invariant [108], and it is given by the Pfaffian of a certain skew-symmetric matrix. The corresponding odeco variety in $S^3(\mathbb{C}^3)$ has codimension 4 and its prime ideal is generated by six quadrics, defined in Example 3.1.18. For higher d and n, the equations defining $\sigma_n(v_d(\mathbb{C}^n))$ are much harder to compute. However, the odeco variety is smaller than $\sigma_n(v_d(\mathbb{C}^n))$ and we believe that the defining equations of its prime ideal are quadrics that are easy to write down. They are shown in Conjecture 3.1.16, and proven to be correct in Section 4.1.

Lemma 3.1.15. *The dimension of the odeco variety in $S^d(\mathbb{C}^n)$ is $\binom{n+1}{2}$.*

Proof. Consider the map

$$\phi : \mathbb{R}^n \times SO_n \to S^d(\mathbb{R}^n) \subset S^d(\mathbb{C}^n)$$

given by

$$(\lambda_1, \ldots, \lambda_n), V \mapsto \sum_{i=1}^{n} \lambda_i v_i^{\otimes d},$$

where v_i is the ith row of the orthogonal matrix V. The image $\mathrm{Im}(\phi)$ of this map is precisely the set of orthogonally decomposable tensors in $S^d(\mathbb{R}^n)$. The odeco variety is $\overline{\mathrm{Im}(\phi)} \subset S^d(\mathbb{C}^n)$. Note that by Theorem 3.1.3, ϕ has a finite fiber (up to permutations of the input). Then, $\dim(\mathrm{Im}(\phi)) = \dim(\mathbb{R}^n \times SO_n) = n + \binom{n}{2} = \binom{n+1}{2}$. Therefore, the dimension of the odeco variety is $\dim\left(\overline{\mathrm{Im}(\phi)}\right) = \binom{n+1}{2}$. □

We are going to conjecture what the defining equations of the odeco variety are. In Theorem 3.1.20 we prove the result for the case $n = 2$. The general proof was found subsequently in collaboration with Ada Boralevi, Jan Draisma, and Emil Horobeţ [23], and is presented in Section 4.1.

Consider a tensor $T \in S^d(\mathbb{C}^n)$ and the corresponding homogeneous polynomial $f_T(x_1, x_2, \ldots, x_n) \in \mathbb{C}[x_1, \ldots, x_n]$ of degree d. To define our equations, it is more convenient to work

with the polynomial version of the tensor. As mentioned before, given $T \in S^d(\mathbb{C}^n)$, the corresponding polynomial can be rewritten as

$$f_T(x_1, \ldots, x_n) = \sum_{j_1,\ldots,j_d} T_{j_1\ldots j_d} x_{j_1} \ldots x_{j_d}$$

$$= \sum_{i_1+\cdots+i_n=d} \binom{d}{i_1, \ldots, i_n} T_{\underbrace{1\ldots1}_{i_1 \text{ times}} \ldots \underbrace{n\ldots n}_{i_n \text{ times}}} x_1^{i_1} \ldots x_n^{i_n} = \sum_{i_1+\cdots+i_n=d} \frac{1}{i_1! \ldots i_n!} u_{i_1,\ldots,i_n} x_1^{i_1} \ldots x_n^{i_n},$$

where

$$u_{i_1,\ldots,i_n} = d! T_{\underbrace{1\ldots1}_{i_1 \text{ times}} \ldots \underbrace{n\ldots n}_{i_n \text{ times}}}.$$

We write the equations defining the odeco variety in terms of the variables $u_{i_1,\ldots,i_n}$. Note that for all such variables $i_1 + \cdots + i_n = d$.

Conjecture 3.1.16. *The prime ideal of the odeco variety inside $S^d(\mathbb{C}^n)$ is generated by*

$$\sum_{s=1}^{n} u_{y+e_s} u_{v+e_s} - u_{w+e_s} u_{z+e_s} = 0, \tag{3.1.5}$$

where $y, v, w, z \in \mathbb{Z}^n_{\geq 0}$ are such that $\sum_i y_i = \sum_i v_i = \sum_i z_i = \sum_i w_i = d-1$ and $y+v = z+w$.

Written in terms of the T-variables, these equations can be expressed as

$$\sum_{s=1}^{n} T_{i_1,\ldots,i_{d-1},s} T_{j_1,\ldots,j_{d-1},s} - T_{k_1,\ldots,k_{d-1},s} T_{l_1,\ldots,l_{d-1},s} = 0, \tag{3.1.6}$$

for all indices such that $\{i_r, j_r\} = \{k_r, l_r\}$, and also up to permuting the indices due to the fact that T is symmetric.

Another way to think about (3.1.6) is as follows. Suppose we contract T along one of its dimensions, say the d-th dimension, resulting into a tensor $T *_d T \in S^2(S^{d-1}(\mathbb{R}^n))$ whose entry indexed by $i_1, \ldots, i_{d-1}, j_1, \ldots, j_{d-1}$ is

$$(T *_d T)_{i_1,\ldots,i_{d-1},j_1,\ldots,j_{d-1}} = \sum_{s=1}^{n} T_{i_1,\ldots,i_{d-1},s} T_{j_1,\ldots,j_{d-1},s}.$$

Then, the equations (3.1.6) are equivalent to saying that $T *_d T$ also lies inside $S^{2(d-1)}(\mathbb{R}^n)$. In Section 4.1.3.1 we will see that, when $d = 3$, these equations are also equivalent to a certain algebra associated to the tensor T being associative.

Example 3.1.17. *When $d = 2$ the elements of $S^2(\mathbb{R}^n)$ are symmetric matrices and the set of equations (3.1.5) is empty, which is equivalent to the fact that all symmetric matrices are odeco.*

In essence, the ideal defined by (3.1.5) is a lifting of the toric ideal defining the Veronese variety $v_{d-1}(\mathbb{C}^n) \subset S^{d-1}(\mathbb{C}^n)$ to non-toric equations on $S^d(\mathbb{C}^n)$.

Example 3.1.18. *Let $d = n = 3$. We will illustrate how to obtain the equations (3.1.5) of the odeco variety in $S^3(\mathbb{C}^3)$ from the equations of the Veronese variety $v_{d-1}(\mathbb{C}^n) = v_2(\mathbb{C}^3)$. Consider the Veronese embedding $v_2 : \mathbb{C}^3 \to S^2(\mathbb{C}^3)$ given by $x \mapsto x^{\otimes 2}$. The image $v_2(\mathbb{C}^3)$ is the set of rank one 3×3 symmetric matrices. The space $S^2(\mathbb{C}^3)$ has coordinates $u_{i_1 i_2 i_3}$, where $i_1 + i_2 + i_3 = 2$. There are six equations that define the prime ideal of the Veronese variety $v_2(\mathbb{C}^3) \subseteq S^2(\mathbb{C}^3)$ and they are*

$$\begin{aligned} u_{200}u_{020} - u_{110}^2 &= 0, & u_{200}u_{011} - u_{110}u_{101} &= 0, \\ u_{200}u_{002} - u_{101}^2 &= 0, & u_{110}u_{002} - u_{101}u_{011} &= 0, \\ u_{101}u_{020} - u_{110}u_{011} &= 0, & u_{020}u_{002} - u_{011}^2 &= 0. \end{aligned} \tag{3.1.7}$$

Each of these equations has the form $u_y u_v - u_w u_z = 0$, where $y, v, w, z \in \mathbb{Z}^3_{\geq 0}$, $\sum_i y = \sum_i v = \sum_i w = \sum_i z = 2$, and $y + v = w + z$. Each such equation leads to one of the equations in (3.1.5) as follows

$$u_y u_v - u_w u_z \mapsto u_{y+e_1}u_{v+e_1} - u_{w+e_1}u_{z+e_1} + u_{y+e_2}u_{v+e_2} - u_{w+e_2}u_{z+e_2} + u_{y+e_3}u_{v+e_3} - u_{w+e_3}u_{z+e_3}.$$

Therefore, using (3.1.7), we obtain the six equations in (3.1.5)

$$\begin{aligned} u_{200}u_{020} - u_{110}^2 &\mapsto u_{300}u_{120} - u_{210}^2 + u_{210}u_{030} - u_{120}^2 + u_{201}u_{021} - u_{111}^2, \\ u_{200}u_{011} - u_{110}u_{101} &\mapsto u_{300}u_{111} - u_{210}u_{201} + u_{210}u_{021} - u_{120}u_{111} + u_{201}u_{012} - u_{111}u_{102}, \\ u_{200}u_{002} - u_{101}^2 &\mapsto u_{300}u_{102} - u_{201}^2 + u_{210}u_{012} - u_{111}^2 + u_{201}u_{003} - u_{102}^2, \\ u_{110}u_{002} - u_{101}u_{011} &\mapsto u_{210}u_{102} - u_{201}u_{111} + u_{120}u_{012} - u_{111}u_{021} + u_{111}u_{003} - u_{102}u_{012}, \\ u_{101}u_{020} - u_{110}u_{011} &\mapsto u_{201}u_{120} - u_{210}u_{111} + u_{111}u_{030} - u_{120}u_{021} + u_{102}u_{021} - u_{111}u_{012}, \\ u_{020}u_{002} - u_{011}^2 &\mapsto u_{120}u_{102} - u_{111}^2 + u_{030}u_{012} - u_{021}^2 + u_{021}u_{003} - u_{012}^2. \end{aligned}$$

We shall return to this example in Section 4.1.3.1.

Lemma 3.1.19. *The equations (3.1.5) vanish on the odeco variety.*

Proof of Lemma 3.1.19. Let $T = \sum_i \lambda_i v_i^{\otimes d}$ be odeco. Then, by definition of the u-variables, at the point T we have

$$u_{y_1 \dots y_n} = d! \sum_{i=1}^n \lambda_i v_{i_1}^{y_1} \cdots v_{i_n}^{y_n} = d! \sum_{i=1}^n \lambda_i v_i^y.$$

Thus, at the point T, the equations (3.1.5), for $y, v, w, z \in \mathbb{Z}^n_{\geq 0}$ with $y + v = w + z$ and $\sum_i y = \sum_i v = \sum_i w = \sum_i z = d - 1$, have the form

$$\begin{aligned}
&\sum_{(} s = 1^n u_{y+e_s} u_{v+e_s} - u_{w+e_s} u_{z+e_s} = \\
&= (d!)^2 \sum_{s=1}^n \big(\sum_{i=1}^n \lambda_i v_i^{y+e_s} \big) \big(\sum_{j=1}^n \lambda_j v_j^{v+e_s} \big) - \big(\sum_{i=1}^n \lambda_i v_i^{w+e_s} \big) \big(\sum_{j=1}^n \lambda_j v_j^{z+e_s} \big) \\
&= (d!)^2 \sum_{s=1}^n \big(\sum_{i=1}^n \lambda_i^2 (\cancel{v_i^{y+v+2e_s}} - \cancel{v_i^{w+z+2e_s}}) + \sum_{i \neq j} \lambda_i \lambda_j (v_i^{y+e_s} v_j^{v+e_s} - v_i^{w+e_s} v_j^{z+e_s}) \big) \\
&= (d!)^2 \sum_{i \neq j} \lambda_i \lambda_j (v_i^y v_j^v - v_i^w v_j^z) \sum_{s=1}^n v_{is} v_{js} = 0,
\end{aligned}$$

where the last row is 0 since v_i and v_j are orthogonal and $\sum_{s=1}^n v_{is} v_{js} = v_i \cdot v_j = 0$

Therefore, (3.1.5) vanish on the odeco variety. □

We are going to select a subset of the equations (3.1.5) that spans the vector space defined by (3.1.5). More precisely, consider

$$f_{y,v,i,j} = \sum_{s=1}^n u_{y+e_s} u_{v+e_s} - u_{y+e_i-e_j+e_s} u_{v-e_i+e_j+e_s}, \tag{3.1.8}$$

for all $i \neq j \in \{1, 2, \ldots, n\}$ and all $y, v \in \mathbb{Z}^n_{\geq 0}$ whose entries sum to $d-1$ and $y_j \geq 1$, $v_i \geq 1$.

We now prove Conjecture 3.1.16 for the case $n = 2$.

Theorem 3.1.20. *When $n = 2$, the equations (3.1.8) form a Gröbner basis with respect to the term order $\prec$ (defined below as a refinement of the weight order* (3.1.10)*) and the dimension of the variety they cut out is $\binom{n+1}{2} = 3$. The ideal defined by (3.1.8) is the prime ideal of the odeco variety.*

Proof. We are going to work over the polynomial ring

$$\begin{aligned}
\mathbb{C}[\mathbf{u}] &:= \mathbb{C}[u_{i_1 i_2} | i_1, i_2 \geq 0 \text{ and } i_1 + i_2 = d] \\
&= \mathbb{C}[u_{d0}, u_{(d-1)1}, \ldots, u_{0d}].
\end{aligned}$$

Then, the equations (3.1.8) are

$$f_{y,v,1,2} = u_{y+e_1} u_{v+e_1} - u_{y+e_1-e_2+e_1} u_{v-e_1+e_2+e_1} + u_{y+e_2} u_{v+e_2} - u_{y+e_1-e_2+e_2} u_{v-e_1+e_2+e_2},$$

where $y, v \in \mathbb{Z}^2_{\geq 0}$, the sum of the entries of each of y and v is $d-1$ and $y_2 \geq 1, v_1 \geq 1$. Let the ideal they generate be

$$I := \langle f_{y,v,1,2} | y, v \in \mathbb{Z}^2_{\geq 0}, \sum_i y_i = \sum_i v_i = d - 1, y_2 \geq 1, v_1 \geq 1 \rangle. \tag{3.1.9}$$

We introduce the following weights on our variables. Let

$$\text{weight}\left(u_{i(d-i)}\right) = i, \tag{3.1.10}$$

for all $i = 0, 1, \ldots, d$. Consider the weighted term order on monomials $\prec$ given by the above weights, refined by the lexicographic term order such that $u_{d0} \succ u_{(d-1)1} \succ \cdots \succ u_{0d}$ in case of equal weights.

We first show that the equations (3.1.8) form a Gröbner basis with respect to $\prec$. Using `Macaulay2`, we have shown that they form a Gröbner basis for $d = 1, 2, \ldots, 9$. Now, consider any $d > 9$. Take $f_{y',v',1,2}$ and $f_{y'',v'',1,2}$. By Buchberger's second criterion, we only need to consider the two polynomials when their initial terms have a common variable. Then, the two polynomials $f_{y',v',1,2}$ and $f_{y'',v'',1,2}$ contain $l \leq 9$ different variables in total. If we restrict our generators (3.1.8) to these l variables only, the restriction of the term order is the same as the term order in the case $d = l - 1$, and we have shown that in this case, the restricted generators form a Gröbner basis. Therefore, we can reduce the S-pair of $f_{y',v',1,2}$ and $f_{y'',v'',1,2}$ to 0 using the generators (3.1.8). Thus, the equations (3.1.8) form a Gröbner basis.

Next, we show that the ideal I generated by (3.1.8) has dimension 3. One way to see this is to use Lemma 3.1.21 together with the fact that I is prime, which is proven below. Another way to see that $\dim I = 3$ is to reason with standard monomials as follows.

Note that because of our choice of term order $\prec$, the initial term of every $f_{u,v,1,2}$ is square-free. The reason is that if $u_{y+e_s} = u_{v+e_s}$, then, $\text{weight}(u_{y+e_1}u_{v+e_1}) = \text{weight}(\ u_{y+e_1-e_2+e_1}\ u_{v-e_1+e_2-e_1}) > \text{weight}(\ u_{y+e_2}u_{v+e_2}) = \text{weight}(u_{y+e_1-e_2+e_2}\ u_{v-e_1+e_2-e_2})$, but $u_{y+e_1-e_2+e_1}$ appears first in $\prec$, so, $u_{y+e_1-e_2+e_1}u_{v-e_1+e_2-e_1}$ is the leading term. The reasoning is similar if $u_{y+e_1-e_2+e_2} = u_{v-e_1+e_2-e_1}$. Therefore, $\text{in}_\prec I$ (and thus I) is a radical ideal.

To show that $\dim I = 3$, let $S = \{u_{i_1(d-i_1)}, u_{i_2(d-i_2)}, u_{i_3(d-i_3)}, u_{i_4(d-i_4)}\}$ be a set of four variables, where $i_1 > i_2 > i_3 > i_4$. We will show that there is a monomial with only variables from S which is not standard. This would mean that $\dim I \leq 3$. Indeed, consider

$$f_{(i_1-1,d-i_1+1),(i_3+1,d-i_3-1),1,2} = u_{(i_1-1)(d-i_1+1)}u_{(i_3+1)(d-i_3+1)} - \underline{u_{i_1(d-i_1)}u_{i_3(d-i_3)}}$$

$$+u_{(i_1-2)(d-i_1+2)}u_{i_2(d-i_2)} - u_{(i_1-1)(d-i_1+1)}u_{(i_2-1)(d-i_2+1)}.$$

Since $i_1 - 2 \geq i_3$, the initial term is $u_{i_1(d-i_1)}u_{i_3(d-i_3)}$. Therefore, $\dim I \leq 3$.

Now, consider the set $S = \{u_{2(d-2)}, u_{1(d-1)}, u_{0d}\}$. Suppose there exists

$$f_{y,v,1,2} = u_{y+e_1}u_{v+e_1} - u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1} + u_{y+e_2}u_{v+e_2} - u_{y+e_1-e_2+e_2}u_{v-e_1+e_2+e_2},$$

such that $\text{in}_\prec(f)$ has both of its variables in S. We know that $\text{in}_\prec(f) = u_{y+e_1}u_{v+e_1}$ or $\text{in}_\prec(f) = u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1}$. Moreover, if $y = (y_1, y_2)$ and $v = (v_1, v_2)$, then, $y_2, v_1 \geq 1$ and $y_1, v_2 \leq d - 2$. Thus, if $\text{in}_\prec(f) = u_{y+e_1}u_{v+e_1}$ and $u_{y+e_1}, u_{v+e_1} \in S$, then, $v = (1, d-2)$ and $y = (1, d-2)$ or $y = (0, d-1)$. Since $f_{y,v,1,2}$ is not the trivial polynomial 0, then, $y \neq (0, d-1)$. Thus, $y = (1, d-2)$. But this is impossible since $\text{in}_\prec(f)$ is square-free for

every generator f. If $\text{in}_\prec(f) = u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1}$ and $u_{y+e_1-e_2+e_1}, u_{v-e_1+e_2+e_1} \in S$, then, $u_{(y_1+2,y_1-1)} \in S$. But $y_1 \geq 1$, so, $y_1 + 2 \geq 3$, therefore, $u_{(y_1+2,y_2-1)} \notin S$. In any case, there can't be a monomial with only variables in S, which is a leading term of an element in I. Thus, $\dim I = 3$.

Another way to see that $\dim I \geq 3$ is by noting that $V(I)$ contains the odeco variety, which has dimension 3 in this case.

Finally, we show that the ideal generated by (3.1.8) is prime. Let J be the ideal generated by the leading binomials of the elements in (3.1.8) with respect to the weight order defined by (3.1.10) (without considering the refinement given by the order of the variables). Denote by g_w the leading term of a polynomial g just with respect to this weight order. Then, $(f_{y,v,1,2})_w = u_{y+e_1}u_{v+e_1} - u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1}$, and $J = \langle u_{y+e_1}u_{v+e_1} - u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1} : y, v \in \mathbb{Z}^2_{\geq 0}, y_1 + y_2 = v_1 + v_2 = d-1, y_2, v_1 \geq 1\rangle$. The ideal J is the prime ideal of the rational normal curve; in particular, it is prime. Moreover, by Proposition 1.13 in [146], $\text{in}_\prec(I) = \text{in}_\prec(J)$. Therefore, $\text{in}_\prec(I)$ is an initial ideal of both I and J. In the following paragraph, we show that J is the initial ideal of I with respect to the weight order given by (3.1.10). Then, since J is prime, it follows that I is prime.

Suppose J is not initial, i.e. there exists $g \in I$ such that $g_w \notin J$. Choose g with $\text{in}_\prec(g)$ as small as possible. Since the elements $f_{u,v,1,2}$ form a Gröbner basis of I, then, there exist y, v such that $\text{in}_\prec(g)$ is divisible by $\text{in}_\prec(f_{y,v,1,2})$. Then, $g = \alpha_{y,v}f_{y,v,1,2} + g_1$, where $\alpha_{y,v}$ is a monomial and $\text{in}_\prec g_1 \prec \text{in}_\prec g$. But note that then, $g_w = \alpha_{y,v}(u_{y+e_1}u_{v+e_1} - u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1}) + (g_1)_w$. Since $u_{y+e_1}u_{v+e_1} - u_{y+e_1-e_2+e_1}u_{v-e_1+e_2+e_1} \in J$ and $g_w \notin J$, then, $(g_1)_w \notin J$. But this is a contradiction since $\text{in}_\prec(g_1) \prec \text{in}_\prec(g)$ and we chose $\text{in}_\prec(g)$ to be as small as possible such that $g_w \notin J$.

Therefore, J is initial. Since it is prime, then, I is also prime. By Lemma 3.1.21, the dimension of the odeco variety for $n = 2$ is 3. Moreover, it is contained in $\mathcal{V}(I)$. Since $\mathcal{V}(I)$ is also irreducible and has dimension 3, then, I is exactly the prime ideal of the odeco variety. □

3.1.3.1 Evidence for Conjecture 3.1.16

Lemma 3.1.21. *The odeco variety is an irreducible component of $\mathcal{V}(I)$, where I is the ideal generated by the equations (3.1.5).*

Proof. We show that the dimension of the component of $\mathcal{V}(I)$ containing the odeco variety is equal to $\binom{n+1}{2}$. This equals the dimension of the odeco variety. Since it is irreducible, then it is an irreducible component of $\mathcal{V}(I)$.

Consider the point $T \in \mathcal{V}(I)$ given by $T_{i\ldots i} = 1$ for all $i = 1, \ldots, n$ and all other entries of T are 0. The polynomial corresponding to T is the standard Fermat polynomial $f_T(x_1, \ldots, x_n) = x_1^d + \cdots + x_n^d$. In the u coordinates, T is represented by the point for which $u_{0\ldots0d0\ldots0} = u_{de_i} = 1$ for $i = 1, \ldots, n$ and all other $u_{i_1\ldots i_n} = 0$.

We can select generators $f_{v,w}$ for I such that $v, w \in \mathbb{Z}^n_{\geq 0}$ with $\sum_i v_i = \sum_i w_i = d-1$ and

$$f_{v,w} = \sum_{i=1}^{s} u_{v+e_s} u_{w+e_s} - u_{\mathrm{sort}(v,w)_1+e_s} u_{\mathrm{sort}(v,w)_2+e_s},$$

where $\mathrm{sort}(v,w)_1$ and $\mathrm{sort}(v,w)_2$ are defined as follows. Given v and w, form the corresponding sequences $t(v) = \underbrace{1\ldots1}_{v_1 \text{ times}} \underbrace{2\ldots2}_{v_2 \text{ times}} \ldots \underbrace{n\ldots n}_{v_n \text{ times}}$ and $t(w) = \underbrace{1\ldots1}_{w_1 \text{ times}} \underbrace{2\ldots2}_{w_2 \text{ times}} \ldots \underbrace{n\ldots n}_{w_n \text{ times}}$. Let $t(v,w) = \mathrm{sort}(t(v) \cup t(w))$ be the sequence obtained by concatenating $t(v)$ and $t(w)$ and then sorting. Let $t(v,w)_1$ be the subsequence of elements in odd positions and $t(v,w)_2$ the subsequence of elements in even positions. Define $u_{\mathrm{sort(v,w)}_1}$ and $u_{\mathrm{sort}(v,w)_2}$ be the corresponding u variables. The fact that the polynomials $f_{u,w}$ generate I follows from Theorem 14.2 in [146].

We form the Jacobian $\mathcal{J}$ of I at the point T. Index the rows of $\mathcal{J}$ by the generators $f_{v,w}$ and index the columns by the variables $u_{i_1,\ldots,i_n}$. Note that $\frac{\partial f}{\partial u_{de_i}}|_T = 0$ since the monomials in $f_{v,w}$ containing u_{de_i} contain another variable $u_{i_1,\ldots,i_n} \neq u_{de_j}$ for all $j = 1,\ldots,n$. Therefore, the column corresponding to u_{de_i} is zero.

Note that the monomials $u_{\mathrm{sort}(v,w)_1+e_s} u_{\mathrm{sort}(v,w)_2+e_s}$ cannot contain a variable u_{de_i} for any v and w that give a nontrivial $f_{u,v}$, so they don't matter in the Jacobian analysis.

Now, the column of $\mathcal{J}$ corresponding to the variable $u_{(d-1)e_i+e_j}$ for $i \neq j$ has 1 only in the rows corresponding to $f_{(d-1)e_i,(d-1)e_j}$ and so does the variable $u_{(d-1)e_j+e_i}$. Therefore, the variables $u_{(d-1)e_i+e_j}$ and the polynomials $f_{(d-1)e_i,(d-1)e_j}$ form a block in $\mathcal{J}$ of rank $\binom{n}{2}$, which equals the number of pairs $i \neq j$.

For any other variable $u_{i_1,\ldots,i_n}$, such that $(i_1,\ldots,i_n) \neq de_i$ or $(d-1)e_i + e_j$, its corresponding column is nonzero only at the rows corresponding to the polynomials $f_{(i_1,\ldots,i_n)-e_s,(d-1)e_s}$ for all s such that $i_s > 0$. Each such polynomial has no other 1's in its row except for the one at $u_{i_1,\ldots,i_n}$. Therefore, each variable $u_{i_1,\ldots,i_n}$, such that $(i_1,\ldots,i_n) \neq de_i$ or $(d-1)e_i + e_j$, contributes a size $1 \times \{\#s : i_s > 0\}$ nonzero block to $\mathcal{J}$, so it contributes 1 to the rank. Therefore, the rank of $\mathcal{J}$ is

$$\#\text{ variables } - \#\{u_{de_i}\} - \#\{u_{(d-1)e_i+e_j : i\neq j}\} + \binom{n}{2}$$

$$= \#\text{ variables } - n - n(n-1) + \binom{n}{2} = \#\text{ variables } - \binom{n+1}{2}.$$

Thus, the rank of the Jacobian at a smooth point in the irreducible component of T is at least $\#$ variables $- \binom{n+1}{2}$, so the dimension of an irreducible component containing T is at most $\binom{n+1}{2}$.

Since the odeco variety is irreducible, has dimension $\binom{n+1}{2}$, contains T, and is contained in $\mathcal{V}(I)$, then it is one of the irreducible components of $\mathcal{V}(I)$. □

Lemma 3.1.21 shows that one only needs to show that the ideal I is prime in order to confirm Conjecture 3.1.16.

Computations

In Figure 3.2 we show some computational checks of the conjecture.

Since the ideal I becomes quite large, as n and d grow, it soon becomes hard to check its primality. It was easy to check the conjecture was correct in the case $n = d = 3$ using `Macaulay2`. The case $n = 3, d = 4$ was checked using the numerical homotopy software Bertini. We were unable to confirm the rest of the results using (short) computations.

n	d	dimension	degree	# min. gens.	conjecture check
3	3	6	10	6	True
3	4	6	35	27	True
3	5	6	84	75	
4	3	≥ 10		20	
4	4	≥ 10		126	
5	3	≥ 15		50	

Figure 3.2: A table of what can be found computationally about the ideal I generated by the equations in (3.1.5).

Acknowledgements

I would like to thank my advisor Bernd Sturmfels for his great help in this project. I would also like to thank Kaie Kubjas and Luke Oeding for helpful comments and Matthew Niemerg for his help with the software Bertini. I was supported by a UC Berkeley Graduate Fellowship and by the National Institute of Mathematical Sciences (NIMS) in Daejeon, Korea.

3.2 Singular Vectors of Orthogonally Decomposable Tensors

Orthogonal decomposition of tensors is a generalization of the singular value decomposition of matrices. In this section, we study the spectral theory of orthogonally decomposable tensors. For such a tensor, we give a description of its singular vector tuples as a variety in a product of projective spaces. This is joint work with Anna Seigal titled *Singular vectors of orthogonally decomposable tensors* [132].

3.2.1 Introduction

The singular value decomposition of a matrix $M \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ expresses it in the form

$$M = V^{(1)}\Sigma(V^{(2)})^T = \sum_{i=1}^{n} \sigma_i v_i^{(1)} \otimes v_i^{(2)}, \tag{3.2.1}$$

where $V^{(1)} \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_1}$ and $V^{(2)} \in \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_2}$ are orthogonal matrices. The vectors $v_1^{(1)}, \ldots, v_{n_1}^{(1)}$ and $v_1^{(2)}, \ldots, v_{n_2}^{(2)}$ are the columns of the matrices $V^{(1)}$ and $V^{(2)}$ respectively. The matrix Σ is diagonal of size $n_1 \times n_2$ with non-negative diagonal entries $\sigma_1, ..., \sigma_n$, where $n = \min\{n_1, n_2\}$. The singular value decomposition of a matrix is extremely useful for studying matrix-shaped data coming from applications. For example, it allows the best low-rank approximation of a matrix to be found.

In light of the excellent properties of the singular value decomposition, and of the prevalence of tensor data coming from applications, it is a topic of major interest to extend the singular value decomposition to tensors. In fact it is even more crucial to find a low rank approximation of a tensor than it is for a matrix: the greater number of dimensions makes tensors in their original form especially computationally intractable. In this section we investigate those tensors for which the singular value decomposition is possible. We note that our singular value decomposition is not valid for all tensors of a given format, which makes it more stringent than that in [110], which is based on flattenings of the tensor.

Definition 3.2.1. *A tensor $T \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \cdots \otimes \mathbb{R}^{n_d}$ is* orthogonally decomposable*, or* odeco*, if it can be written as*

$$T = \sum_{i=1}^{n} \sigma_i v_i^{(1)} \otimes v_i^{(2)} \otimes \cdots \otimes v_i^{(d)},$$

where $n = \min\{n_1, \ldots, n_d\}$, the scalars $\sigma_i \in \mathbb{R}$, and the vectors $v_1^{(j)}, v_2^{(j)}, \ldots, v_n^{(j)} \in \mathbb{R}^{n_j}$ are orthonormal for every fixed $j \in \{1, \ldots, d\}$.

We remark that in the above decomposition for T it is sufficient to sum up to $n = \min\{n_1, \ldots, n_d\}$ since there are at most n_j orthonormal vectors in $\mathbb{R}^{n_j}$ for every $j = 1, \ldots, d$. Such a decomposition will in general be unique up to re-ordering the summands.

Odeco tensors have been studied in the past due to their appealing properties [8, 6, 98, 99, 131, 160]. Finding the decomposition of a general tensor is NP-hard [92], however finding the decomposition of an odeco tensor can be done efficiently via a few different methods [100, 160].

The variety of odeco tensors is studied in Section 4.1, and the eigenvectors of symmetric odeco tensors of format $n \times \cdots \times n$ were studied in Section 3.1. Here we focus on odeco tensors of format $n_1 \times \cdots \times n_d$ that need not be symmetric, and whose dimensions n_i need not be equal. As with matrices, when the dimensions n_i are not equal, it is no longer possible to define eigenvectors. The right notion is now that of a singular vector tuple.

Definition 3.2.2. *A* singular vector tuple *of a tensor $T \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ is a d-tuple of nonzero vectors $(x^{(1)}, \ldots, x^{(d)}) \in \mathbb{C}^{n_1} \times \cdots \times \mathbb{C}^{n_d}$ such that*

$$T(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots, x^{(d)}) \text{ is parallel to } x^{(j)}, \text{ for all } j = 1, \ldots, d. \tag{3.2.2}$$

The left hand side of equation (3.2.2) *is the vector obtained by contracting T by the vector $x^{(k)}$ along its k-th dimension for all $k \neq j$.*

Since this setup is invariant under scaling each vector $x^{(j)}$, we consider the singular vector tuple $(x^{(1)}, \ldots, x^{(d)})$ to lie in the product of projective spaces $\mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}$.

The singular vector tuples of a tensor can also be characterized via a variational approach, as in [112]. They are the critical points of the optimization problem

$$\begin{aligned} \text{maximize} \quad & T(x^{(1)}, \ldots, x^{(d)}) \\ \text{subject to} \quad & ||x^{(1)}|| = \cdots = ||x^{(d)}|| = 1, \end{aligned}$$

where we note that the global maximizer gives the best rank-one approximation of the tensor.

Given a decomposition of an odeco tensor $T = \sum_{i=1}^{n} \sigma_i v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}$, it is straightforward to see that the tuples $(v_i^{(1)}, \ldots, v_i^{(d)})$ corresponding to the rank-one tensors in the decomposition are singular vector tuples. For generic matrices $M \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ the rank-one terms in the singular value decomposition constitute all of the singular vector pairs. In contrast, odeco tensors have additional singular vector tuples that do not appear as terms in the decomposition.

Remark 3.2.3. *When $(x^{(1)}, \ldots, x^{(d)})$ is a singular vector tuple of T, we distinguish between the cases $T(x^{(1)}, \ldots, x^{(d)}) = 0$ and $T(x^{(1)}, \ldots, x^{(d)}) \neq 0$. This is equivalent to whether or not the vectors in* (3.2.2) *are zero for all $j = 1, \ldots, d$ (by the definition of a singular vector tuple). In the former case, the singular vector tuple is a* base point *of the following map of projective space induced by T:*

$$\mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1} \to \mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}$$

$$(x^{(1)}, \ldots, x^{(d)}) \mapsto \left(T(\cdot, x^{(2)}, \ldots, x^{(d)}), \ldots, T(x^{(1)}, \ldots, x^{(d-1)}, \cdot)\right).$$

In the latter case the singular vector tuple is a fixed point *of this map.*

Our main theorem is the following description of the singular vector tuples of an odeco tensor:

Theorem 3.2.4. *The projective variety of singular vector tuples of an odeco tensor* $T \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \cdots \otimes \mathbb{R}^{n_d}$ *is a subvariety of* $\mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}$ *and consists of*

$$\frac{(2^{d-1}(d-2)+1)^n - 1}{2^{d-1}(d-2)}$$

fixed points, and an arrangement of base points. The base points comprise $\binom{d}{2}^n - c(d-1)^n + \binom{c}{2}$ *components, each of dimension* $\sum_{j=1}^{d}(n_j - 1) - 2n$*, that are products of linear subspaces of each* $\mathbb{P}^{n_j-1}$*. Here,* $n = min\{n_1, \ldots, n_d\}$ *and* $c = \#\{j : n_j = n\}$*.*

In particular, for all but a few small cases, the singular vector tuples of an odeco tensor comprise a positive-dimensional variety. In contrast, the variety of singular vector tuples of a generic tensor is zero-dimensional [69]. It is interesting to study how the positive-dimensional components of the singular vector variety for an odeco tensor adopt generic behavior under a small perturbation. Note the contrast to the variety of eigenvectors of a general symmetric odeco tensor, which is also zero-dimensional by Theorem 3.1.8.

The rest of this section is organized as follows. In Subsection 3.2.2, we use the theory of binomial ideals [60] to describe the singular vector tuples of an odeco tensor. In Subsection 3.2.3 we conclude the proof of our theorem by describing the positive-dimensional components of the variety of singular vector tuples. Finally, in Subsection 3.2.4, we explore the structure of these components in more detail by studying specific examples.

3.2.2 Description of the Singular Vector Tuples

In this section we give a formula for the singular vector tuples of an odeco tensor. We start by considering a diagonal odeco tensor.

Lemma 3.2.5. *Let* $S \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ *be the tensor*

$$S = \sum_{i=1}^{n} \sigma_i e_i^{(1)} \otimes \cdots \otimes e_i^{(d)},$$

where $\sigma_1, \ldots, \sigma_n \neq 0$*, the vector* $e_i^{(j)}$ *is the* i*th basis vector in* $\mathbb{R}^{n_j}$*, and* $n = \min\{n_1, \ldots, n_d\}$*. The singular vector tuples* $(x^{(1)}, \ldots, x^{(d)}) \in \mathbb{P}^{n_1-1} \times \ldots \times \mathbb{P}^{n_d-1}$ *of* S *are given as follows.*

Type I: Tuples $(x^{(1)}, \ldots, x^{(d)})$ *of the form*

$$\sigma_{\tau(1)}^{-\frac{1}{d-2}} \left(e_{\tau(1)}^{(1)}, e_{\tau(1)}^{(2)}, \ldots, e_{\tau(1)}^{(d)} \right) + \sum_{i=1}^{m} \eta_i \sigma_{\tau(i)}^{-\frac{1}{d-2}} \left(e_{\tau(i)}^{(1)}, \chi_i^{(2)} e_{\tau(i)}^{(2)}, \ldots, \chi_i^{(d)} e_{\tau(i)}^{(d)} \right) \tag{3.2.3}$$

where $1 \leq m \leq n$*, the scalars* $\chi_i^{(j)} \in \{\pm 1\}$ *are such that* $\prod_{j=2}^{d} \chi_i^{(j)} = 1$ *for every* $i = 1, ..., m$*, each scalar* η_i *is a* $(2d-4)$*-th root of unity, and* τ *is any permutation on* $\{1, \ldots, n\}$*.*

Type II: *All tuples* $(x^{(1)}, \ldots, x^{(d)})$ *such that the* $n \times d$ *matrix* $X = (x_i^{(j)})_{1 \leq i \leq n, 1 \leq j \leq d}$ *has at least two zeros in each row. Since each* $x^{(j)} \in \mathbb{P}^{n_j - 1}$*, we further require that no* $x^{(j)}$ *has all coordinates equal to zero.*

Before proving Lemma 3.2.5, we illustrate it by way of the following example:

Example 3.2.6. *Consider the odeco tensor* $S = e_1 \otimes e_1 \otimes e_1 + e_2 \otimes e_2 \otimes e_2 \in \mathbb{R}^2 \otimes \mathbb{R}^3 \otimes \mathbb{R}^3$*. Its Type I singular vector tuples are*

$$\left(e_1^{(1)}, e_1^{(2)}, e_1^{(3)}\right), \left(e_2^{(1)}, e_2^{(2)}, e_2^{(3)}\right)$$
$$\left(e_1^{(1)} + e_2^{(1)}, e_1^{(2)} + e_2^{(2)}, e_1^{(3)} + e_2^{(3)}\right), \left(e_1^{(1)} + e_2^{(1)}, e_1^{(2)} - e_2^{(2)}, e_1^{(3)} - e_2^{(3)}\right),$$
$$\left(e_1^{(1)} - e_2^{(1)}, e_1^{(2)} + e_2^{(2)}, e_1^{(3)} - e_2^{(3)}\right), \left(e_1^{(1)} - e_2^{(1)}, e_1^{(2)} - e_2^{(2)}, e_1^{(3)} + e_2^{(3)}\right),$$

The Type II singular vectors make five copies of $\mathbb{P}^1$*, namely*

$$\left(\square e_1^{(1)} + \square e_2^{(1)}, e_3^{(2)}, e_3^{(3)}\right), \left(e_1^{(1)}, \square e_2^{(2)} + \square e_3^{(2)}, e_3^{(3)}\right), \left(e_1^{(1)}, e_3^{(2)}, \square e_2^{(3)} + \square e_3^{(3)}\right),$$
$$\left(e_1^{(1)}, \square e_1^{(2)} + \square e_3^{(2)}, e_3^{(3)}\right), \left(e_2^{(1)}, e_3^{(2)}, \square e_1^{(3)} + \square e_3^{(3)}\right),$$

where two $\square$*'s in a vector indicate a copy of* $\mathbb{P}^1$ *on those two coordinates. The five copies of* $\mathbb{P}^1$ *intersect in two triple intersections, as seen in Figure 3.3.*

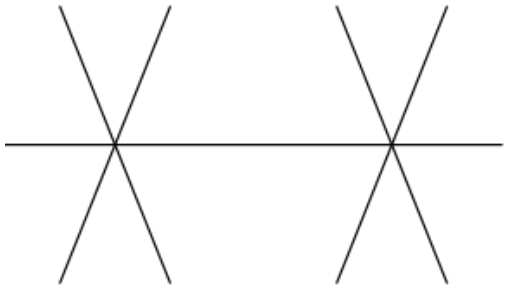


Figure 3.3: The Type II singular vectors: five copies of $\mathbb{P}^1$ meeting at two triple intersections

According to [69], the generic number of singular vector tuples of a tensor of this size is 15, so the five copies of $\mathbb{P}^1$ *degenerate from nine points. For example, consider the family of perturbed tensors*

$$S_\epsilon = S + \epsilon T,$$

where T is the $2 \times 3 \times 3$ tensor with slices $T_{1,\cdot,\cdot}$ and $T_{2,\cdot,\cdot}$ given by

$$T_{1,\cdot,\cdot} = \begin{pmatrix} 0 & 40 & 10 \\ 100 & 3 & 3 \\ 3 & 2 & 6 \end{pmatrix}, \qquad T_{2,\cdot,\cdot} = \begin{pmatrix} 7 & 1 & 1 \\ 8 & 0 & 2 \\ 2 & 2 & 3 \end{pmatrix}.$$

For ϵ on the order of 10^{-6} we attain nine points: one point near each copy of $\mathbb{P}^1$, and two points of multiplicity 2 near each triple intersection.

We will return this example in Section 3.2.4.

Proof of Lemma 3.2.5. By definition, $(x^{(1)}, \ldots, x^{(d)})$ is a singular vector tuple of S if and only if for each $j = 1, \ldots, d$ the following matrix has rank at most one:

$$M_{S,j} = \left[S(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots, x^{(d)}) \quad | \quad x^{(j)} \right] = \begin{bmatrix} \sigma_1 x_1^{(1)} \cdots \hat{x}_1^{(j)} \cdots x_1^{(d)} & x_1^{(j)} \\ \vdots & \vdots \\ \sigma_n x_n^{(1)} \cdots \hat{x}_n^{(j)} \cdots x_n^{(d)} & x_n^{(j)} \end{bmatrix}$$

where $\hat{x}_i^{(j)}$ denotes the omission of $x_i^{(j)}$ from the product.

We examine the structure of the singular vectors tuples of S by looking at the following three cases.

Case 1: Consider the variables $x_i^{(1)}, \ldots, x_i^{(d)}$, where $i \in \{1, 2, \ldots, n\}$ is fixed. Suppose that exactly one of the variables $x_i^{(j)}$ is 0, i.e. $x_i^{(k)} \neq 0$ for all $k \neq j$. The i-th row of the matrix $M_{S,j}$ has first entry $\sigma_i x_i^{(1)} \ldots \hat{x}_i^{(j)} \ldots x_i^{(d)} \neq 0$ and second entry $x_i^{(j)} = 0$. Therefore, in order for this matrix to have rank 1, we need the whole second column to be zero, i.e. $x_1^{(j)} = \cdots = x_n^{(j)} = 0$. Since $x^{(j)} \in \mathbb{P}^{n_j - 1}$, this can only happen if $n_j > n$ and one of the last $n_j - n$ coordinates of $x^{(j)}$ is nonzero. But the contraction $S(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots x^{(d)})$ lies in the span of $e_1^{(j)}, \ldots, e_n^{(j)}$, so in order for it to be parallel to $x^{(j)}$ it has to be 0. In particular, its i-th entry $\sigma_i x_i^{(1)} \ldots \hat{x}_i^{(j)} \ldots x_i^{(d)}$ has to be 0. Contradiction! Therefore, we can't have exactly one of the variables $x_i^{(1)}, \ldots, x_i^{(d)}$ equal to 0.

Case 2: Suppose that for some i at least two of the entries $x_i^{(1)}, \ldots, x_i^{(d)}$, but not all of them, are equal to 0. This means that the entry in the i-th row and the first column of $M_{S,k}$ is 0 for every k, and if $x_i^{(k)} \neq 0$ (and we assumed that one such k exists) then the entry in the i-th row and the second column is not 0. For such a k, the whole first column of $M_{S,k}$ must be 0 in order that it have rank 1. Therefore, for every i, at least two of the entries $x_i^{(1)}, \ldots, x_i^{(d)}$ are equal to 0. Conversely, if for every i at least two of the entries $x_i^{(1)}, \ldots, x_i^{(d)}$ are equal to 0 in such a way that $x^{(j)} \in \mathbb{P}^{n_j - 1}$, then, $(x^{(1)}, \ldots, x^{(d)})$ is a singular vector tuple of S. This gives the singular vector tuples of Type II, also known as the base points.

Case 3: It remains to consider the situation where, for every i, either $x_i^{(1)} = \ldots = x_i^{(d)} = 0$ or none of the variables $x_i^{(1)}, \ldots, x_i^{(d)}$ are 0. After reordering, assume that $x_i^{(1)} = \ldots = x_i^{(d)} = 0$ for $m + 1 \leq i \leq n$, for some $m \leq n$, and $x_i^{(1)}, \ldots, x_i^{(d)} \neq 0$ for $1 \leq i \leq m$.

The condition for being a singular vector tuple now yields the following system of polynomial equations in the Laurent polynomial ring $\mathbb{C}\left[x_i^{(j)}, \frac{1}{x_i^{(j)}} : 1 \leq i \leq m, 1 \leq j \leq d\right]$:

$$I = \left\langle \sigma_i x_i^{(1)} \dots \hat{x}_i^{(j)} \dots x_i^{(d)} x_l^{(j)} - \sigma_l x_l^{(1)} \dots \hat{x}_l^{(j)} \dots x_l^{(d)} x_i^{(j)} : 1 \leq j \leq d, 1 \leq i, l \leq m \right\rangle \quad (3.2.4)$$

To solve this system of equations, we use the theory of binomial ideal decomposition developed in [60].

Consider the lattice

$$L_\rho = \left\langle \sum_{k=1}^{d} (e_i^{(k)} - e_l^{(k)}) - 2(e_i^{(j)} - e_l^{(j)}) : 1 \leq j \leq d, 1 \leq i, l \leq m \right\rangle \subseteq \mathbb{Z}^{d \times m}$$

where $e_b^{(a)}$ is the elementary basis vector in $\mathbb{Z}^{d \times m}$ with a 1 in coordinate (a, b). Let $\rho : L_\rho \to \mathbb{C}^*$ denote the partial character

$$\rho\left(\sum_{k=1}^{d} (e_i^{(k)} - e_l^{(k)}) - 2(e_i^{(j)} - e_l^{(j)})\right) = \frac{\sigma_l}{\sigma_i} \quad \forall\ 1 \leq j \leq d, 1 \leq i, l, \leq m \quad (3.2.5)$$

Then the lattice ideal $I(\rho) = \langle x^v - \rho(v) : v \in L_\rho \rangle$ is our ideal I, where x^v denotes taking the variables $x_i^{(j)}$ in the ring to the powers indicated by the lattice element v.

We have the inclusion $L_\rho \subseteq L = \langle e_i^{(j)} - e_l^{(j)} : 1 \leq j \leq d, 1 \leq i, l \leq m \rangle$. Therefore by [60, Theorem 2.1],

$$I(\rho) = \bigcap_{\rho' \text{ extends } \rho \text{ to } L} I(\rho').$$

To decompose the ideal $I(\rho)$, we therefore seek to characterize the partial characters ρ' of L which extend ρ. Summing (3.2.5) over $1 \leq j \leq d$ gives the formula

$$\rho\left(\sum_{j=1}^{d}\left(\sum_{k=1}^{d} (e_i^{(k)} - e_l^{(k)}) - 2(e_i^{(j)} - e_l^{(j)})\right)\right) = \left(\frac{\sigma_l}{\sigma_i}\right)^d$$

which, after simplifying, yields $\rho\left((d-2)\sum_{k=1}^{d}(e_i^{(k)} - e_l^{(k)})\right) = \left(\frac{\sigma_l}{\sigma_i}\right)^d$. Therefore, any ρ' extending ρ satisfies

$$\rho'\left(\sum_{k=1}^{d} (e_i^{(k)} - e_l^{(k)})\right) = \phi_{il}\left(\frac{\sigma_l}{\sigma_i}\right)^{\frac{d}{d-2}} \quad (3.2.6)$$

where ϕ_{il} is a $(d-2)$-th root of unity. By rearranging (3.2.5), we moreover obtain

$$\rho'\left(2(e_i^{(j)} - e_l^{(j)})\right) = \rho'\left(\sum_{k=1}^{d} (e_i^{(k)} - e_l^{(k)})\right)\left(\frac{\sigma_i}{\sigma_l}\right). \quad (3.2.7)$$

Combining (3.2.6) and (3.2.7) yields

$$\rho'\left(2(e_i^{(j)} - e_l^{(j)})\right) = \phi_{il}\left(\frac{\sigma_l}{\sigma_i}\right)^{\frac{2}{d-2}}.$$

Thus,

$$\rho'\left(e_i^{(j)} - e_l^{(j)}\right) = \phi_{il}^{(j)}\left(\frac{\sigma_l}{\sigma_i}\right)^{\frac{1}{d-2}}$$

where $\phi_{il}^{(j)}$ are $2(d-2)$-th roots of unity such that $(\phi_{il}^{(j)})^2 = \phi_{il}$ for all $j = 1, \ldots, d$. It remains to find the relations satisfied by $\phi_{il}^{(j)}$ as i, l, j vary so that the original equation (3.2.5) is satisfied. When we plug in to that equation, we get

$$\prod_{k=1}^{d}\left(\phi_{il}^{(k)}\left(\frac{\sigma_l}{\sigma_i}\right)^{\frac{1}{d-2}}\right)\phi_{il}^{-1}\left(\frac{\sigma_l}{\sigma_i}\right)^{-\frac{2}{d-2}} = \frac{\sigma_l}{\sigma_i},$$

which is equivalent to

$$\prod_{k=1}^{d}\phi_{il}^{(k)} = \phi_{il}. \tag{3.2.8}$$

To satisfy these conditions, we can express everything in the following way. Let $\eta_{il} = \phi_{il}^{(j)}$ be a $(2d-4)$-th root of unity. Since $\eta_{il}^2 = (\phi_{il}^{(j)})^2 = \phi_{il}$, then $\phi_{il}^{(j)} = \eta_{il}\chi_{il}^{(j)}$, where $\chi_{il}^{(j)} = \pm 1$. Then, equation (3.2.8) becomes

$$\eta_{il}^d\prod_{j=2}^{d}\chi_{il}^{(j)} = \phi_{il} = \eta_{il}^2.$$

Equivalently,

$$\eta_{il}^{d-2} = \prod_{j=2}^{d}\chi_{il}^{(j)}.$$

Note that since η_{il} is a $(2d-4)$-th root of unity, we have $\eta_{il}^{d-2} = \pm 1$.

Finally, since $(e_i^{(j)} - e_l^{(j)}) + (e_l^{(j)} - e_h^{(j)}) + (e_h^{(j)} - e_i^{(j)}) = 0$, applying ρ gives

$$1 = \chi_{il}^{(j)}\eta_{il}\left(\frac{\sigma_l}{\sigma_i}\right)^{\frac{1}{d-2}}\chi_{lh}^{(j)}\eta_{lh}\left(\frac{\sigma_h}{\sigma_l}\right)^{\frac{1}{d-2}}\chi_{hi}^{(j)}\eta_{hi}\left(\frac{\sigma_i}{\sigma_h}\right)^{\frac{1}{d-2}}.$$

We now have all the relations required to find the ideals $I(\rho')$:

$$I(\rho') = \left\langle x_i^{(j)} - \chi_{i1}^{(j)}\eta_{i1}\left(\frac{\sigma_1}{\sigma_i}\right)^{\frac{1}{d-2}}x_1^{(j)} : 1 \le i \le m, 1 \le j \le d\right\rangle$$

where $\chi_{i1}^{(j)} \in \{\pm 1\}$ with $\prod_{j=2}^{d} \chi_{i1}^{(j)} = 1$ and η_{i1} are $(2d-4)$-th roots of unity. Setting $\chi_i^{(j)} = \chi_{i1}^{(j)}$ and $\eta_i = \eta_{i1}$, and taking I to be the intersection of the $I(\rho')$, we obtain the required form of our singular vector tuples:

$$I = \bigcap_{\eta,\chi} \left\langle x_i^{(j)} - \chi_i^{(j)} \eta_i \left(\frac{\sigma_1}{\sigma_i}\right)^{\frac{1}{d-2}} x_1^{(j)} : 1 \leq i \leq m, 1 \leq j \leq d \right\rangle.$$

Here $\chi_i^{(1)} = 1$ for all i. The zeros of this ideal are the singular vector tuples of Type 1, also known as the fixed points.

□

Now, we proceed to the main result of this section. We describe the singular vector tuples of a general odeco tensor.

Proposition 3.2.7. *Let $T = \sum_{i=1}^{n} \sigma_i v_i^{(1)} \otimes \cdots \otimes v_i^{(d)} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ be an odeco tensor such that $v_1^{(j)}, \ldots, v_n^{(j)} \in \mathbb{R}^{n_j}$ are orthonormal vectors. Let $V^{(j)} \in \mathbb{R}^{n_j} \otimes \mathbb{R}^{n_j}$ be any orthogonal matrix whose first n columns are $v_1^{(j)}, \ldots, v_n^{(j)}$. Then, the singular vector tuples of T are given by $(V^{(1)}x^{(1)}, \ldots, V^{(d)}x^{(d)})$ where $(x^{(1)}, \ldots, x^{d)})$ is a singular vector tuple of the diagonal tensor $S = \sum_{i=1}^{n} \sigma_i e_i^{(1)} \otimes \cdots \otimes e_i^{(d)}$ described in Lemma 3.2.5. In other words, the singular vectors of T are as follows:*

Type I: *Tuples $\left(V^{(1)}x^{(1)}, \ldots, V^{(d)}x^{(d)}\right)$, such that $(x^{(1)}, \ldots, x^{d)})$ is a Type I singular vector of the diagonal odeco tensor in Lemma 3.2.3.*

Type II: *Tuples $\left(V^{(1)}x^{(1)}, \ldots, V^{(d)}x^{(d)}\right)$, where the matrix $X = (x_i^{(j)})_{ij}$ has at least two zeros in each row such that none of the vectors $x^{(j)} \in \mathbb{P}^{n_j - 1}$ is identically zero.*

Proof. Assume that $(y^{(1)}, \ldots, y^{(d)})$ is a singular vector tuple of T. Equivalently, for all $1 \leq j \leq d$, the vector $T(y^{(1)}, \ldots, y^{(j-1)}, \cdot, y^{(j+1)}, \ldots, y^{(d)})$ is parallel to $y^{(j)}$. Unpacking the definition of the contraction, we obtain

$$T(y^{(1)}, \ldots, y^{(j-1)}, \cdot, y^{(j+1)}, \ldots, y^{(d)}) = \sum_{i=1}^{n} \sigma_i \left(\prod_{k \neq j} (v_i^{(k)} \cdot y^{(k)}) \right) v_i^{(j)} \tag{3.2.9}$$

The inner-product term $(v_i^{(k)} \cdot y^{(k)})$ is the i-th element in the vector $x^{(k)} := (V^{(k)})^T y^{(k)}$, where $V^{(k)}$ is any orthogonal matrix with first n columns equal to $v_1^{(k)}, \ldots, v_n^{(k)}$. We can re-write the right hand side of (3.2.9) in terms of the $x^{(k)}$, $1 \leq k \leq d$, as

$$\sum_{i=1}^{n} \sigma_i \left(\prod_{k \neq j} x_i^{(k)} \right) v_i^{(j)} = V^{(j)} \left(\sum_{i=1}^{n} \sigma_i \left(\prod_{k \neq j} x_i^{(k)} \right) e_i^{(j)} \right).$$

Therefore,

$$T(y^{(1)}, \ldots, y^{(j-1)}, \cdot, y^{(j+1)}, \ldots, y^{(d)}) = V^{(j)} S(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots, x^{(d)}),$$

where $S = \sum_{i=1}^{n} \sigma_i e_i^{(1)} \otimes \cdots \otimes e_i^{(d)}$. Since $V^{(j)}$ is orthogonal, $T(y^{(1)}, \ldots, y^{(j-1)}, \cdot, y^{(j+1)}, \ldots, y^{(d)})$ and $y^{(j)} = V^{(j)} x^{(j)}$ are parallel if and only if $S(x^{(1)}, \ldots, x^{(j-1)}, \cdot, x^{(j+1)}, \ldots, x^{(d)})$ and $x^{(j)}$ are parallel. Therefore, equivalently $(x^{(1)}, \ldots, x^{(d)})$ is a singular vector tuple of S, and the solutions for all such $(x^{(1)}, \ldots, x^{(d)})$ are given in Lemma 3.2.5. □

3.2.3 Proof of the Main Theorem

Proof of Theorem 3.2.4. The count for the contribution of the fixed points to the projective variety of singular vector tuples is obtained as follows directly from Proposition 3.2.7. For any choice of $m \in \{1, \ldots, n\}$, a subset of $\{1, \ldots, n\}$ of size m, scalars η_i which are $(2d-4)$-th roots of unity (where $i \in \{2, \ldots, m\}$), and $\chi_i^{(j)} \in \{\pm 1\}$ such that $\prod_{j=2}^{d} \chi_i^{(j)} = 1$ (where $i \in \{2, \ldots, m\}$ and $j \in \{2, \ldots, d\}$), we have one singular vector tuple. Therefore, the total number of singular vector tuples of Type I is

$$\sum_{m=1}^{n} \binom{n}{m} (2d-4)^{m-1} 2^{(m-1)(d-2)} = \sum_{m=1}^{n} (d-2)^{m-1} 2^{(m-1)(d-1)}$$

$$= \frac{\sum_{m=1}^{n} (d-2)^m 2^{(d-1)m}}{2^{d-1}(d-2)} = \frac{(2^{d-1}(d-2)+1)^n - 1}{2^{d-1}(d-2)}.$$

It remains to study the contribution made by the Type II singular vector tuples which constitute the base locus. By Proposition 3.2.7, we can restrict our attention to the tensor $S = \sum_{i=1}^{n} \sigma_i e_i \otimes \cdots \otimes e_i$, since its singular vector tuples differ from those of a general tensor only by an orthogonal change of coordinates in each factor.

We first study the case in which all dimensions are equal, $n_1 = \cdots = n_d = n$. Here, the tuple $(x^{(1)}, \ldots, x^{(d)})$ is a Type II singular vector tuple if and only if the matrix $X = (x_i^{(j)})$ has at least two zeros in every row and none of the vectors $x^{(j)}$ is identically zero. This configuration is a subvariety of $\mathbb{P}^{n-1} \times \cdots \times \mathbb{P}^{n-1}$. Its ideal is given by

$$\sum_{i=1}^{n} \langle x_i^{(1)} \cdots \hat{x}_i^{(j)} \cdots x_i^{(d)} : j = 1, \ldots, d \rangle \quad = \quad \sum_{i=1}^{n} \bigcap_{1 \le j < k \le d} \langle x_i^{(j)}, x_i^{(k)} \rangle. \tag{3.2.10}$$

We count the number of components in this subvariety by looking at the Chow ring of $\mathbb{P}^{n-1} \times \cdots \times \mathbb{P}^{n-1}$, which is $\mathbb{Z}[t_1, \ldots, t_d]/(t_1^n, \ldots, t_d^n)$. Each t_j represents the class of a hyperplane in $\mathbb{P}^{n_j - 1}$, the jth projective space in the product. The equivalence class of the variety $\mathcal{V}\left(\langle x_i^{(j)}, x_i^{(k)} \rangle\right)$ is given by $t_j t_k$. We consider the variety

$$\mathcal{V}\left(\bigcap_{1 \le j < k \le d} \langle x_i^{(j)}, x_i^{(k)} \rangle \right) = \bigcup_{1 \le j < k \le d} \mathcal{V}\left(\langle x_i^{(j)}, x_i^{(k)} \rangle \right) \tag{3.2.11}$$

which yields our variety of interest when we intersect over i. Its equivalence class is given by $\sum_{1 \le j < k \le d} t_j t_k$. From this, we see that the equivalence class in the Chow ring of the total

configuration is given by

$$p(t_1, \ldots, t_d) = \left(\sum_{1 \le j < k \le d} t_j t_k \right)^n . \tag{3.2.12}$$

Therefore, to count the number of linear spaces that constitute the Type II singular vector tuples, we wish to count the number of monomials of the polynomial (3.2.12) as an element of the Chow ring. Equivalently we count the terms in the expansion, as an element of $\mathbb{Z}[t_1, \ldots, t_d]$, that are not divisible by t_j^d for any j.

A monomial in the expanded form of (3.2.12) is produced by multiplying one of the $\binom{d}{2}$ terms in each of the n factors. This produces the first term, $\binom{d}{2}^n$, in the expression for the number of components in the base locus. We must now subtract those terms that are divisible by t_j^n for some fixed j. These are formed by selecting the terms $t_j t_{k_1}, \ldots, t_j t_{k_n}$ from consecutive factors. There are $d-1$ choices for each k_s, and d choices for the fixed j, yielding at first glance $d(d-1)^n$ terms of this format. However, we have double-counted those terms of the form $t_j^n t_k^n$ for fixed j and k, of which there are $\binom{d}{2}$. Combining these terms gives the correct specialization of our desired formula to the case $c = \#\{j : n_j = n\} = d$:

$$\binom{d}{2}^n - d(d-1)^n + \binom{d}{2} \tag{3.2.13}$$

The codimension of the ideal in (3.2.10) is $2n$, so our linear spaces enumerated above are of dimension $d(n-1) - 2n$.

The case of non-equal dimensions follows similarly: consider $S = \sum_{i=1}^n \sigma_i e_i^{(1)} \otimes \cdots \otimes e_i^{(d)}$ of format $n_1 \times \cdots \times n_d$ where $n = \min\{n_1, \ldots, n_d\}$ and $c = \#\{j : n_j = n\}$. To count the number of maximal-dimensional linear spaces, we consider the same polynomial (3.2.12) in the Chow ring $\mathbb{Z}[t_1, \ldots, t_d]/(t_1^{n_1}, \ldots, t_d^{n_d})$, and we now want to count the number of terms which are not divisible by $t_j^{n_j}$ for any $j = 1, \ldots, d$.

From the form of p in (3.2.12), we see that it is impossible for a term to be divisible by $t_j^{n_j}$ for any $n_j > n$. Our previous formula (3.2.13) therefore generalizes to

$$\binom{d}{2}^n - c(d-1)^n + \binom{c}{2}$$

and the dimension of each components is $\sum_{j=1}^d (n_j - 1) - 2n$. This concludes the proof. □

3.2.4 Further Explorations of the Type II Singular Vectors

In this section we turn our attention to the Type II singular vector tuples of the odeco tensor $S = \sum_{i=1}^n e_i^{(1)} \otimes \cdots \otimes e_i^{(d)}$, where S is of format $n_1 \times \cdots \times n_d$ and $n = \min\{n_1, \ldots, n_d\}$.

We can associate to each projective space $\mathbb{P}^{n_j - 1}$ the simplex $\Delta_{n_j - 1}$ and consider our linear spaces as polyhedral subcomplexes (prodsimplicial complexes) in the boundary of the

product of simplices $\Delta_{n_1-1} \times \ldots \times \Delta_{n_d-1}$. The number of components in the variety of Type II singular vector tuples is the number of facets in this complex.

We first return to Example 3.2.6, in which we had six Type I singular vector tuples, and the Type II singular vector tuples made up five copies of $\mathbb{P}^1$. In Figure 3.4, we draw the polyhedral complex in $\Delta_1 \times \Delta_2 \times \Delta_2$ corresponding to the Type II singular vector tuples. Motivated by this example, we investigate the shape of the Type II singular vector tuples of other small odeco tensors.

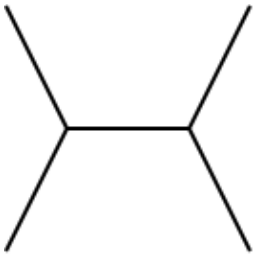


Figure 3.4: The Type II singular vectors tuples of a $2 \times 3 \times 3$ odeco tensor, drawn as a polyhedral complex

It is interesting to stratify odeco tensors according to the dimension of their Type II singular vectors, using the following proposition:

Proposition 3.2.8. *For each dimension k, the odeco tensors whose Type II singular vector tuples have dimension k come from a finite list of possible sizes $n_1 \times \cdots \times n_d$.*

Proof. By Theorem 3.2.4, we seek the solutions of $n_1, \ldots, n_d$ with $n_j \geq 2$ and $d \geq 3$ to the equation

$$\sum_{j=1}^{d}(n_j - 1) - 2n = k \tag{3.2.14}$$

where $n = \min\{n_1, \ldots, n_d\}$. An odeco tensor of size $n_1 \times \cdots \times n_d$ will then have Type II singular vector tuples consisting of product of linear spaces of dimension k. Without loss of generality, we assume that $n_1 \leq \ldots \leq n_d$, and hence $n = n_1$. Let the constant α be such that $n_2 = n + \alpha$. For fixed α, rearranging (3.2.14) shows that we seek to solve the equation

$$\sum_{j=3}^{d}(n_j - 1) = k + 2 - \alpha. \tag{3.2.15}$$

This has finitely many solutions, since the right hand side is a fixed number, and each summand on the left hand side has strictly positive integer size. From the form of the right hand side, we see that there will be solutions for only finitely many values of α. In conclusion, there are only finitely many size combinations $n_1 \times \cdots \times n_d$ which yield Type II singular vector tuples of dimension k. □

For example, odeco tensors whose Type II singular vector tuples constitute a zero-dimensional projective variety have possible sizes:

$$\{2 \times 2 \times 2, 3 \times 3 \times 3, 2 \times 2 \times 2 \times 2\}.$$

Theorem 3.2.4 tells us how many singular vector tuples there are of Types I and II, which are entered in the first two columns of the table below. The number of singular vector tuples of a generic tensor of a given format is given by [69, Theorem 1], and this is entered into the last column of the table. We observe that odeco tensors whose Type II singular vector tuples consist solely of points attain the generic count.

Tensor Size	Type I Count	Type II Count	Generic Count
$2 \times 2 \times 2$	6	0	6
$3 \times 3 \times 3$	31	6	37
$2 \times 2 \times 2 \times 2$	18	6	24

Now we consider odeco tensors whose Type II singular vector tuples make a one-dimensional projective variety. They are of one of the following formats:

$$\{2 \times 3 \times 3, 2 \times 2 \times 4, 3 \times 3 \times 4, 4 \times 4 \times 4, 2 \times 2 \times 2 \times 3, 2 \times 2 \times 2 \times 2 \times 2\}.$$

Their singular vector tuples consists of a finite collection of points (Type I) and a collection of copies of $\mathbb{P}^1$ in the product of projective spaces $\mathbb{P}^{n_1-1} \times \ldots \times \mathbb{P}^{n_d-1}$ (Type II). When two copies of $\mathbb{P}^1$ meet, they do so at a triple intersection point. The data for these tensor formats is recorded in the table below. Under a small perturbation, each copy of $\mathbb{P}^1$ contributes one singular vector tuple, and two arise from each triple intersection. We observe that summing the Type I count, the number of copies of $\mathbb{P}^1$, and twice the number of triple intersections yields the generic count.

Tensor Size	Type I Count	$\#\mathbb{P}^1$s	#Triple Intersections	Generic Count
$2 \times 3 \times 3$	6	5	2	15
$2 \times 2 \times 4$	6	2	0	8
$3 \times 3 \times 4$	31	12	6	55
$4 \times 4 \times 4$	156	36	24	240
$2 \times 2 \times 2 \times 3$	18	12	6	42
$2 \times 2 \times 2 \times 2 \times 2$	50	30	20	120

We explored the $2 \times 3 \times 3$ case in more detail in Example 3.2.6. In the $3 \times 3 \times 4$ and $2 \times 2 \times 2 \times 3$ cases the simplicial complexes of the Type II singular vector tuples are the same shape. They consist of the 12 copies of $\mathbb{P}^1$ meeting at six triple intersections pictured in Figure 3.5.

In the case of $2 \times 2 \times 2 \times 2 \times 2$ odeco tensors, we have 30 copies of $\mathbb{P}^1$ that meet at 20 triple intersection points as seen in the non-planar arrangement pictured in Figure 3.6. In

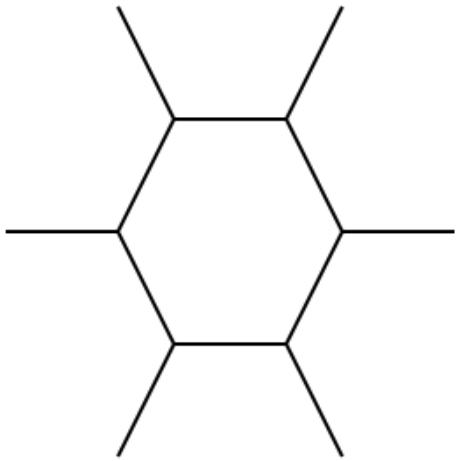


Figure 3.5: The 12 copies of $\mathbb{P}^1$ with six triple intersection points, for $3 \times 3 \times 4$ tensors and $2 \times 2 \times 2 \times 3$ tensors

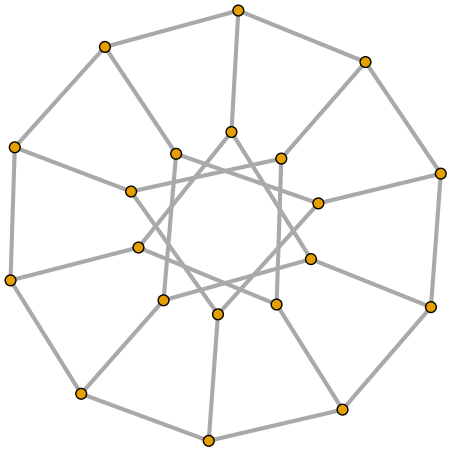


Figure 3.6: The 30 copies of $\mathbb{P}^1$ with 20 triple intersection points, for $2 \times 2 \times 2 \times 2 \times 2$ tensors

the case of $4 \times 4 \times 4$ odeco tensors, we have 36 copies of $\mathbb{P}^1$ meeting at 24 triple intersection points as pictured in Figure 3.7.

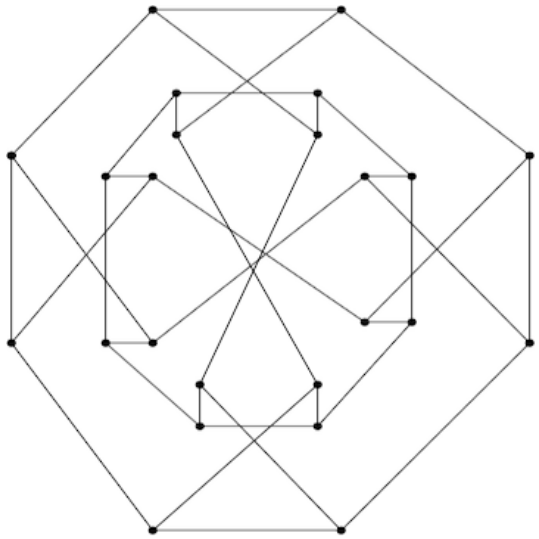


Figure 3.7: The 36 copies of $\mathbb{P}^1$ with 24 triple intersection points, for $4 \times 4 \times 4$ tensors

3.3 Conclusion

In this chapter we studied orthogonally decomposable tensors. In particular, we described the eigenvectors of symmetric odeco tensors, and the singular vector tuples of non-symmetric odeco tensors. For a general tensor it is hard both computationally and algebraically to find its eigenvectors or singular vector tuples. As we saw, odeco tensors have very appealing structure, and we can express their eigenvectors and singular vector tuples in terms of the elements in their decomposition.

In the next chapter we continue to explore the properties of odeco tensors by describing the *odeco variety*.

Chapter 4

Varieties of Tensors

4.1 The Variety of Orthogonally Decomposable Tensors

While every matrix admits a singular value decomposition, in which the terms are pairwise orthogonal, higher-order tensors typically do not admit such an orthogonal decomposition. Those that do have attracted attention from theoretical computer science and scientific computing. We complement this existing body of literature with an algebro-geometric analysis of the set of orthogonally decomposable tensors.

More specifically, we prove that they form a real-algebraic variety defined by polynomials of degree at most four. The exact degrees, and the corresponding polynomials, are different in each of the two scenarios: ordinary or symmetric. A key feature of our approach is a surprising connection between orthogonally decomposable tensors and semisimple associative algebras.

This section is based on part of joint work with Ada Boralevi, Jan Draisma and Emil Horobeţ titled *Orthogonal and unitary tensor decomposition from an algebraic perspective* [23].

4.1.1 Introduction and results

By the singular value decomposition, any real $m \times n$-matrix A can be written as $A = \sum_{i=1}^{k} u_i v_i^T$, where $u_1, \ldots, u_k \in \mathbb{R}^m$ and $v_1, \ldots, v_k \in \mathbb{R}^n$ are sets of nonzero, pairwise orthogonal vectors. The singular values $||u_i|| \cdot ||v_i||$, including their multiplicities, are uniquely determined by A, and if these are all distinct, then so are the terms $u_i v_i^T$. If $m = n$ and A is symmetric, then the u_i and v_i can be chosen equal.

In this section we consider *higher-order* tensors in a tensor product $V^{(1)} \otimes \cdots \otimes V^{(d)}$ of finite-dimensional vector spaces $V^{(i)}$ over $\mathbb{R}$ where the tensor product is also over $\mathbb{R}$. We assume that each $V^{(i)}$ is equipped with a positive-definite inner product $(\cdot|\cdot)$.

Definition 4.1.1. A tensor in $V^{(1)} \otimes \cdots \otimes V^{(d)}$ is called *orthogonally decomposable* (*odeco*) if it can be written as

$$\sum_{i=1}^{k} v_i^{(1)} \otimes \cdots \otimes v_i^{(d)},$$

where for each j the vectors $v_1^{(j)}, \ldots, v_k^{(j)}$ are nonzero and pairwise orthogonal in $V^{(j)}$.

Note that orthogonality implies that the number k of terms is at most the minimum of the dimensions of the $V^{(i)}$, so odeco tensors form a rather low-dimensional subvariety of the space of all tensors; see Proposition 4.1.7.

Next we consider tensor powers of a single, finite-dimensional $\mathbb{R}$-space V. We write $S^d(V)$ for the subspace of $V^{\otimes d}$ consisting of all *symmetric tensors*, i.e., those fixed by all permutations of the tensor factors.

Definition 4.1.2. A tensor in $S^d(V)$ is called *symmetrically odeco* if it can be written as

$$\sum_{i=1}^{k} \pm v_i^{\otimes d}$$

where the vectors $v_1, \ldots, v_k$ are nonzero, pairwise orthogonal vectors in V.

The signs are only required if d is even, as they can otherwise be absorbed into the v_i by taking a d-th root of -1. Clearly, a symmetrically odeco tensor is symmetric and odeco in the earlier sense. The converse also holds; see Proposition 4.1.16.

By quantifier elimination, it follows that the set of odeco tensors is a semi-algebraic set in $V^{(1)} \otimes \cdots \otimes V^{(d)}$, i.e., a finite union of subsets described by polynomial equations and (weak or strict) polynomial inequalities. A simple compactness argument (see Proposition 4.1.5) also shows that they form a closed subset in the Euclidean topology, so that only weak inequalities are needed. However, our main result says that, in fact, only *equations* are needed, and that the same holds in the symmetrically case as well.

Theorem 4.1.3 (Main Theorem)**.** *For each integer $d \geq 3$, and for all finite-dimensional inner product spaces $V^{(1)}, \ldots, V^{(d)}$ and V over $\mathbb{R}$, the odeco tensors in $V^{(1)} \otimes \cdots \otimes V^{(d)}$, and the symmetrically odeco tensors in $S^d(V)$, form real algebraic varieties defined by polynomials of degree 2.*

Remark 4.1.4. Several remarks are in order:

1. Unlike for $d = 2$, for $d \geq 3$ the decomposition in Definitions 4.1.1, 4.1.2 is always unique in the sense that the terms are uniquely determined, regardless of whether some of their norms coincide; see Proposition 4.1.6.

2. The polynomial equations defining the variety of symmetric odeco tensors are the same as the ones we saw in Conjecture 3.1.16. We will describe the polynomials defining the

variety of ordinary odeco tensors in detail later on. The high-level perspective in both cases is that the equations of degree two guarantee that a particular algebra associated to a tensor is associative.

3. The degree 2 is minimal in the sense that there are no linear equations.

4. More generally, we do not know whether the equations that we give generate the prime ideal of all polynomial equations vanishing on our real algebraic varieties when $d \geq 3$.

The remainder of this section is organized as follows. In Subsection 4.1.2 we discuss some background and earlier literature.

In Subsection 4.1.3 we prove the Main Theorem for tensors of order three. The proofs for symmetrically odeco three-tensors are the simplest, and those for ordinary odeco three-tensors build upon them. Then, in Subsection 4.1.4 we derive the theorem for higher-order ordinary and symmetric tensors. We conclude in Subsection 4.1.5 with some open questions.

4.1.2 Background

In this section we collect background results on orthogonally decomposable tensors, and connect our results to earlier work on them.

Proposition 4.1.5. *The set of (ordinary or symmetrically) odeco tensors is closed in the Euclidean topology.*

Proof. We give the argument for symmetrically odeco tensors; the same works in the other case. Thus consider the space $V = \mathbb{R}^n$ with the standard inner product, let O_n be the orthogonal group, and consider the map

$$\varphi : \mathrm{O}_n \times \mathbb{P}V \to \mathbb{P}S^d(V),\ ((v_1|\ldots|v_n), [\lambda_1 : \ldots : \lambda_n]) \mapsto \left[\sum_{i=1}^{n} \lambda_i v_i^{\otimes d}\right].$$

Here $\mathbb{P}$ stands for projective space and where v_i is the i-th column of the orthogonal matrix v. The key point is that this map is well-defined and continuous, since the expression between the last square brackets is never zero by linear independence of the $v_i^{\otimes d}$. Now φ is a continuous map whose source is a compact topological space, hence $\operatorname{im}\varphi$ is a closed subset of $\mathbb{P}S^d(V)$. But then the pre-image of $\operatorname{im}\varphi$ in $S^d(V) \setminus \{0\}$ is also closed, and so is the union of this pre-image with $\{0\}$. This is the set of symmetrically odeco tensors in $S^d(V)$. □

Proposition 4.1.6. *For $d \geq 3$, any (ordinary or symmetrically) odeco tensor has a* unique *orthogonal decomposition.*

In the ordinary case this was proved in [161, Theorem 3.2].

Proof. We give the argument for ordinary odeco tensors. Consider an orthogonal decomposition

$$T = \sum_{i=1}^{k} v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}$$

of an odeco tensor $T \in V^{(1)} \otimes \cdots \otimes V^{(d)}$. Contracting T with an arbitrary tensor $S \in V^{(3)} \otimes \cdots \otimes V^{(d)}$ via the inner products on $V^{(3)}, \ldots, V^{(d)}$ leads to a tensor

$$T' = \sum_{i=1}^{k} \lambda_i v_i^{(1)} \otimes v_i^{(2)}$$

where λ_i is the inner product of S with $v_i^{(3)} \otimes \cdots \otimes v_i^{(d)}$. Now the above is a singular value decomposition for the two-tensor T', of which, for S sufficiently general, the singular values $|\lambda_i| \cdot ||v_i^{(1)}|| \cdot ||v_i^{(2)}||$ are all distinct. Thus $v_1^{(1)}, \ldots, v_k^{(1)}$ are, up to nonzero scalars, uniqely determined as the singular vectors (corresponding to the nonzero singular values) of the pairing of T with a sufficiently general S. And these vectors determine the corresponding terms, since the i-th term equals $v_i^{(1)}$ tensor the pairing of T with $v_i^{(1)}$, divided by $||v_i^{(1)}||^2$.

The arguments in the symmetric case are almost identical. We stress that, as permuting the first two factors commutes with contracting the last $d-2$ factors, the contraction of a symmetric tensor is a symmetric matrix. □

Note that the proof of this proposition yields a simple randomized algorithm for deciding whether a tensor is odeco, and for finding a decomposition when it exists. At the heart of this algorithm is the computation of an ordinary singular-value decomposition for a small matrix. For much more on algorithmic issues see [16, 100, 135, 161].

The uniqueness of the orthogonal decomposition makes it easy to compute the dimensions of the real-algebraic varieties in our Main Theorem.

Proposition 4.1.7. *Let $n := \dim V$, $l := \lfloor \frac{n}{d} \rfloor$, and assume that the dimensions $n_i := \dim V^{(i)}$ are in increasing order $n_1 \leq \ldots \leq n_d$. Then, the dimensions of the real-algebraic varieties of symmetric odeco tensors is*

$$n + \binom{n}{2},$$

and that of ordinary odeco tensors is

$$n_1 + \sum_{j=1}^{d} \frac{n_1(2n_j - n_1 - 1)}{2}.$$

Proof. In the symmetric case, a symmetrically odeco tensor encodes n pairwise perpendicular points in $\mathbb{P}V$. For the first point we have $n-1$ degrees of freedom. The second point is chosen from the projective space orthogonal to the first point, so this yields $n-2$ degrees

of freedom, etc. Summing up, we obtain $\binom{n}{2}$ degrees of freedom over K for the points. In addition, we have n scalars from $\mathbb{R}$ for the individual terms. Since each odeco tensor has a unique decomposition, the dimension of the odeco variety is the same as the dimension of the space of n pairwise orthogonal points and n scalars.

The computation for the ordinary case is the same, except that only n_1 pairwise perpendicular projective points are chosen from each $V^{(j)}$. □

Over the last two decades, orthogonal tensor decomposition has been studied intensively from a scientific computing perspective (see, e.g., [44, 99, 98, 42, 100]). The paper [42] gives a characterization of orthogonally decomposable tensors in terms of their *higher-order SVD* [48], which is different from the real-algebraic characterization in our Main Theorem. One of the interesting properties of an orthogonal tensor decomposition with k terms is that discarding the r terms with smallest norm yields the best rank-r approximation to the tensor; see [154], where it is also proved that in general, tensors are not *optimally truncatable* in this manner.

In general, tensor decomposition is NP-hard [92]. The decomposition of odeco tensors, however, can be found efficiently. The vectors in the decomposition of an odeco tensor are exactly the attraction points of the *tensor power method* and are called *robust* eigenvectors. Because of their efficient decomposition, odeco tensors have been used in machine learning, in particular for learning latent variables in statistical models [8]. More recent work in this direction concerns overcomplete latent variable models [7].

In Conjecture 3.1.16, we presented the equations defining the variety of symmetrically odeco tensors. Formulated for the case of ordinary tensors instead, this conjecture is as follows. Let $V^{(1)}, \ldots, V^{(d)}$ be real inner product spaces, and consider an odeco tensor $T \in V^{(1)} \otimes \cdots \otimes V^{(d)}$ with orthogonal decomposition $T = \sum_{i=1}^{k} v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}$. Now take two copies of T, and contract these in their l-th components via the inner product $V^{(l)} \times V^{(l)} \to \mathbb{R}$. By orthogonality of the $v_i^{(l)}$, $i = 1, \ldots, k$, after regrouping the tensor factors, the resulting tensor is

$$\sum_{i=1}^{k} \left(||v_i^{(l)}||^2 \bigotimes_{j \neq l} (v_i^{(j)} \otimes v_i^{(j)}) \right) \in \bigotimes_{j \neq l} (V^{(j)} \otimes V^{(j)});$$

we write $T *_l T$ for this tensor. It is clear from this expression that $T *_l T$ is multi-symmetric in the sense that it lies in the subspace $\bigotimes_{j \neq l} S^2(V^{(j)})$. In [131] I conjecture that this (or rather, its analogue in the symmetric setting) characterizes odeco tensors. This is now a theorem, which follows from the proof of our main theorem (see Remark 4.1.14).

Theorem 4.1.8. *$T \in V^{(1)} \otimes \cdots \otimes V^{(d)}$ is odeco if and only if for all $l = 1, \ldots, d$ we have*

$$T *_l T \in \bigotimes_{j \neq l} S^2(V^{(j)}).$$

This concludes the discussion of background to our results. We now proceed to prove the main theorem in the case of order-three tensors.

4.1.3 Tensors of order three

In all our proofs below, we will encounter a finite-dimensional vector space A over $\mathbb{R}$ equipped with a positive-definite inner product $(\cdot|\cdot)$, as well as a bi-additive product $A \times A \to A,\ (x, y) \mapsto x \cdot y$ which is bilinear. The product will be commutative. Moreover, the inner product will be *compatible* with the product in the sense that $(x \cdot y|z) = (z \cdot x|y)$. An *ideal* in $(A, \cdot)$ is a $\mathbb{R}$-subspace I such that $I \cdot A \subseteq I$—by commutativity we then also have $A \cdot I \subseteq I$—and A is called *simple* if $A \neq \{0\}$ and A contains no nonzero proper ideals. We have the following well-known result.

Lemma 4.1.9. *The orthogonal complement $I^\perp$ of any ideal I in A is an ideal, as well. Consequently, A splits as a direct sum of pairwise orthogonal simple ideals.*

Proof. We have $(A \cdot I^\perp|I) = (I \cdot A|I^\perp) = \{0\}$. The second statement follows by induction on $\dim A$. Therefore, $A \cdot I^\perp \subseteq I^\perp$, and since $I^\perp$ is a subspace of A, it follows by definition that $I^\perp$ is an ideal of A. □

4.1.3.1 Symmetrically odeco three-tensors

In this subsection, we fix a finite-dimensional real inner product space V and characterize odeco tensors in $S^3(V)$. We have $S^3(V) \subseteq V^{\otimes 3} \cong (V^*)^{\otimes 2} \otimes V$, where the isomorphism comes from the linear isomorphism $V \to V^*,\ v \mapsto (v|\cdot)$. Thus a general tensor $T \in S^3(V)$ gives rise to a bilinear map $V \times V \to V,\ (u, v) \mapsto u \cdot v$, which has the following properties:

1. $u \cdot v = v \cdot u$ for all $u, v \in V$ (commutativity, which follows from the fact that T is invariant under permuting the first two factors); and
2. $(u \cdot v|w) = (u \cdot w|v)$ (compatibility with the inner product, which follows from the fact that T is invariant under permuting the last two factors).

Thus T gives V the structure of an $\mathbb{R}$-algebra equipped with a compatible inner product. The following lemma describes the quadratic equations from the Main Theorem.

Lemma 4.1.10. *If T is symmetrically odeco, then $(V, \cdot)$ is associative.*

Proof. Write $T = \sum_{i=1}^k v_i^{\otimes 3}$ where $v_1, \ldots, v_k$ are pairwise orthogonal nonzero vectors. Then we find, for $x, y, z \in V$, that

$$x \cdot (y \cdot z) = x \cdot \left(\sum_i (v_i|y)(v_i|z)v_i \right) = \sum_i (v_i|x)(v_i|y)(v_i|z)(v_i|v_i) = (x \cdot y) \cdot z,$$

where we have used that $(v_i|v_j) = 0$ for $i \neq j$ in the second equality. □

Proposition 4.1.11. *Conversely, if $(V, \cdot)$ is associative, then T is symmetrically odeco.*

Proof. By Lemma 4.1.9, V has an orthogonal decomposition $V = \bigoplus_i U_i$ where the subspaces U_i are (nonzero) simple ideals of V. Correspondingly, T decomposes as an element of $\bigoplus_i S^3(U_i)$. Thus it suffices to prove that each U_i is one-dimensional. This is certainly the case when the multiplication $U_i \times U_i \to U_i$ is zero, because then any one-dimensional subspace of U_i is an ideal in V, hence equal to U_i by simplicity. If the multiplication map is nonzero, then pick an element $x \in U_i$ such that the multiplication $M_x : U_i \to U_i,\ y \mapsto x \cdot y$ is nonzero. Then $\ker M_x$ is an ideal in V, because for $z \in V$ we have

$$x \cdot (\ker M_x \cdot z) = (x \cdot \ker M_x) \cdot z = \{0\},$$

where we use associativity. By simplicity of U_i, $\ker M_x = \{0\}$. Now define a new bilinear multiplication $*$ on U_i via $y * z := M_x^{-1}(y \cdot z)$. This multiplication is commutative, has x as a unit element, and we claim that it is also associative. Indeed,

$$((x \cdot y) * z) * (x \cdot v) = M_x^{-1}(M_x^{-1}((x \cdot y) \cdot z) \cdot (x \cdot v)) = y \cdot z \cdot v = (x \cdot y) * (z * (x \cdot v)),$$

where we used associativity and commutativity of $\cdot$ in the second equality. Since any element is a multiple of x, this proves associativity. Moreover, $(U_i, *)$ is simple; indeed, if I is ideal, then $M_x^{-1}(U_i \cdot I) \subseteq I$ and hence

$$U_i \cdot (x \cdot I) = (U_i \cdot x) \cdot I = U_i \cdot I \subseteq x \cdot I,$$

so that $x \cdot I$ is an ideal in $(U_i, \cdot)$; and therefore $I = \{0\}$ or $I = U_i$.

Now $(U_i, *)$ is a simple, associative $\mathbb{R}$-algebra with 1, hence isomorphic to a matrix algebra over a division ring. As it is also commutative, it is isomorphic to either $\mathbb{R}$ or $\mathbb{C}$. If it were isomorphic to $\mathbb{C}$, then it would contain a square root of -1, i.e., an element y with $y*y = -x$, so that $y \cdot y = -x \cdot x$. But then

$$0 < (x \cdot y | x \cdot y) = (y \cdot y | x \cdot x) = -(x \cdot x | x \cdot x) < 0,$$

a contradiction. We conclude that U_i is one-dimensional, as desired. □

Lemma 4.1.10 and Proposition 4.1.11 imply the Main Theorem for symmetrically odeco three-tensors, because the identity $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ expressing associativity translates into quadratic equations for the tensor T.

Example 4.1.1. *We now show how we can obtain the equations from Conjecture 3.1.16 using the statement of Proposition 4.1.11. Let $T \in S^3(\mathbb{R}^n)$ be a symmetric $n \times n \times n$ tensor. The algebra $(V, \cdot)$ associated to T is associative if and only if $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for all $x, y, z \in V$. We claim that it is enough to consider standard basis vectors $x = e_i, y = e_j, z = e_k$. Then,*

$$[(e_i \cdot e_j) \cdot e_k]_l = [T_{i,j,\cdot} \cdot e_k]_l = \sum_{s,t=1}^{n} T_{s,t,l} T_{i,j,s} e_{k,t} = \sum_{s=1}^{n} T_{s,k,l} T_{i,j,s} = \sum_{s=1}^{n} T_{k,l,s} T_{i,j,s},$$

and

$$[e_i \cdot (e_j \cdot e_k)]_l = [e_i \cdot T_{j,k,\cdot}]_l = \sum_{t,s=1}^{n} T_{t,s,l} e_{i,t} T_{j,k,s} = \sum_{s=1}^{n} T_{i,s,l} T_{j,k,s} = \sum_{s=1}^{n} T_{i,l,s} T_{j,k,s}.$$

Therefore,

$$\sum_{s=1}^{n} T_{k,l,s} T_{i,j,s} = \sum_{s=1}^{n} T_{i,l,s} T_{j,k,s},$$

for every i, j, k, l, which is the same condition we discovered in equation (3.1.6)*. Moreover, note that the left hand side and the right hand side can be rewritten as*

$$(T *_3 T)_{k,l,i,j} = (T *_3 T)_{i,l,j,k}.$$

Since T is symmetric, we get exactly the condition that $T *_3 T \in S^4(\mathbb{R}^n)$.

4.1.3.2 Ordinary odeco three-tensors

In this subsection, we consider a general tensor T in a tensor product $U \otimes V \otimes W$ of real, finite-dimensional inner product spaces. Via the inner products, T gives rise to a bilinear map $U \times V \to W$, and similarly with the three spaces permuted. Consider the external direct sum $A := U \oplus V \oplus W$ of U, V, W, and equip A with the inner product $(\cdot|\cdot)$ that restricts to the given inner products on U, V, W and that makes these spaces pairwise perpendicular.

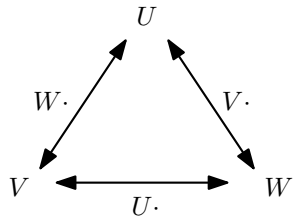


Figure 4.1: $U \cdot (V + W) = W + V$, and similarly with U, V, W permuted.

Taking cue from the symmetric case, we construct a bilinear product $\cdot : A \times A \to A$ as follows: the product in A of two elements in U, or two elements in V, or in W, is defined as zero; $\cdot$ restricted to $U \times V$ is the map into W given by T; etc.—see Figure 4.1. The tensor in $S^3(A)$ describing the multiplication is the symmetric embedding of T from [127].

As in the symmetrically odeco case, the algebra has two fundamental properties:

1. it is commutative: $x \cdot y = y \cdot x$ by definition; and
2. the inner product is compatible: $(x \cdot y|z) = (x \cdot z|y)$. For instance, if $x \in U, y \in V, z \in W$, then both sides equal the inner product of the tensor $x \otimes y \otimes z$ with T; and if $y, z \in W$, then both sides are zero both for $x \in U$ (so that $x \cdot y, x \cdot z \in V$, which is perpendicular to W) and for $x \in W$ (so that $x \cdot y = x \cdot z = 0$) and for $x \in V$ (so that $x \cdot y, x \cdot z \in U \perp W$).

We are now interested in *homogeneous* ideals $I \subseteq A$ only, i.e., ideals such that $I = (I \cap U) \oplus (I \cap V) \oplus (I \cap W)$. We call A *simple* if it is nonzero and does not contain proper, nonzero homogeneous ideals. We will call an element of A homogeneous if it belongs to one of U, V, W. Next, we derive a polynomial identity for odeco tensors.

Lemma 4.1.12. *If T is odeco, then for all homogeneous x, y, z where x and z belong to the same space (U, V, or W), we have $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.*

We will refer to this property as *partial associativity*.

Proof. If x, y, z all belong to the same space, then both products are zero. Otherwise, by symmetry, it suffices to check the case where $x, z \in U$ and $y \in V$. Let $T = \sum_i u_i \otimes v_i \otimes w_i$ be an orthogonal decomposition of T. Then we have

$$(x \cdot y) \cdot z = \left(\sum_i (u_i|x)(v_i|y) w_i \right) \cdot z = \sum_i (u_i|x)(v_i|y)(w_i|w_i)(z|u_i) = x \cdot (y \cdot z),$$

where we have used that $(w_i|w_j) = 0$ for $i \neq j$ in the second equality. □

Proposition 4.1.13. *Conversely, if $(A, \cdot)$ is partially associative, then T is odeco.*

Proof. By a version of Lemma 4.1.9 restricted to homogeneous ideals, A is the direct sum of pairwise orthogonal, simple homogeneous ideals I_i. Accordingly, T lies in $\bigoplus_i (I_i \cap U) \otimes (I_i \cap V) \otimes (I_i \cap W)$. Thus it suffices to prove that T is odeco under the additional assumption that A itself is simple and that $\cdot$ is not identically zero.

By symmetry, we may assume that $V \cdot (U+W) \neq \{0\}$. For $u \in U$, let $M_u : V+W \to W+V$ be multiplication with u. By commutativity and partial associativity, the M_u, for $u \in U$, all commute. By compatibility of $(\cdot|\cdot)$, each M_u is symmetric with respect to the inner product on $V + W$, and hence orthogonally diagonalizable.

Consequently, $V+W$ splits as a direct sum of pairwise orthogonal simultaneous eigenspaces

$$(V + W)_\lambda := \{v + w \in V + W \mid u \cdot (v + w) = \lambda(u)(w + v) \text{ for all } u \in U\},$$

where λ runs over U^*. Suppose we are given $v + w \in (V + W)_\lambda$ and $v' + w' \in (V + W)_\mu$ with $\lambda \neq \mu$. Then $v + w$ and $v' + w'$ are perpendicular and for each $u \in V$ we have

$$(u|(v + w) \cdot (v' + w')) = (u \cdot (v + w)|v' + w') = \lambda(u)(v + w|v' + w') = 0,$$

hence $(v + w) \cdot (v' + w') = 0$. We conclude that for each λ the space

$$(V + W)_\lambda \oplus [(V + W)_\lambda \cdot (V + W)_\lambda]$$

is a homogeneous ideal in A. By simplicity and the fact that $M_u \neq 0$ for at least some u, A is equal to this ideal for some nonzero λ. Pick an $x \in U$ such that $\lambda(x) = 1$, so that $x \cdot (v + w) = w + v$ for all $v \in V$, $w \in W$. In particular, for $v, v' \in V$ we have

$(M_x v | M_x v') = (M_x^2 v | v') = (v|v')$, so that the restrictions $M_x : V \to W$ and $M_x : W \to V$ are mutually inverse isometries.

By the same construction, we find an element $z \in W$ such that $z \cdot (u + v) = v + u$ for all $u \in U,\ v \in V$. Let T' be the image of T under the linear map $M_x \otimes I_V \otimes M_z : U \otimes V \otimes W \to V \otimes V \otimes V$. We claim that T' is symmetrically odeco. Indeed, let $* : V \times V \to V$ denote the bilinear map associated to T'. We verify the conditions from Section 4.1.3.1. First,

$$v * v' = (x \cdot v) \cdot (z \cdot v') = z \cdot ((x \cdot v) \cdot v') = z \cdot ((v' \cdot x) \cdot v) = (z \cdot v) \cdot (v' \cdot x) = v' * v,$$

where we have repeatedly used commutativity and partial associativity (e.g., in the second equality, to the elements $x \cdot v, z$ belonging to the same space W). Second, we have

$$(v * v' | v'') = ((x \cdot v) \cdot (z \cdot v') | v'') = ((x \cdot v) | v' \cdot (z \cdot v'')) = (v | (x \cdot v') \cdot (z \cdot v'')) = (v | v' * v'').$$

Hence T' is, indeed, and element of $S^3(V)$. Finally, we have

$$\begin{aligned}(v * v') * v'' &= (x \cdot ((x \cdot v) \cdot (z \cdot v'))) \cdot (z \cdot v'') = x \cdot ((z \cdot v'') \cdot ((x \cdot v) \cdot (z \cdot v'))) \\ &= x \cdot (((z \cdot v'') \cdot (x \cdot v)) \cdot (z \cdot v')) = x \cdot ((v * v'') \cdot (z \cdot v')) = (v * v'') * v',\end{aligned}$$

which, together with commutativity, implies associativity of $*$. Hence T' is (symmetrically) odeco by Proposition 4.1.13, and hence so is its image T under the tensor product $M_x \otimes I_V \otimes M_z$ of linear isometries. □

Remark 4.1.14. The condition that $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for, say, $x, z \in W$ and $y \in V$ translates into the condition that the contraction $T *_1 T \in (V \otimes V) \otimes (W \otimes W)$ lies in $S^2(V) \otimes S^2(W)$. This can be seen by proceeding analogously to Example 4.1.1. Thus Proposition 4.1.13 implies Theorem 4.1.8 in the case of three factors. The case of more factors follows from the case of three factors and flattening as in Proposition 4.1.15.

4.1.4 Higher-order tensors

In this section, building on the case of order three, we prove the Main Theorem for tensors of arbitrary order.

4.1.4.1 Ordinary tensors

Let $V^{(1)}, \dots, V^{(d)}$ be finite dimensional inner product spaces over $\mathbb{R}$. The key observation is the following. Let $J_1 \cup \cdots \cup J_e = \{1, \dots, d\}$ be a partition of $\{1, \dots, d\}$. Then the natural *flattening* map

$$V^{(1)} \otimes \cdots \otimes V^{(d)} \to (\bigotimes_{j \in J_1} V^{(j)}) \otimes \cdots \otimes (\bigotimes_{j \in J_e} V^{(j)})$$

sends the set of order-d odeco tensors into the set of order-e odeco tensors, where the inner product on each factor $\bigotimes_{j \in J_\ell} V^{(j)}$ is the one induced from the inner products on the factors. The following proposition gives a strong converse to this observation.

Proposition 4.1.15. *Let $T \in V^{(1)} \otimes \cdots \otimes V^{(d)}$ be a tensor, where $d \geq 4$. Suppose that the flattenings of T with respect to the three partitions*

(i) $\{1\}, \ldots, \{d-3\}, \{d-2\}, \{d-1, d\}$,

(ii) $\{1\}, \ldots, \{d-3\}, \{d-2, d-1\}, \{d\}$, *and*

(iii) $\{1\}, \ldots, \{d-3\}, \{d-2, d\}, \{d-1\}$

are all odeco. Then so is T.

The lower bound of 4 in this proposition is essential, because any flattening of a three-tensor is a matrix and hence odeco, but as we have seen in Section 4.1.3 not every three-tensor is odeco.

Proof. As the first two flattenings are odeco, we have orthogonal decompositions

$$T = \sum_{i=1}^{k} T_i \otimes u_i \otimes A_i = \sum_{\ell=1}^{r} T'_\ell \otimes B_\ell \otimes w_\ell$$

where $A_1, \ldots, A_k \in V^{(d-1)} \otimes V^{(d)}$ are pairwise orthogonal and nonzero, and so are $u_1, \ldots, u_k \in V^{(d-2)}$, and the T_i are of the form $z_{i1} \otimes \cdots \otimes z_{i(d-3)}$ where for each j the z_{ij}, $i = 1, \ldots, k$ are pairwise orthogonal and nonzero. Similarly for the factors in the second expression. Contracting T with T_i in the first $d-3$ factors yields a single term on the left (here we use that $d > 3$):

$$(T_i|T_i)u_i \otimes A_i = \sum_{\ell=1}^{r} (T'_\ell|T_i) B_\ell \otimes w_\ell.$$

Since the let-hand side is nonzero, there exists at least one index ℓ such that $(T'_\ell|T_i)$ is nonzero. For such an index ℓ contract both sides with w_ℓ. We find that $B_\ell = u_i \otimes v_\ell$ with $v_\ell = A_i w_\ell \in V^{(d-1)}$. This means that B_ℓ is of rank one. Since there is at least one such index ℓ and since the u_i are linearly independent for distinct i, and since , we find that the set of ℓ with $(T'_\ell|T_i) \neq 0$ is disjoint from the set defined similarly for another value of i. Hence, $r \geq k$. By swapping the roles of the two decompositions we also find the opposite equality, so that $r = k$, and after relabelling we find that $B_i = u_i \otimes v_i$ for $i = 1, \ldots, k$ and certain nonzero vectors $v_i = A_i w_i$.Hence we find

$$T = \sum_{i=1}^{k} T'_i \otimes u_i \otimes v_i \otimes w_i,$$

where we do not yet know whether the v_i are pairwise perpendicular. However, applying the same reasoning to the second and third decompositions in the lemma, we get that $B_i = u'_i \otimes v'_i$, and we obtain another decomposition

$$T = \sum_{i=1}^{k} T'_i \otimes u'_i \otimes v'_i \otimes w_i,$$

where we know that the v_i' are pairwise perpendicular, but not that the u_i' are. Contracting with T_i' we find that, in fact, both decompositions are equal and the v_i are pairwise perpendicular, as required. □

Proof of the Main Theorem (Theorem 4.1.3) for ordinary tensors. It follows from Lemma 4.1.12 and Proposition 4.1.13, that ordinary odeco tensors of order three are characterized by degree-two equations. By Proposition 4.1.15 and the remarks preceding it, a higher-order tensor is odeco if and only if certain of its flattenings are odeco. Thus the equations characterizing lower-order odeco tensors pull back, along linear maps, to equations characterizing higher-order odeco tensors. □

4.1.4.2 Symmetric tensors

In this section, V is a finite-dimension vector space over $\mathbb{R}$.

Proposition 4.1.16. *For $d \geq 3$, a tensor $T \in S^d(V)$ is symmetrically odeco if and only if it is odeco when considered as an ordinary tensor in $V^{\otimes d}$.*

Proof. The "only if" direction is immediate, since a symmetric orthogonal decomposition is *a fortiori* an ordinary orthogonal decomposition. For the converse, consider an orthogonal decomposition

$$T = \sum_{i=1}^{k} v_i^{(1)} \otimes \cdots \otimes v_i^{(d)},$$

where the $v_i^{(j)}$ are nonzero vectors, pairwise perpendicular for fixed j. Since T is symmetric, we have

$$T = \sum_{i} v_i^{(\pi(1))} \otimes \cdots \otimes v_i^{(\pi(d))} \tag{4.1.1}$$

for each $\pi \in S_d$. By uniqueness of the decomposition (Proposition 4.1.6), the terms in this latter decomposition are the same, up to a permutation, as the terms in the original decomposition. In particular, the unordered cardinality-k sets of projective points $Q_j := \{[v_1^{(j)}], \ldots, [v_k^{(j)}]\} \subseteq \mathbb{P}V$ are identical for all $j = 1, \ldots, d$.

Consider the integer $k \times d$-matrix A with entries in $[k] := \{1, \ldots, k\}$ determined by $a_{ij} = m$ if $[v_i^{(j)}] = [v_m^{(1)}]$. The matrix A has all integers $1, \ldots, k$ in each column, and they are in increasing order in the first column. Furthermore, A has the property that for each $d \times d$-permutation matrix π there exists a $k \times k$-permutation matrix σ such that $\sigma A = A\pi$. This is because if we permute the d columns of A by π, then, we can permute its rows so that the first column has the numbers $1, \ldots, k$ in increasing order. This is how we obtain the $k \times k$ permutation matrix σ.

To conclude the proof we only need to prove that, for $d \geq 3$, the only such $k \times d$ matrix is the matrix whose i-th row consists entirely of copies of i.

To show this, for $j \in \{2, \ldots, d\}$ pick $\pi_j = (1, j)$ to be the transposition switching 1 and j. Let the columns of A be $\mathrm{id}, \tau_2, \ldots, \tau_d$ thought of as permutations of $[k]$. By the property imposed on A there exists a σ_j such that $\sigma_j A = A\pi_j$. In particular, the first column of $(A\pi_j)_1$, which is τ_j, has to equal to the first column of $\sigma_j A$, which is σ_j. So $\tau_j = \sigma_j$ for all $j \in \{2, \ldots, d\}$. Since $d \geq 3$, one can pick an index l which is fixed by π_j, so that the l-th column of $\sigma_j A = A\pi_j$, which is τ_l, equals to $\sigma_j \tau_l$. But then $\sigma_j = \mathrm{id} = \tau_j$., and therefore the i-th row of A consists completely of the number i. This concludes the proof of Proposition 4.1.16. □

Proof of the Main Theorem (Theorem 4.1.3) for symmetric tensors. By Proposition 4.1.16, the equations for odeco tensors in $V \otimes \cdots \otimes V$ pull back to equations characterizing symmetrically odeco tensors in $S^d V$ via the inclusion of the latter space into the former. Thus the Main Theorem for symmetric tensors follows from the Main Theorem (Theorem 4.1.3) for ordinary tensors, proved in the previous subsection. □

Remark 4.1.17. The proof of the Proposition 4.1.13 in Section 4.1.3 for ordinary odeco three-tensors relies on the proof of Proposition 4.1.11 for symmetrically odeco three-tensors, so the proof above does not render the proof of Proposition 4.1.13 superfluous.

4.1.5 Concluding remarks

We have established quadratic real-algebraic characterizations of orthogonally decomposable tensors in the symmetric and ordinary case. While this is quite a satisfactory result, we still don't know if the equations that we have found generate the ideals of the real-algebraic varieties at hand? We are somewhat optimistic, because of evidence in [130] for the case of symmetrically odeco $2 \times 2 \times \cdots \times 2$-tensors.

Acknowledgements

For completing this project, we would like to thank Nick Vannieuwenhoven for several remarks on a previous draft. We would also like to thank the organizers of the Fall 2014 workshop "Tensors in Computer Science and Geometry" at the Simons Institute for the Theory of Computing, where this project started.

4.2 Frame Decomposable Tensors

A symmetric tensor of small rank decomposes into a configuration of only few vectors. We study the variety of tensors for which this configuration is a unit norm tight frame. This section is based on parts of joint work with Luke Oeding and Bernd Sturmfels titled *Decomposing tensors into frames* [121].

4.2.1 Introduction

A fundamental problem in computational algebraic geometry, with a wide range of applications, is the low rank decomposition of symmetric tensors; see e.g. [8, 26, 45, 120, 131]. If $T = (t_{i_1 i_2 \cdots i_d})$ is a symmetric tensor in $\mathrm{Sym}_d(\mathbb{C}^n)$, then such a decomposition takes the form

$$T \;=\; \sum_{j=1}^{r} \lambda_j v_j^{\otimes d}. \tag{4.2.1}$$

Here $\lambda_j \in \mathbb{C}$ and $v_j = (v_{1j}, v_{2j}, \ldots, v_{nj}) \in \mathbb{C}^n$ for $j = 1, 2, \ldots, r$. The smallest r for which a representation (4.2.1) exists is the *rank* of T. In particular, each $v_j^{\otimes d}$ is a tensor of rank 1.

An equivalent way to represent a symmetric tensor T is as the homogeneous polynomial

$$T \;\;=\;\; \sum_{i_1,\ldots,i_d=1}^{n} t_{i_1 i_2 \cdots i_d} \cdot x_{i_1} x_{i_2} \cdots x_{i_d}. \tag{4.2.2}$$

If $d = 2$, then (4.2.2) is the identification of symmetric matrices with quadratic forms. Written as a polynomial, the right hand side of (4.2.1) is a linear combination of powers of linear forms:

$$T \;\;=\;\; \sum_{j=1}^{r} \lambda_j (v_{1j} x_1 + v_{2j} x_2 + \cdots + v_{nj} x_n)^d. \tag{4.2.3}$$

The decomposition in (4.2.1) and (4.2.3) is called *Waring decomposition*. When $d = 2$, it corresponds to orthogonal diagonalization of symmetric matrices. We could subsume the constants λ_i into the vectors v_i but we prefer to leave (4.2.1) and (4.2.3) as is, for reasons to be seen shortly. The (projective) variety of all such symmetric tensors is the *r-th secant variety of the Veronese variety*. The vast literature on the geometry and equations of this variety (cf. [108]) forms the mathematical foundation for low rank decomposition algorithms for symmetric tensors.

In many situations one places further restrictions on the summands in (4.2.1) and (4.2.3), such as being real and nonnegative. Applications to machine learning in [8] concern the case when $r = n$ and the vectors $v_1, \ldots, v_n$ form an orthonormal basis of $\mathbb{R}^n$. Sections 3.1 and 4.1 characterize the *odeco variety* of all tensors that admit such an orthogonal decomposition.

The present section takes this one step further by connecting tensors to *frame theory* [30, 29, 38, 57, 145]. We examine the scenario when the v_j form a *f*inite *u*nit *n*orm *t*ight *f*rame

(or funtf) of $\mathbb{R}^n$, an object of recent interest at the interface of applied functional analysis and algebraic geometry. Consider a configuration $V = (v_1, \ldots, v_r) \in (\mathbb{R}^n)^r$ of r labeled vectors in $\mathbb{R}^n$. We also regard this as an $n \times r$-matrix $V = (v_{ij})$. We call V a *funtf* if

$$V \cdot V^T = \frac{r}{n} \cdot \mathrm{Id}_n \qquad \text{and} \qquad \sum_{j=1}^{n} v_{ij}^2 = 1 \quad \text{for } i = 1, 2, \ldots, r. \tag{4.2.4}$$

This is an inhomogeneous system of $n^2 + r$ quadratic equations in nr unknowns. The *funtf variety*, denoted $\mathcal{F}_{r,n}$ as in [29], is the subvariety of complex affine space $\mathbb{C}^{n \times r}$ defined by (4.2.4). For the state of the art we refer to the article [29] by Cahill, Mixon and Strawn, and the references therein. A detailed review, with some new perspectives, will be given in Subsection 4.2.2.

We homogenize the funtf variety by attaching a scalar λ_i to each vector v_i. The result maps into the projective space $\mathbb{P}(\mathrm{Sym}_d(\mathbb{C}^n)) = \mathbb{P}^{\binom{n-1+d}{d}-1}$ of symmetric tensors, via the formulas (4.2.1) and (4.2.3). Our aim is to study the closure of the image of that map. This is denoted $\mathcal{T}_{r,n,d}$. We call it the *variety of frame decomposable tensors*, or the *fradeco variety*. Here r, n, d are positive integers with $r \geq n$. For $r = n$, $\mathcal{T}_{n,n,d}$ is the odeco variety from Sections 3.1 and 4.1.

Example 4.2.1. *Let $n = 3, d = 4$, and consider the symmetric $3\times3\times3\times3$-tensor*

$$\begin{aligned} T \;=\; & 59(x_1^4 + x_2^4 + x_3^4) \,-\, 16(x_1^3x_2 + x_1x_2^3 + x_1^3x_3 + x_2^3x_3 + x_1x_3^3 + x_2x_3^3) \\ & +\, 66(x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2) \,+\, 96(x_1^2x_2x_3 + x_1x_2^2x_3 + x_1x_2x_3^2). \end{aligned} \tag{4.2.5}$$

This ternary quartic lies in $\mathcal{T}_{4,3,4}$, i.e. this tensor has fradeco rank $r = 4$. *To see this, note that*

$$\begin{aligned} T \;=\; & \tfrac{1}{12}(-5x_1 + x_2 + x_3)^4 + \tfrac{1}{12}(x_1 - 5x_2 + x_3)^4 \\ & +\tfrac{1}{12}(x_1 + x_2 - 5x_3)^4 + \tfrac{1}{12}(3x_1 + 3x_2 + 3x_3)^4. \end{aligned} \tag{4.2.6}$$

The corresponding four vectors, appropriately scaled, form a finite unit norm tight frame:

$$V \;=\; \frac{1}{3\sqrt{3}} \begin{pmatrix} -5 & 1 & 1 & 3 \\ 1 & -5 & 1 & 3 \\ 1 & 1 & -5 & 3 \end{pmatrix} \;\in\; \mathcal{F}_{4,3}. \tag{4.2.7}$$

The fradeco variety $\mathcal{T}_{4,3,4}$ is a projective variety of dimension 6 and degree 74 in $\mathbb{P}^{14}$. It is parametrized by applying rotation matrices $\rho \in SO_3$ to all ternary quartics of the form

$$T \,=\, \lambda_1(-5x_1+x_2+x_3)^4+\lambda_2(x_1-5x_2+x_3)^4+\lambda_3(x_1+x_2-5x_3)^4+\lambda_4(3x_1+3x_2+3x_3)^4. \tag{4.2.8}$$

Our objective is to find the output (4.2.6) from the input (4.2.5). In this particular case, the decomposition can be found easily using Sylvester's classical Catalecticant Algorithm*, as explained in [120, Section 2.2]. In general, this will be more difficult to do.* ♢

The *fradeco rank* of a symmetric tensor $T \in \mathrm{Sym}_d(\mathbb{R}^n)$ is defined as the smallest r such that $T \in \mathcal{T}_{r,n,d}$. This property does not imply that T also has a frame decomposition (4.2.1) of length $r+1$. Indeed, we often have $\mathcal{T}_{r,n,d} \not\subset \mathcal{T}_{r+1,n,d}$. For instance, the odeco quartic $x_1^4 + x_2^4 + x_3^4$ lies in $\mathcal{T}_{3,3,4} \backslash \mathcal{T}_{4,3,4}$, by the constraint in Example 4.2.23. See also Example 4.2.17.

This section is organized as follows. In Subsection 4.2.2 we give an introduction to the algebraic geometry of the funtf variety $\mathcal{F}_{r,n}$. This lays the foundation for the subsequent study of fradeco tensors. Subsection 4.2.3 is concerned with the case of symmetric $2\times 2\times \cdots \times 2$-tensors T. These correspond to binary forms ($n = 2$). We characterize frame decomposable tensors in terms of rank conditions on matrices. In Subsection 4.2.4 we investigate the general case $n \geq 3$, and we present what we know about the fradeco varieties $\mathcal{T}_{r,n,d}$. Subsection 4.2.5 is devoted to numerical algorithms for studying $\mathcal{T}_{r,n,d}$ and for decomposing its elements into frames.

4.2.2 Finite unit norm tight frames

In this subsection we discuss various representations of the funtf variety $\mathcal{F}_{r,n}$. This may serve as an invitation to the emerging interaction between algebraic geometry and frame theory.

Each variety studied in this section is defined over the real field $\mathbb{R}$ and is the Zariski closure of its set of real points. This Zariski closure lives in affine or projective space over $\mathbb{C}$. For instance, SO_n is the group of $n\times n$ rotation matrices ρ, and such matrices have entries in $\mathbb{R}$. However, when referring to SO_n as an algebraic variety we mean the irreducible subvariety of $\mathbb{C}^{n\times n}$ defined by the polynomial equations $\rho \cdot \rho^T = \mathrm{Id}_n$ and $\det(\rho) = 1$. Likewise, a funtf V is a real $n \times r$ matrix, but the funtf variety $\mathcal{F}_{r,n}$ lives in $\mathbb{C}^{n\times r}$. It consists of all complex solutions to the quadratic equations (4.2.4). In the frame theory literature [29, 30, 57, 145] there is also a complex Hermitian version of $\mathcal{F}_{r,n}$, but it will not be considered in this section.

It is important to distinguish $\mathcal{F}_{r,n}$ from the variety of *Parseval frames*, here denoted $\mathcal{P}_{r,n}$. The latter is much easier than the former. The variety $\mathcal{P}_{r,n}$ is defined by the matrix equation

$$V \cdot V^T = \mathrm{Id}_n.$$

The real points on $\mathcal{P}_{r,n}$ are smooth and Zariski dense, and they form the Stiefel manifold of all orthogonal projections $\mathbb{R}^r \to \mathbb{R}^n$. Hence $\mathcal{P}_{r,n}$ is irreducible of dimension $nr - \binom{n+1}{2}$.

One feature that distinguishes $\mathcal{P}_{r,n}$ from $\mathcal{F}_{r,n}$ is the existence of a canonical map $\mathcal{P}_{r,n+1} \to \mathcal{P}_{r,n}$. Indeed, by Naimark's Theorem [39], every Parseval frame is the orthogonal projection of an orthonormal basis of $\mathbb{R}^r$, so we can add a row to $V \in \mathcal{P}_{r,n}$ and get a matrix in $\mathcal{P}_{r,n+1}$. There is no analogous statement for the variety $\mathcal{F}_{r,n}$. We begin with the following result.

Theorem 4.2.2. *The dimension of the funtf variety $\mathcal{F}_{r,n}$ is*

$$\dim(\mathcal{F}_{r,n}) \;=\; (n-1)\cdot(r - \frac{n}{2} - 1) \qquad \text{provided } r > n \geq 2. \tag{4.2.9}$$

It is irreducible when $r \geq n+2 > 4$.

r	n	$\dim \mathcal{F}_{r,n}$	$\deg \mathcal{F}_{r,n}$	# components & degrees
3	2	1	$8 \cdot 2$	8 components, each degree 2
4	2	2	$12 \cdot 4$	12 components, each degree 4
5	2	3	112	irreducible
6	2	4	240	irreducible
7	2	5	496	irreducible
4	3	3	$16 \cdot 8$	16 components, each degree 8
5	3	5	1024	irreducible
6	3	7	2048	irreducible
7	3	9	4096	irreducible
5	4	6	$32 \cdot 40$	32 components, each degree 40
6	4	9	20800	irreducible
7	4	12	65536	irreducible

Table 4.1: Dimension and degree of the funtf variety in some small cases

Proof. Cahill, Mixon and Strawn [29, Theorem 1.4] proved that $\mathcal{F}_{r,n}$ is irreducible when $r \geq n+2 > 4$. The dimension formula comes from two articles: one by Dykema and Strawn [57, Theorem 4.3(ii)] regarding the case when r and n are relatively prime, and one by Strawn [145, Corollary 3.5] which studies the local geometry for all r, n. In these articles it is shown that the real points in $\mathcal{F}_{r,n}$ have a dense open subset that forms a manifold of dimension $(n-1) \cdot (r - \frac{n}{2} - 1)$. The arguments in [29] show that the real points are Zariski dense in the complex variety $\mathcal{F}_{r,n}$. Hence (4.2.9) is the correct formula for the dimension of $\mathcal{F}_{r,n}$. □

Next to the dimension, the most important invariant of an algebraic variety is its *degree*. By this we mean the degree of its projective closure [46, §8.4]. This can be computed using symbolic software for Gröbner bases, or using numerical algebraic geometry software. The dimension and degree of $\mathcal{F}_{r,n}$ for small r, n in Table 4.1 were computed using `Bertini` [15].

The case $r = n+1$ is special. Here, the funtf variety decomposes into 2^{n+1} irreducible components, each of which is affinely isomorphic to the $\binom{n}{2}$-dimensional variety SO_n. This will be explained in Corollary 4.2.11. The next example discusses one other exceptional case.

Example 4.2.3 ($r = 4, n = 2$)**.** *Following (4.2.4), the defining ideal of the funtf variety $\mathcal{F}_{4,2}$ equals*

$$\langle\, v_{11}^2 + v_{12}^2 + v_{13}^2 + v_{14}^2 - 2,\, v_{11}v_{21} + v_{12}v_{22} + v_{13}v_{23} + v_{14}v_{24} \rangle \,+\, \langle\, v_{1j}^2 + v_{2j}^2 - 1 \,:\, j = 1,2,3,4 \,\rangle.$$

Note that this contains $v_{21}^2 + v_{22}^2 + v_{23}^2 + v_{24}^2 - 2$. Using Gröbner basis software, such as `Macaulay2` *[81], one checks that this ideal equals the intersection of the six given quadrics, it is radical, and its degree is* 48*. Primary decomposition reveals that this ideal is the intersection of* 12 *prime ideals, each of degree* 4*. One of these associated primes is*

$$\langle\, v_{11} - v_{22}, v_{12} + v_{21},\, v_{13} - v_{24}, v_{23} + v_{14},\, v_{23}^2 + v_{24}^2 - 1, v_{21}^2 + v_{22}^2 - 1 \rangle.$$

The irreducible variety of this particular prime ideal consists of the 2×4*-matrices*

$$V \quad = \quad \big(R_1 \mid R_2\big)\,,$$

where R_1 *and* R_2 *are rotation matrices of format* 2×2*. The other* 11 *components are obtained by replacing* R_i *with* $-R_i$ *and permuting columns. The image of* V *under the map to binary forms is a linear combination of two odeco forms, one given by* R_1 *and the other by* R_2*.* $\diamondsuit$

The real points of $\mathcal{F}_{r,n}$ live in $(\mathbb{S}^{n-1})^r$ where $\mathbb{S}^{n-1} = \{u \in \mathbb{R}^n : \sum_{i=1}^n u_i^2 = 1\}$ denotes the unit sphere. However, the vectors on these spheres will get scaled by the multipliers $\lambda_i^{1/d}$ in (4.2.3) when we pass to the fradeco variety $\mathcal{T}_{r,n,d}$. To achieve better geometric properties and computational speed, we map each real sphere $\mathbb{S}^{n-1}$ to complex projective $(n-1)$-space $\mathbb{P}^{n-1}$.

The *projective funtf variety* $\mathcal{G}_{r,n}$ is the image of $\mathcal{F}_{r,n}$ in $(\mathbb{P}^{n-1})^r$. To describe its equations, we use an $n \times r$-matrix $V = (v_{ij})$ of unknowns as before, but now the i-th column of V represents coordinates on the i-th factor of $(\mathbb{P}^{n-1})^r$. We introduce the $r \times r$ diagonal matrix

$$D \quad = \quad \operatorname{diag}\Big(\sum_{i=1}^n v_{i1}^2\,,\,\sum_{i=1}^n v_{i2}^2\,,\,\ldots\,,\,\sum_{i=1}^n v_{ir}^2\Big). \tag{4.2.10}$$

The variety $\mathcal{G}_{r,n}$ is defined by the following matrix equation:

$$V \cdot D^{-1} \cdot V^T \,=\, \frac{r}{n} \cdot \mathrm{Id}_n. \tag{4.2.11}$$

Each entry on the left hand side is a homogeneous rational function of degree 0. In fact, these functions are multihomogeneous: they define rational functions on $(\mathbb{P}^{n-1})^r$.

The challenge is to clear denominators in (4.2.11), so as to obtain a system of polynomial equations that defines $\mathcal{G}_{r,n}$ as a subvariety of $(\mathbb{P}^{n-1})^r$. Next we solve this problem for $n = 2$.

For planar frames, equation (4.2.11) translates into the vanishing of the two rational functions

$$P \,=\, \sum_{j=1}^r \frac{2v_{1j}^2}{v_{1j}^2+v_{2j}^2} \,-\, r \qquad \text{and} \qquad Q \,=\, \sum_{j=1}^r \frac{2v_{1j}v_{2j}}{v_{1j}^2+v_{2j}^2}. \tag{4.2.12}$$

Consider the numerator of the rational function

$$P - iQ \quad = \quad \sum_{j=1}^r \frac{v_{1j}^2 - 2iv_{1j}v_{2j} - v_{2j}^2}{v_{1j}^2+v_{2j}^2} \quad = \quad \sum_{j=1}^r \frac{v_{1j}-v_{2j}i}{v_{1j}+v_{2j}i}, \qquad \text{where } i = \sqrt{-1}.$$

Let $\widetilde{P}$ and $\widetilde{Q}$ denote the real part and the imaginary part of that numerator. These are two multilinear polynomials of degree r with integer coefficients in $v_{11}, v_{12}, \ldots, v_{2r}$. They define a complete intersection, and, by construction, this is precisely our funtf variety in $(\mathbb{P}^1)^r$:

Lemma 4.2.4. *The projective funtf variety* $\mathcal{G}_{r,2}$ *is a complete intersection of codimension* 2 *in* $(\mathbb{P}^1)^r$*, namely, it is the zero set of the two multilinear forms* $\widetilde{P}$ *and* $\widetilde{Q}$*.*

Here are explicit formulas for the multilinear forms that define $\mathcal{G}_{r,2}$ when $r \leq 5$:

Example 4.2.5. *If $r = 3$, then $\widetilde{P} = 3v_{11}v_{12}v_{13} + v_{11}v_{22}v_{23} + v_{21}v_{12}v_{23} + v_{21}v_{22}v_{13}$ and $\widetilde{Q} = v_{11}v_{12}v_{23}+v_{11}v_{22}v_{13}+v_{21}v_{12}v_{13}+3v_{21}v_{22}v_{23}$. If r=4, then $\widetilde{P} = 4(v_{11}v_{12}v_{13}v_{14}-v_{21}v_{22}v_{23}v_{24})$ and*

$$\begin{aligned}\widetilde{Q} \;=\; & 2v_{11}v_{12}v_{13}v_{24} + 2v_{11}v_{12}v_{23}v_{14} + 2v_{11}v_{22}v_{13}v_{14} + 2v_{11}v_{22}v_{23}v_{24} + \\ & 2v_{21}v_{12}v_{13}v_{14} + 2v_{21}v_{12}v_{23}v_{24} + 2v_{21}v_{22}v_{13}v_{24} + 2v_{21}v_{22}v_{23}v_{14}.\end{aligned}$$

If $r = 5$, then

$$\begin{aligned}\widetilde{P} \;=\; & 5v_{11}v_{12}v_{13}v_{14}v_{15} - v_{11}v_{12}v_{13}v_{24}v_{25} - v_{11}v_{12}v_{23}v_{14}v_{25} - v_{11}v_{12}v_{23}v_{24}v_{15} \\ & -v_{11}v_{22}v_{13}v_{14}v_{25} - v_{11}v_{22}v_{13}v_{24}v_{15} - v_{11}v_{22}v_{23}v_{14}v_{15} - 3v_{11}v_{22}v_{23}v_{24}v_{25} \\ & -v_{21}v_{12}v_{13}v_{14}v_{25} - v_{21}v_{12}v_{13}v_{24}v_{15} - v_{21}v_{12}v_{23}v_{14}v_{15} - 3v_{21}v_{12}v_{23}v_{24}v_{25} \\ & -v_{21}v_{22}v_{13}v_{14}v_{15} - 3v_{21}v_{22}v_{13}v_{24}v_{25} - 3v_{21}v_{22}v_{23}v_{14}v_{25} - 3v_{21}v_{22}v_{23}v_{24}v_{15},\end{aligned}$$

and $\widetilde{Q}$ is obtained from $\widetilde{P}$ by switching the two rows of V. ♢

Such formulas are useful for parametrizing frames. We write the equations for $\mathcal{G}_{r,2}$ as

$$\begin{pmatrix}\widetilde{P}\\ \widetilde{Q}\end{pmatrix} \;=\; \begin{pmatrix}m_{11} & m_{12}\\ m_{21} & m_{22}\end{pmatrix} \cdot \begin{pmatrix}v_{1r}\\ v_{2r}\end{pmatrix} \;=\; \begin{pmatrix}0\\ 0\end{pmatrix}.$$

The matrix entries m_{ij} are multilinear forms in $(v_{11} : v_{21})$, $(v_{12} : v_{22})$, …, $(v_{1,r-1} : v_{2,r-1})$. Using the quadratic formula, we solve the following equation for one of its unknowns:

$$m_{11}m_{22} \;=\; m_{12}m_{21}. \tag{4.2.13}$$

This defines a hypersurface in $(\mathbb{P}^1)^{r-1}$, from which we can now easily sample points. The point in the remaining rth factor $\mathbb{P}^1$ is then recovered by setting $v_{1r} = m_{12},\ v_{2r} = -m_{11}$.

For $n \geq 3$, we do not know the generators of the multihomogeneous prime ideal of $\mathcal{G}_{r,n}$. Here are two instances where `Macaulay2` [81] succeeded in computing these ideals:

Example 4.2.6. *The variety $\mathcal{G}_{4,3}$ is a threefold in $(\mathbb{P}^2)^4$. Its ideal is generated by 34 quartics. Among them are the equations that define the six coordinate projections into $(\mathbb{P}^2)^2$, like*

$$\begin{aligned}&8(v_{11}^2v_{12}^2 + v_{21}^2v_{22}^2 + v_{31}^2v_{32}^2) + 18(v_{11}v_{21}v_{12}v_{22} + v_{11}v_{31}v_{12}v_{32} + v_{21}v_{31}v_{22}v_{32}) \\ &-v_{11}^2v_{22}^2 - v_{11}^2v_{32}^2 - v_{21}^2v_{12}^2 - v_{21}^2v_{32}^2 - v_{31}^2v_{12}^2 - v_{31}^2v_{22}^2.\end{aligned}$$

Example 4.2.7. *Let $r = 5$ and $n = 3$. By saturating the denominators in (4.2.11), we found that the ideal of $\mathcal{G}_{5,3}$ is generated by a 120-dimensional SO_3-invariant space of sextics. The following polynomial (with 60 terms of $\mathbb{Z}^5$-degree $(2,2,2,0,0)$) is a highest weight vector:*

$$\begin{aligned}&50v_{11}^2v_{12}^2v_{13}^2 + 5v_{11}^2v_{12}^2v_{23}^2 + 5v_{11}^2v_{12}^2v_{33}^2 + 45v_{11}^2v_{12}v_{22}v_{13}v_{23} + 45v_{11}^2v_{12}v_{32}v_{13}v_{33} + 5v_{11}^2v_{22}^2v_{13}^2 + 5v_{11}^2v_{22}^2v_{23}^2 - 4v_{11}^2v_{22}^2v_{33}^2 \\ &+18v_{11}^2v_{22}v_{32}v_{23}v_{33} + 5v_{11}^2v_{32}^2v_{13}^2 - 4v_{11}^2v_{32}^2v_{23}^2 + 5v_{11}^2v_{32}^2v_{33}^2 + 45v_{11}v_{21}v_{12}^2v_{13}v_{23} + 45v_{11}v_{21}v_{12}v_{22}v_{13}^2 + 18v_{11}v_{21}v_{32}^2v_{13}v_{23} \\ &+45v_{11}v_{21}v_{12}v_{22}v_{23}^2 + 18v_{11}v_{21}v_{12}v_{22}v_{33}^2 + 27v_{11}v_{21}v_{12}v_{32}v_{23}v_{33} + 45v_{11}v_{21}v_{22}^2v_{13}v_{23} + 27v_{11}v_{21}v_{22}v_{32}v_{13}v_{33} \\ &+45v_{11}v_{31}v_{12}^2v_{13}v_{33} + 27v_{11}v_{31}v_{12}v_{22}v_{23}v_{33} + 45v_{11}v_{31}v_{12}v_{32}v_{13}^2 + 18v_{11}v_{31}v_{12}v_{32}v_{23}^2 + 45v_{11}v_{31}v_{12}v_{32}v_{33}^2 - 4v_{21}^2v_{12}^2v_{33}^2 \\ &+18v_{11}v_{31}v_{22}^2v_{13}v_{33} + 27v_{11}v_{31}v_{22}v_{32}v_{13}v_{23} + 45v_{11}v_{31}v_{32}^2v_{13}v_{33} + 5v_{21}^2v_{12}^2v_{13}^2 + 5v_{21}^2v_{12}^2v_{23}^2 + 45v_{21}^2v_{12}v_{22}v_{13}v_{23} \\ &+18v_{21}^2v_{12}v_{32}v_{13}v_{33} + 5v_{21}^2v_{22}^2v_{13}^2 + 50v_{21}^2v_{22}^2v_{23}^2 + 5v_{21}^2v_{22}^2v_{33}^2 + 45v_{21}^2v_{22}v_{32}v_{23}v_{33} - 4v_{21}^2v_{32}^2v_{13}^2 + 5v_{21}^2v_{32}^2v_{23}^2 + 5v_{21}^2v_{32}^2v_{33}^2 \\ &+18v_{21}v_{31}v_{12}^2v_{23}v_{33} + 27v_{21}v_{31}v_{12}v_{22}v_{13}v_{33} + 27v_{21}v_{31}v_{12}v_{32}v_{13}v_{23} + 45v_{21}v_{31}v_{22}^2v_{23}v_{33} + 18v_{21}v_{31}v_{22}v_{32}v_{13}^2 \\ &+45v_{21}v_{31}v_{22}v_{32}v_{23}^2 + 45v_{21}v_{31}v_{22}v_{32}v_{33}^2 + 45v_{21}v_{31}v_{32}^2v_{23}v_{33} + 5v_{31}^2v_{12}^2v_{13}^2 - 4v_{31}^2v_{12}^2v_{23}^2 + 5v_{31}^2v_{12}^2v_{33}^2 + 18v_{31}^2v_{12}v_{22}v_{13}v_{23} \\ &+45v_{31}^2v_{12}v_{32}v_{13}v_{33} - 4v_{31}^2v_{22}^2v_{13}^2 + 5v_{31}^2v_{22}^2v_{23}^2 + 5v_{31}^2v_{22}^2v_{33}^2 + 45v_{31}^2v_{22}v_{32}v_{23}v_{33} + 5v_{31}^2v_{32}^2v_{13}^2 + 5v_{31}^2v_{32}^2v_{23}^2 + 50v_{31}^2v_{32}^2v_{33}^2.\end{aligned}$$

The ideal of $\mathcal{G}_{5,3} \subset (\mathbb{P}^2)^5$ *has* 10 *generators like this, each spanning a one-dimensional graded component. It has* 30 *components of degrees like* $(2,2,1,1,0)$*, each generated by a polynomial with* 78 *terms. Finally, it has five* 16*-dimensional components of degrees like* $(2,1,1,1,1)$. $\diamondsuit$

In order to sample points from the funtf variety $\mathcal{F}_{r,n}$, we can also use the following parametrization found in [30, 145]. We write $V = (U', W)$, where U' is an $n \times n$-matrix and W is an $(r-n)\times n$-matrix. For the columns of W we take arbitrary points on the unit sphere $\mathbb{S}^{n-1}$. In practice, it is convenient to fix a rational parametrization of $\mathbb{S}^{n-1}$, so as to ensure that W has rational entries w_{ij}. For instance, for $n = 3$ we use the following formulas:

$$w_{1j} = \frac{2\lambda_j\mu_j}{\lambda_j^2+\mu_j^2+\nu_j^2},\ w_{2j} = \frac{2\lambda_j\nu_j}{\lambda_j^2+\mu_j^2+\nu_j^2},\ w_{3j} = \frac{\lambda_j^2-\mu_j^2-\nu_j^2}{\lambda_j^2+\mu_j^2+\nu_j^2},\quad \text{where } \lambda_j,\mu_j,\nu_j \in \mathbb{Z}. \tag{4.2.14}$$

After these choices have been made, we fix the following $n \times n$-matrix with entries in $\mathbb{Q}$:

$$S \ = \ \frac{r}{n} \cdot \mathrm{Id}_n \ - \ W \cdot W^T. \tag{4.2.15}$$

It now remains to study all $n \times n$-matrices $U = (u_{ij})$ that satisfy

$$U \cdot D^{-1} \cdot U^T \ = \ S, \text{ where } \ D \ = \ \mathrm{diag}\big(\sum_{i=1}^n u_{i1}^2\,,\,\ldots\,,\,\sum_{i=1}^n u_{in}^2\big).$$

For any such U we get a funtf $V = (U', W) \in \mathcal{F}_{r,n}$ by setting $U' = U \cdot D^{-1/2}$. For random choices in (4.2.14), the matrix S is invertible, and the previous equation is equivalent to

$$D \ = \ U^T \cdot S^{-1} \cdot U. \tag{4.2.16}$$

This identity of symmetric matrices defines $\binom{n+1}{2}$ equations in the entries u_{ij} of U. The equation in position (i,j) is bilinear in $(u_{1i}, u_{2i}, \ldots, u_{ni})$ and $(u_{1j}, u_{2j}, \ldots, u_{nj})$. We solve the system (4.2.16) iteratively for the columns of U. We begin with the (1,1) entry of (4.2.16). There are $n-1$ degrees of freedom to fill in the first column of U, then $n-2$ degrees of freedom to fill in the second column, etc. This involves repeatedly solving quadratic equations in one variable, so each solution lives in a tower of quadratic extensions over $\mathbb{Q}$. In summary:

Proposition 4.2.8. *Let the columns of the* $(r-n)\times n$ *matrix* W *be arbitrary points on the unit sphere* $\mathbb{S}^{n-1}$ *coming from a rational parametrization such as in* (4.2.14)*. Then, the equations (4.2.15) and (4.2.16) represent a parametrization of* $\mathcal{F}_{r,n}$*.*

The rotation group SO_n acts by left multiplication on the funtf variety $\mathcal{F}_{r,n}$. There is a natural way to construct the quotient $\mathcal{F}_{r,n}/\mathrm{SO}_n$ as an algebraic variety, namely by mapping it into the Grassmannian $\mathrm{Gr}(n,r)$ of n-dimensional subspaces of $\mathbb{C}^r$. This is described by Cahill and Strawn in [30, Section 3.1], and we briefly develop some basic algebraic properties.

We here define $\mathrm{Gr}(n,r)$ to be the image of the *Plücker map* $\mathbb{C}^{n\times r} \to \mathbb{C}^{\binom{r}{n}}$ that takes an $n \times r$-matrix V to its vector $p = p(V)$ of $n \times n$-minors. The coordinates p_I of p are

indexed by the set $\binom{[r]}{n}$ of n-element subsets of $[r] = \{1, 2, \ldots, r\}$. With this definition, $\mathrm{Gr}(n, r)$ is the affine subvariety of $\mathbb{C}^{\binom{r}{n}}$ defined by the *quadratic Plücker relations*, such as $p_{12}p_{34} - p_{13}p_{24} + p_{14}p_{23} = 0$ for $n = 2, r = 4$. The dimension of $\mathrm{Gr}(n, r)$ is $(r-n)n+1$. Note that if $VV^T = (r/n) \cdot \mathrm{Id}_n$, then the *Cauchy-Binet formula* (cf. [30, Prop. 6]) implies

$$\sum_{I \in \binom{[r]}{n}} p_I^2 \;=\; \left(\frac{r}{n}\right)^n. \tag{4.2.17}$$

The real points in $\mathrm{Gr}(n, r)$, up to scaling, correspond to n-dimensional subspaces of $\mathbb{R}^r$.

Proposition 4.2.9. *The image of $\mathcal{F}_{r,n}$ under the Plücker map is an affine variety of dimension $(r-n)n - r + 2$ in the Grassmannian $Gr(n, r) \subset \mathbb{C}^{\binom{r}{n}}$. It is defined by the equations*

$$\sum_{I: i \in I} p_I^2 \;=\; \left(\frac{r}{n}\right)^{n-1} \qquad \text{for } i = 1, 2, \ldots, r. \tag{4.2.18}$$

The real points in this image correspond to SO_n-orbits of n-dimensional frames in $\mathcal{F}_{r,n}$.

Note that adding up the r relations in (4.2.18) and dividing by n gives precisely (4.2.17).

Proof. Both $\mathcal{F}_{r,n}$ and the constraints (4.2.18) are invariant under SO_n. Suppose that $V \in \mathbb{C}^{n \times r}$ satisfies $VV^T = (r/n) \cdot \mathrm{Id}_n$. We may assume (modulo SO_n) that the i-th column of V is $(\alpha, 0, \ldots, 0)^T$ for some $\alpha \in \mathbb{C}$. Let $\tilde{V}$ be the matrix obtained from V by deleting the first row and i-th column. Then $\tilde{V} \cdot \tilde{V}^T = (r/n) \cdot \mathrm{Id}_{n-1}$. Any p_I with $i \in I$ equals α times the maximal minor of $\tilde{V}$ indexed by $I \backslash \{i\}$. Applying (4.2.17) to $\tilde{V}$, this gives

$$\sum_{I: i \in I} p_I^2 \;=\; \alpha^2 \cdot \left(\frac{r}{n}\right)^{n-1}.$$

Hence (4.2.18) holds if and only if $\alpha = \pm 1$, and this holds for all i if and only if V lies in $\mathcal{F}_{r,n}$. The dimension formula follows from Theorem 4.2.2 because SO_n acts faithfully on $\mathcal{F}_{r,n}$. □

Example 4.2.10. *Let $n = 2$. If $r = 5$, then our construction realizes $\mathcal{F}_{5,2}/SO_2$ as an irreducible surface of degree 80 in $\mathbb{C}^{10}$. Its prime ideal is generated by the ten quadratic polynomials*

$$p_{14}p_{23} - p_{13}p_{24} + p_{12}p_{34},\ p_{15}p_{23} - p_{13}p_{25} + p_{12}p_{35},\ p_{15}p_{24} - p_{14}p_{25} + p_{12}p_{45},\ p_{15}p_{34} - p_{14}p_{35} + p_{13}p_{45},\ p_{25}p_{34} - p_{24}p_{35} + p_{23}p_{45},\ p_{12}^2 + p_{13}^2 + p_{14}^2 + p_{15}^2 - 5/2,\ p_{12}^2 + p_{23}^2 + p_{24}^2 + p_{25}^2 - 5/2,\ p_{13}^2 + p_{23}^2 + p_{34}^2 + p_{35}^2 - 5/2,\ p_{14}^2 + p_{24}^2 + p_{34}^2 + p_{45}^2 - 5/2,\ p_{15}^2 + p_{25}^2 + p_{35}^2 + p_{45}^2 - 5/2.$$

If $r = 4$, then $\mathcal{F}_{4,2}/SO_2$ is a reducible curve of degree 24 in $\mathbb{C}^6$. Its defining equations are

$$p_{14}p_{23} - p_{13}p_{24} + p_{12}p_{34} = 0,\ p_{12}^2 + p_{13}^2 + p_{14}^2 = p_{12}^2 + p_{23}^2 + p_{24}^2 = p_{13}^2 + p_{23}^2 + p_{34}^2 = p_{14}^2 + p_{24}^2 + p_{34}^2 = 2.$$

As in Example 4.2.3, this curve breaks into 12 components. One of these 12 irreducible curves is $\{ p \in \mathbb{C}^6 : p_{12} = p_{34} = 1,\ p_{13} = p_{24},\ p_{14} = -p_{23},\ p_{23}^2 + p_{24}^2 = 1 \}$. ♢

The analogous decomposition is found easily for the case $r = n+1$. Here, there are no Plücker relations, so $\mathrm{Gr}(n, n+1) \simeq \mathbb{S}^n$. For convenience of notation, we set $q_i = p_{[n+1]\setminus\{i\}}$ in (4.2.18). The quotient space $\mathcal{F}_{n+1,n}/\mathrm{SO}_n$ is the subvariety of $\mathbb{C}^{n+1}$ defined by the equations

$$q_1^2 + q_2^2 + \cdots + q_n^2 + q_{n+1}^2 \;=\; (n+1)^{n-1}/n^n + \; q_i^2 \qquad \text{for } i = 1, 2, \ldots, n+1.$$

These are equivalent to the following equations, which imply Corollary 4.2.11:

$$q_1^2 \;=\; q_2^2 \;=\; q_3^2 \;=\; \cdots \;=\; q_{n+1}^2 \;=\; (n+1)^{n-1}/n^{n+1}.$$

Corollary 4.2.11. *The quotient space $\mathcal{F}_{n+1,n}/SO_n$ is a variety consisting of 2^{n+1} isolated points in $\mathbb{R}^{n+1} = Gr(n, n+1)$, namely those points with coordinates $\pm(n+1)^{(n-1)/2}/n^{(n+1)/2}$.*

Any of the 2^{n+1} components of $\mathcal{F}_{n+1,n}$ can be used to parametrize our variety $\mathcal{T}_{n+1,n,d}$.

Example 4.2.12. *Let $n = 3$. The point $p = \sqrt{3}(\frac{4}{9}, \frac{4}{9}, \frac{4}{9}, \frac{4}{9})$ in $Gr(3, 4)$ corresponds to the SO_3-orbit of the frame V in Example 4.2.1. The variety $\mathcal{G}_{4,3}$ can be parametrized as follows:*

$$V = (v_{ij}) \;=\; \begin{bmatrix} 1-2y^2-2z^2 & 2xy-2zw & 2xz+2yw \\ 2xy+2zw & 1-2x^2-2z^2 & 2yz-2xw \\ 2xz-2yw & 2yz+2xw & 1-2x^2-2y^2 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 & -5 \\ 3 & 1 & -5 & 1 \\ 3 & -5 & 1 & 1 \end{bmatrix} \begin{bmatrix} \nu_1 & 0 & 0 & 0 \\ 0 & \nu_2 & 0 & 0 \\ 0 & 0 & \nu_3 & 0 \\ 0 & 0 & 0 & \nu_4 \end{bmatrix}.$$

The 3×3-matrix on the left is the familiar parametrization of SO_3 via unit quaternions. This gives the parametrization of the fradeco variety $\mathcal{T}_{4,3,d}$ seen in (4.2.8). ♢

The embedding of $\mathcal{F}_{r,n}/\mathrm{SO}_n$ into $\mathrm{Gr}(r, n)$ via (4.2.18) connects frame theory with *matroid theory*. The matroid of V is given by the set of Plücker coordinates p_I that are zero. If all Plücker coordinates are nonzero, then the matroid is uniform. It is a natural to ask which matroids are realizable over $\mathbb{R}$ when the additional constraints (4.2.18) are imposed.

The discussion in [30, Section 3.2] relates frame theory to the study of *orbitopes* [136]. Cahill and Strawn set up an optimization problem for computing Parseval frames that are most uniform. Their formulation in [30, p. 24] is a linear program over the *Grassmann orbitope*, which is the convex hull of $\mathrm{Gr}(n, r)$ intersected with (4.2.17). The same optimization problem makes sense with $\mathrm{Gr}(n, r)$ replaced by $\mathcal{F}_{r,n}/\mathrm{SO}_n$, or, algebraically, with (4.2.17) replaced by (4.2.18). If $n = 2$, then the former problem is a *semidefinite program*. This is the content of [136, Theorem 7.3]. For $n \geq 3$, the situation is more complicated, but the considerable body of results coming from calibrated manifolds, such as [136, Theorem 7.5], should still be helpful.

4.2.3 Binary forms

We now commence our study of the fradeco variety $\mathcal{T}_{r,n,d}$. In this subsection we focus on the case $n = 2$ of binary forms that are decomposable into small frames. The case $r = 2$

is the odeco surface known from [131, §3]. Proposition 3.6 in [131] gives an explicit list of quadrics that forms a Gröbner basis for the prime ideal of $\mathcal{T}_{2,2,d}$, and these are here expressed as the 2×2-minors of a certain $3 \times (d-3)$-matrix $\mathcal{M}_4$. What follows is our main result in subsection 4.2.3. We are using coordinates $(t_0 : \cdots : t_d)$ for the space $\mathbb{P}^d = \mathbb{P}(\mathrm{Sym}_d(\mathbb{C}^2))$ of binary forms. In the notation of (4.2.2), the coordinate t_i would be $t_{111\cdots1222\cdots2}$ with i indices 1 and $d-i$ indices 2.

Theorem 4.2.13. *Fix $r \in \{3, 4, \ldots, 9\}$ and $d \geq 2r-2$. There exists a matrix $\mathcal{M}_r$ such that:*

(a) Its maximal minors form a Gröbner basis for the prime ideal of $\mathcal{T}_{r,2,d}$.

(b) It has $r-1$ rows and $d-r+1$ columns, and the entries are linear forms in $t_0, \ldots, t_d$.

(c) Each column involves r of the unknowns t_i, and they are identical up to index shifts.

These matrices can be chosen as follows:

$$\mathcal{M}_3 = \begin{pmatrix} t_0-3t_2 & t_1-3t_3 & t_2-3t_4 & t_3-3t_5 & \cdots & t_{d-3}-3t_{d-1} \\ 3t_1-t_3 & 3t_2-t_4 & 3t_3-t_5 & 3t_4-t_6 & \cdots & 3t_{d-2}-t_d \end{pmatrix} \tag{4.2.19}$$

$$\mathcal{M}_4 = \begin{pmatrix} t_0+t_4 & t_1+t_5 & t_2+t_6 & t_3+t_7 & \cdots & t_{d-4}+t_d \\ t_1-t_3 & t_2-t_4 & t_3-t_5 & t_4-t_6 & \cdots & t_{d-3}-t_{d-1} \\ t_2 & t_3 & t_4 & t_5 & \cdots & t_{d-2} \end{pmatrix} \tag{4.2.20}$$

$$\mathcal{M}_5 = \begin{pmatrix} t_0+5t_2 & t_1+5t_3 & t_2+5t_4 & t_3+5t_5 & \cdots & t_{d-5}+5t_{d-3} \\ t_1-3t_3 & t_2-3t_4 & t_3-3t_5 & t_4-3t_6 & \cdots & t_{d-4}-3t_{d-2} \\ 3t_2-t_4 & 3t_3-t_5 & 3t_4-t_6 & 3t_5-t_7 & \cdots & 3t_{d-3}-t_{d-1} \\ 5t_3+t_5 & 5t_4+t_6 & 5t_5+t_7 & 5t_6+t_8 & \cdots & 5t_{d-2}+t_d \end{pmatrix} \tag{4.2.21}$$

$$\mathcal{M}_6 = \begin{pmatrix} t_0+3t_2 & t_1+3t_3 & t_2+3t_4 & t_3+3t_5 & \cdots & t_{d-6}+3t_{d-4} \\ t_1+t_5 & t_2+t_6 & t_3+t_7 & t_4+t_8 & \cdots & t_{d-5}+t_{d-1} \\ t_2-t_4 & t_3-t_5 & t_4-t_6 & t_5-t_7 & \cdots & t_{d-4}-t_{d-2} \\ t_3 & t_4 & t_5 & t_6 & \cdots & t_{d-3} \\ 3t_4+t_6 & 3t_5+t_7 & 3t_6+t_8 & 3t_7+t_9 & \cdots & 3t_{d-2}+t_d \end{pmatrix} \tag{4.2.22}$$

The first column of $\mathcal{M}_7$ is $(3t_0+7t_2, t_1+5t_3, t_2-3t_4, 3t_3-t_5, 5t_4+t_6, 7t_5+3t_7)^T$, the first column of $\mathcal{M}_8$ is $(t_0+2t_2, t_1+3t_3, t_4, t_3-t_5, t_2+t_6, 3t_5+t_7, 2t_6+t_8)^T$, and the first column of $\mathcal{M}_9$ is $(5t_0+9t_2, 3t_1+7t_3, t_2+5t_4, t_3-3t_5, 3t_4-t_6, 5t_5+t_7, 7t_6+3t_8, 9t_7+5t_9)^T$.

We conjecture that the same result holds for all r, and we explain what we currently know after the proof. Let us begin with a lemma concerning the dimension of our variety.

Lemma 4.2.14. *The fradeco variety $\mathcal{T}_{r,2,d}$ is irreducible and has dimension $\min(2r-3, d)$.*

Proof. For $d \geq 5$, the funtf variety $\mathcal{F}_{r,2} \subset (\mathbb{S}^1)^r$ is irreducible, by Theorem 4.2.2, and hence so is its closure $\mathcal{G}_{r,2}$ in $(\mathbb{P}^1)^r$. While the two special varieties $\mathcal{F}_{3,2}$ and $\mathcal{F}_{4,2}$ are reducible, the analyzes in Example 4.2.3 and Corollary 4.2.11 show that $\mathcal{G}_{3,2}$ and $\mathcal{G}_{4,2}$ are irreducible.

Regarding $\mathcal{G}_{r,2}$ as an affine variety in $\mathbb{C}^{2\times r}$, we obtain $\mathcal{T}_{r,2,d}$ as its image under the map

$$t_i \;=\; v_{11}^i v_{21}^{d-i} + v_{12}^i v_{22}^{d-i} + v_{13}^i v_{23}^{d-i} + \cdots + v_{1r}^i v_{2r}^{d-i} \qquad \text{for } i = 0, 1, \ldots, d. \tag{4.2.23}$$

This proves that $\mathcal{T}_{r,2,d}$ is irreducible. To see that it has the expected dimension, consider the r-th secant variety of the rational normal curve in $\mathbb{P}^d$, which is the image of the map $\mathbb{C}^{2\times r} \dashrightarrow \mathbb{P}^d$ given by (4.2.23). It is known that this secant variety has the expected dimension, namely $\min(2r-1, d)$, and the fiber dimension of the map (4.2.23) does not jump unless some 2×2-minor of $V = (v_{ij})$ is zero. Since $\operatorname{codim}(\mathcal{G}_{r,2}) = 2$, by Lemma 4.2.4, the claim follows. □

Proof of Theorem 4.2.13. We first show that the maximal minors of our matrices $\mathcal{M}_r$ vanish on the fradeco variety $\mathcal{T}_{r,2,d}$ for $r = 3, 4, \ldots, 9$. After substituting the parametrization (4.2.23) for $t_0, t_1, \ldots, t_d$, we can decompose these matrices as follows:

$$\mathcal{M}_r \;=\; M_r \cdot \begin{pmatrix} v_{11}^{d-r} & v_{11}^{d-r-1}v_{21} & v_{11}^{d-r-2}v_{21}^2 & \cdots & v_{21}^{d-r} \\ v_{12}^{d-r} & v_{12}^{d-r-1}v_{22} & v_{12}^{d-r-2}v_{22}^2 & \cdots & v_{22}^{d-r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{1r}^{d-r} & v_{1r}^{d-r-1}v_{2r} & v_{1r}^{d-r-2}v_{2r}^2 & \cdots & v_{2r}^{d-r} \end{pmatrix},$$

where

$$M_3 \;=\; \begin{pmatrix} (v_{22}^2 - 3v_{11}^2)v_{21} & (v_{22}^2 - 3v_{12}^2)v_{22} & (v_{23}^2 - 3v_{13}^2)v_{23} \\ (3v_{21}^2 - v_{11}^2)v_{11} & (3v_{22}^2 - v_{12}^2)v_{12} & (3v_{23}^2 - v_{13}^2)v_{13} \end{pmatrix},$$

$$M_4 = \begin{pmatrix} v_{21}^4 + v_{11}^4 & v_{22}^4 + v_{12}^4 & v_{23}^4 + v_{13}^4 & v_{24}^4 + v_{14}^4 \\ v_{11}v_{21}^3 - v_{11}^3 v_{21} & v_{12}v_{22}^3 - v_{12}^3 v_{22} & v_{13}v_{23}^3 - v_{13}^3 v_{23} & v_{14}v_{24}^3 - v_{14}^3 v_{24} \\ v_{11}^2 v_{21}^2 & v_{12}^2 v_{22}^2 & v_{13}^2 v_{23}^2 & v_{14}^2 v_{24}^2 \end{pmatrix},$$

$$M_5 = \begin{pmatrix} v_{21}^5 + 5v_{11}^5 & v_{22}^5 + 5v_{12}^5 & v_{23}^5 + 5v_{13}^5 & v_{24}^5 + 5v_{14}^5 & v_{25}^5 + 5v_{15}^5 \\ v_{11}v_{21}^4 - 3v_{11}^3 v_{21}^2 & v_{12}v_{22}^4 - 3v_{12}^3 v_{22}^2 & v_{13}v_{23}^4 - 3v_{13}^3 v_{23}^2 & v_{14}v_{24}^4 - 3v_{14}^3 v_{24}^2 & v_{15}v_{25}^4 - 3v_{15}^3 v_{25}^2 \\ 3v_{11}^2 v_{21}^3 - v_{11}^4 v_{21} & 3v_{12}^2 v_{22}^3 - v_{12}^4 v_{22} & 3v_{13}^2 v_{23}^3 - v_{13}^4 v_{23} & 3v_{14}^2 v_{24}^3 - v_{14}^4 v_{24} & 3v_{15}^2 v_{25}^3 - v_{15}^4 v_{25} \\ 5v_{11}^3 v_{21}^2 + v_{11}^5 & 5v_{12}^3 v_{22}^2 + v_{12}^5 & 5v_{13}^3 v_{23}^2 + v_{13}^5 & 5v_{14}^3 v_{24}^2 + v_{14}^5 & 5v_{15}^3 v_{25}^2 + v_{15}^5 \end{pmatrix},$$

and similarly for M_6, M_7, M_8 and M_9. We claim that the matrices M_r have rank $< r-1$ whenever $V \in \mathcal{F}_{r,2}$. Equivalently, the $(r-1)\times(r-1)$ minors of M_r lie in the ideal of $\mathcal{G}_{r,2}$. It suffices to consider the leftmost such minor since all minors are equivalent under permuting the columns of V. For each $r \leq 9$, we check that the determinant of that minor factors as

$$(m_{11}m_{22} - m_{12}m_{21}) \cdot \prod_{1\leq i<j\leq r-1} (v_{1i}v_{2j} - v_{2i}v_{1j}), \tag{4.2.24}$$

where the left factor is the polynomial of degree $2r-2$ given in (4.2.13). That polynomial vanishes on $\mathcal{G}_{r,2}$. This implies $\operatorname{rank}(M_r) \leq r-2$ on $\mathcal{G}_{r,2}$, and hence $\operatorname{rank}(\mathcal{M}_r) \leq r-2$ on $\mathcal{T}_{r,2,d}$.

Fix the lexicographic term order on $\mathbb{C}[t_0, t_1, \ldots, t_d]$. We can check that, for each $r \in \{3, 4, ..., 9\}$, the leading monomial of the leftmost maximal minor of $\mathcal{M}_r$ equals $t_0 t_2 t_4 \cdots t_{r-2}$. Hence all $\binom{d-r+1}{r-1}$ maximal minors of $\mathcal{M}_r$ are squarefree, and they generate the ideal

$$I_{r,d} \;=\; \big\langle\, t_{i_1} t_{i_2} t_{i_3} \cdots t_{i_{r-1}} \;:\; 2 \leq i_1{+}2 \leq i_2,\, i_2{+}2 \leq i_3,\, i_3{+}2 \leq i_4,\, \ldots,\, i_{r-2}{+}2 \leq i_{r-1} \leq d{-}2 \,\big\rangle.$$

This squarefree monomial ideal is pure of codimension $d-2r+3$ and it has degree $\binom{d-r+1}{r-2}$. This follows from [118, Theorem 1.6]. Indeed, in Murai's theory, our ideal $I_{r,d}$ is obtained from the power of the maximal ideal by applying the stable operator given by $a = (2, 4, 6, \ldots)$.

Combinatorial analysis reveals that the ideal $I_{r,d}$ is the intersection of the prime ideals

$$\big\langle t_{j_0}, t_{j_1}, t_{j_2}, t_{j_3}, \ldots, t_{j_{d-2r+2}} \big\rangle,$$

where $j_0, j_2, j_4, \ldots$ are even, $j_1, j_3, j_5, \ldots$ are odd, and $0 \leq j_0 < j_1 < j_2 < \cdots < j_{d-2r+2} \leq d$. Note that number of such sequences is $\binom{d-r+1}{d-2r+3} = \binom{d-r+1}{r-2}$. Hence the codimension and degree of $I_{r,d}$ are as expected for the ideal of maximal minors of an $(r{-}1) \times (d{-}r{+}1)$-matrix with linear entries [84, Ex. 19.10]. The monomial ideal $I_{r,d}$ is Cohen-Macaulay because its corresponding simplicial complex is shellable (cf. [142, §III.2]). Indeed, if we list the associated primes in a dictionary order for all sequences $j_0 j_1 j_2 \cdots j_{d-2r+2}$ as above, then this gives a shelling order.

Using Buchberger's S-pair criterion, we check that the maximal minors of $\mathcal{M}_r$ form a Gröbner basis. We only need to consider pairs of minors whose leading terms share variables. Up to symmetry, there are only few such pairs, so this is an easy check for each fixed $r \leq 9$.

Since $I_{r,d}$ is radical of codimension $d-2r+3$, we conclude that the ideal of maximal minors of $\mathcal{M}_r$ is radical and has the same codimension. However, that ideal of minors is contained in the prime ideal of $\mathcal{T}_{r,2,d}$, which has codimension $d-2r+3$ by Lemma 4.2.14.

Therefore, we now know that $\mathcal{T}_{r,2,d}$ is one of the irreducible components of the variety of maximal minors of $\mathcal{M}_r$. To conclude the proof we need to show that the latter variety is irreducible, so they are equal. To see this, we fix r and we proceed by induction on d. For $d = 2r-2$, when $\mathcal{M}_r$ is a square matrix, this can be checked directly. To pass from d to $d+1$, we factor the matrix as M_r times the rank r Hankel matrix associated with a funtf V. Increasing the value of d to $d+1$ multiplies the i-th row of the Hankel matrix by v_{i1} and it adds one more column. This gives us the value for the new variable t_{d+1}. Now, since that variable occurs linearly in the maximal minors, its value is unique. This implies that the unique rank $r-2$ extension from the old to the new $\mathcal{M}_r$ must come from the funtf V. □

We established Theorem 4.2.13 assuming that $r \leq 9$, but we believe that it holds for all r:

Conjecture 4.2.15. *For all $r \geq 3$ there exists a matrix $\mathcal{M}_r$ which satisfies the properties (a),(b) and (c) in Theorem 4.2.13. When r is odd, the matrix $\mathcal{M}_r$ may be chosen as*

$$\begin{pmatrix} r-4 & 0 & r & & & & & & \\ & r-6 & 0 & r-2 & & & & & \\ & & & \ddots & & & & & \\ & & & -1 & 0 & 3 & & & \\ & & & & 3 & 0 & -1 & & \\ & & & & & 5 & 0 & 1 & \\ & & & & & & & \ddots & \\ & & & & & & & r & 0 & r-4 \end{pmatrix} \begin{pmatrix} t_0 & t_1 & \cdots & t_{d-r} \\ t_1 & t_2 & \cdots & t_{d-r+1} \\ \vdots & \vdots & \ddots & \vdots \\ t_{r-2} & t_{r-1} & \cdots & t_{d-2} \\ t_{r-1} & t_r & \cdots & t_{d-1} \\ t_r & t_{r+1} & \cdots & t_d \end{pmatrix}.$$

We do not know yet what the general formula for $\mathcal{M}_r$ should be when r is even. The following systematic construction led to the matrices M_r and $\mathcal{M}_r$ in all cases known to us. Let $\widetilde{P}$ and $\widetilde{Q}$ be the multilinear forms in Lemma 4.2.4 that define $\mathcal{G}_{r,2}$. Let F_j denote the polynomial of degree $2r-2$ obtained by eliminating v_{1j} and v_{2j} from $\widetilde{P}$ and $\widetilde{Q}$. Let G_j denote the product of all $\binom{r-1}{2}$ minors $v_{1k}v_{2l} - v_{1l}v_{2k}$ of V where $j \notin \{k,l\}$. Each product F_jG_j is a polynomial of degree $r(r-1)$. Note that F_r is $m_{11}m_{22} - m_{12}m_{21}$ in (4.2.13), and F_rG_r is (4.2.24). Now, the ideal $\langle F_1G_1, F_2G_2, \ldots, F_rG_r\rangle$ is Cohen-Macaulay of codimension 2. By the Hilbert-Burch Theorem, the F_jG_j are the maximal minors of an $(r-1)\times r$-matrix M_r, which can be extracted from the minimal free resolution of $\langle F_1G_1, \ldots, F_rG_r\rangle$. This is precisely our matrix. In order to extend Theorem 4.2.13, and to find the desired $\mathcal{M}_r$ for even r, we need that all entries of the Hilbert-Burch matrix M_r have the same degree r.

Remark 4.2.16. *The singular locus of $\mathcal{T}_{r,2,d}$ is defined by the $(r-2)\times(r-2)$-minors of $\mathcal{M}_r$. It would be interesting to study this subvariety of $\mathbb{P}^d$ and how it relates to singularities of $\mathcal{F}_{r,2}$. For instance, for $r=4$, this singular locus is precisely the odeco variety $\mathcal{T}_{2,2,d}$, and, using Theorem 3.1.20, we can see that its prime ideal is generated by the 2×2-minors of $\mathcal{M}_4$.*

In Subsection 4.2.5 we shall see how the matrices $\mathcal{M}_r$ can be used to find a frame decomposition of a given symmetric $2\times 2\times\cdots\times 2$-tensor T. We close with an example that shows how this task differs from the easier problem of constructing a rank r Waring decomposition of T.

Example 4.2.17. *Let $r=4$, $d=8$, and consider the sum of two odeco tensors*

$$T \;=\; x^8+y^8 \,+\, (x-y)^8+(x+y)^8 \;=\; 3x^8+56x^6y^2+140x^4y^4+56x^2y^6+3y^8.$$

The coordinates of this tensor are $t_0=t_8=3$, $t_2=t_4=t_6=2$, and $t_1=t_3=t_5=t_7=0$. Here, the 3×5-matrix $\mathcal{M}_4$ has rank 2. This verifies that T lies in $\mathcal{T}_{4,2,8}$, in accordance with Example 4.2.3. However, the 4×4-matrix $\mathcal{M}_5$ is invertible. This means that T does not lie in $\mathcal{T}_{5,2,8}$. In other words, there is no funtf among the rank 5 Waring decompositions of T. $\diamondsuit$

4.2.4 Ternary Forms and Beyond

We now move on to higher dimensions $n \geq 3$. Our object of study is the fradeco variety

$$\mathcal{T}_{r,n,d} \subset \mathbb{P}(\mathrm{Sym}_d(\mathbb{C}^n)).$$

A very first question is: What is the dimension of $\mathcal{T}_{r,n,d}$? In Lemma 4.2.14, we saw that $\dim(\mathcal{T}_{r,2,d}) = 2r-3$. The following proposition generalizes that formula to arbitrary n:

Proposition 4.2.18. *For all $r > n$ and $d \geq 3$, the dimension of $\mathcal{T}_{r,n,d}$ is bounded above by*

$$min\left\{(n-1)(r-n) + \frac{(n-1)(n-2)}{2} + r - 1\,,\ \binom{n+d-1}{d} - 1\right\}. \tag{4.2.25}$$

Proof. The right number is the dimension of the ambient space, so this is an upper bound. The left number is the dimension of $\mathcal{F}_{r,n} \times \mathbb{P}^{r-1}$, by the formula in Theorem 4.2.2. The formula (4.2.3) expresses our variety as the (closure of the) image of a polynomial map

$$\mathcal{F}_{r,n} \times \mathbb{P}^{r-1} \longrightarrow \mathcal{T}_{r,n,d}. \tag{4.2.26}$$

The dimension of the image of this map is bounded above by the dimension of the domain. □

Remark 4.2.19. *When $T_{r,n,d}$ is not the ambient space,* (4.2.25) *is the same as* $\dim \mathcal{P}_{r,n}$.

We conjecture that the true dimension always agrees with the expected dimension:

Conjecture 4.2.20. *The dimension of the variety $\mathcal{T}_{r,n,d}$ is equal to (4.2.25) for all $r > n$ and $d \geq 3$.*

This conjecture is subtler than it may seem. Let $\sigma_r\nu_d\mathbb{P}^{n-1}$ denote the Zariski closure of the set of tensors of rank $\leq r$ in $\mathbb{P}(\mathrm{Sym}_d(\mathbb{C}^n))$. Geometrically, this is the r-th secant variety of the d-th Veronese embedding of $\mathbb{P}^{n-1}$. It is known that $\sigma_r\nu_d\mathbb{P}^{n-1}$ has the expected dimension in almost all cases. The Alexander-Hirschowitz Theorem (cf. [27, 108]) states that, assuming $d \geq 3$, the dimension of $\sigma_r\nu_d\mathbb{P}^{n-1}$ is lower than expected in precisely four cases:

$$(r,n,d) \ \in\ \big\{(5,3,4),\,(7,5,3),\,(9,4,4),\,(14,5,4)\big\}. \tag{4.2.27}$$

One might think that in these cases also the fradeco subvariety $\mathcal{T}_{r,n,d}$ has lower than expected dimension. However, the results summarized in Theorem 4.2.21 suggest that this is not the case.

Theorem 4.2.21. *Consider the fradeco varieties $\mathcal{T}_{r,n,d}$ in the cases when $n \geq 3$ and $1 \leq dim(\mathcal{T}_{r,n,d}) \cdot codim(\mathcal{T}_{r,n,d}) \leq 100$. Table 4.2 gives their degrees and some defining polynomials. The last column shows the minimal generators of lowest possible degrees in the ideal of $\mathcal{T}_{r,n,d}$.*

variety	dim	codim	degree	known equations
$\mathcal{T}_{4,3,3}$	6	3	17	3 cubics, 6 quartics
$\mathcal{T}_{4,3,4}$	6	8	74	6 quadrics, 37 cubics
$\mathcal{T}_{4,3,5}$	6	14	191	27 quadrics, 104 cubics
$\mathcal{T}_{5,3,4}$	9	5	210	1 cubic, 6 quartics
$\mathcal{T}_{5,3,5}$	9	11	1479	20 cubics, 213 quartics
$\mathcal{T}_{6,3,4}$	12	2	99	none in degree ≤ 5
$\mathcal{T}_{6,3,5}$	12	8	4269	one quartic
$\mathcal{T}_{7,3,5}$	15	5	≥ 38541	none in degree ≤ 4
$\mathcal{T}_{8,3,5}$	18	2	690	none in degree ≤ 5
$\mathcal{T}_{10,3,6}$	24	3	≥ 16252	none in degree ≤ 7
$\mathcal{T}_{5,4,3}$	10	9	830	none in degree ≤ 4
$\mathcal{T}_{6,4,3}$	14	5	1860	none in degree ≤ 3
$\mathcal{T}_{7,4,3}$	18	1	194	one in degree 194

Table 4.2: A census of small fradeco varieties

Computational Proof. The dimensions are consistent with Conjecture 4.2.20. They were verified by computing tangent spaces at a generic point using `Bertini` and `Matlab`. The degrees were computed with the monodromy loop method described in Subsubsection 4.2.5.6. The numerical Hilbert function method in Subsubsection 4.2.5.7 was used to determine how many polynomials of a given degree vanish on $\mathcal{T}_{r,n,d}$. This was followed up with computations in exact arithmetic in `Maple` and `Macaulay2`. These confirmed the earlier numerical results, and they enabled us to find the explicit polynomials in $\mathbb{Q}[T]$ that are listed in Examples 4.2.22, 4.2.23 and 4.2.24. In the cases where we report no equations occurring below a certain degree, this is a combination of Corollary 4.2.28 and the numerical Hilbert function computation. □

We shall now discuss some of the cases appearing in Theorem 4.2.21 in more detail.

Example 4.2.22. *The* 6*-dimensional variety* $\mathcal{T}_{4,3,3} \subset \mathbb{P}^9$ *has the parametrization*

$$\begin{aligned}
t_{300} &= v_{11}^3 + v_{12}^3 + v_{13}^3 + v_{14}^3, \\
t_{030} &= v_{21}^3 + v_{22}^3 + v_{23}^3 + v_{24}^3, \\
t_{003} &= v_{31}^3 + v_{32}^3 + v_{33}^3 + v_{34}^3, \\
t_{012} &= v_{21}v_{31}^2 + v_{22}v_{32}^2 + v_{23}v_{33}^2 + v_{24}v_{34}^2, \\
t_{021} &= v_{21}^2v_{31} + v_{22}^2v_{32} + v_{23}^2v_{33} + v_{24}^2v_{34}, \\
t_{102} &= v_{11}v_{31}^2 + v_{12}v_{32}^2 + v_{13}v_{33}^2 + v_{14}v_{34}^2, \\
t_{120} &= v_{11}v_{21}^2 + v_{12}v_{22}^2 + v_{13}v_{23}^2 + v_{14}v_{24}^2, \\
t_{201} &= v_{11}^2v_{31} + v_{12}^2v_{32} + v_{13}^2v_{33} + v_{14}^2v_{34}, \\
t_{210} &= v_{11}^2v_{21} + v_{12}^2v_{22} + v_{13}^2v_{23} + v_{14}^2v_{24}, \\
t_{111} &= v_{11}v_{21}v_{31} + v_{12}v_{22}v_{32} + v_{13}v_{23}v_{33} + v_{14}v_{24}v_{34}.
\end{aligned} \tag{4.2.28}$$

Here the matrix $V = (v_{ij})$ is given by the parametrization of $\mathcal{G}_{4,3}$ seen in (4.2.8) of Example 4.2.1.

Using exact linear algebra in `Maple`*, we find that the ideal of $\mathcal{T}_{4,3,3}$ contains no quadrics, but it contains three linearly independent cubics and 36 quartics. One of the cubics is*

$$C_{123} + 2C_{145} + 2C_{345} - C_{126} - C_{236} - 4C_{456}, \tag{4.2.29}$$

where C_{ijk} denotes the determinant of the 3×3 submatrix with columns i, j, k in

$$C = \begin{pmatrix} t_{300} & t_{210} & t_{120} & t_{201} & t_{111} & t_{102} \\ t_{210} & t_{120} & t_{030} & t_{111} & t_{021} & t_{012} \\ t_{201} & t_{111} & t_{021} & t_{102} & t_{012} & t_{003} \end{pmatrix}.$$

The other two cubics are obtained from this one by permuting the indices. The resulting three cubics define a complete intersection in $\mathbb{P}^9$. However, that complete intersection strictly contains $\mathcal{T}_{4,3,3}$ because the three cubics have only 30 multiples in degree 4, whereas we know that 36 quartics vanish on $\mathcal{T}_{4,3,3}$. Using `Macaulay2`*, we identified six minimal ideal generators in degree 4, and we found that the nine known generators generate a Cohen-Macaulay ideal of codimension 3 and degree 17. Using* `Bertini`*, we independently verified that fradeco variety $\mathcal{T}_{4,3,3}$ has degree 17. This implies that we have found the correct prime ideal.* ♢

Example 4.2.23. *The variety $\mathcal{T}_{4,3,4}$ is also 6-dimensional, and it lives in the $\mathbb{P}^{14}$ of ternary quartics. The parametrization is as in (4.2.28) but with quartic monomials instead of cubic. Among the ideal generators for $\mathcal{T}_{4,3,4}$ are six quadrics and 37 cubics. One of the quadrics is*

$$\begin{aligned} &8(t_{013}^2 - t_{004}t_{022}) + 8(t_{031}^2 - t_{022}t_{040}) + 8(t_{211}^2 - t_{202}t_{220}) + 18(t_{112}^2 - t_{103}t_{121}) + 18(t_{121}^2 - t_{112}t_{130}) \\ &+(t_{004}t_{040} + 19t_{022}^2 - 20t_{013}t_{031}) + (t_{004}t_{220} + t_{022}t_{202} - 2t_{013}t_{211}) + (t_{040}t_{202} + t_{022}t_{220} - 2t_{031}t_{211}). \end{aligned}$$

A `Bertini` *computation suggests that the known generators suffice to cut out $\mathcal{T}_{4,3,4}$. We also note that the 27 quadrics for $\mathcal{T}_{4,3,5}$ come from the 6 quadrics for $\mathcal{T}_{4,3,4}$. For instance, replacing each variable t_{ijk} by $t_{i,j,k+1}$ yields the quadric $8t_{014}^2 + 8t_{032}^2 + \cdots + 19t_{023}^2$ for $\mathcal{T}_{4,3,5}$.* ♢

Example 4.2.24. *The fradeco variety $\mathcal{T}_{5,3,4}$ is especially interesting because $(5, 3, 4)$ appears on the Alexander-Hirschowitz list (4.2.27). The unique cubic that vanishes on $\mathcal{T}_{5,3,4}$ is*

$$\begin{aligned} &46t_{022}t_{202}t_{220} + 73t_{112}t_{121}t_{211} - 4t_{004}t_{040}t_{400} + 19[t_{013}t_{130}t_{301}]_2 - 50[t_{004}t_{112}^2]_3 - 22[t_{004}t_{220}^2]_3 \\ &-18[t_{022}t_{211}^2]_3 + 50[t_{004}t_{022}t_{202}]_3 + 26[t_{004}t_{130}t_{310}]_3 + 100[t_{013}t_{103}t_{112}]_3 - 53[t_{013}t_{121}t_{310}]_3 \\ &+5[t_{004}t_{022}t_{400}]_6 - 50[t_{013}^2t_{202}]_6 - 5[t_{013}^2t_{220}]_6 + 45[t_{004}t_{031}t_{211}]_6 - 40[t_{022}t_{202}^2]_6 + 5[t_{004}t_{022}t_{220}]_6 \\ &+40[t_{022}t_{112}^2]_6 - 5[t_{004}t_{130}^2]_6 - 45[t_{004}t_{121}^2]_6 - 10[t_{004}t_{112}t_{130}]_6 - 45[t_{013}t_{022}t_{211}]_6 + 35[t_{013}t_{031}t_{202}]_6 \\ &+10[t_{013}t_{103}t_{130}]_6 + 10[t_{013}t_{112}t_{121}]_6 - 80[t_{013}t_{112}t_{301}]_6 + 80[t_{013}t_{202}t_{211}]_6 + 8[t_{013}t_{211}t_{220}]_6. \end{aligned}$$

This polynomial has 128 terms: each bracket denotes an orbit of monomials under the S_3-action, and the subscript is the orbit size. In addition, six fairly large quartics vanish on $\mathcal{T}_{5,3,4}$. The seven known generators cut out a reducible variety of dimension 9 in $\mathbb{P}^{14}$. The fradeco variety $\mathcal{T}_{5,3,4}$ is the unique top-dimensional component. But, using `Bertini`*, we found two extraneous components of dimension 7. Their degrees are 120 and 352 respectively.* ♢

We close this subsection by examining the geometric interplay between fradeco varieties and secant varieties. We write $\sigma_r\nu_d\mathbb{P}^{n-1}$ for the r-th secant variety of the d-th Veronese embedding of $\mathbb{P}^{n-1}$. This lives in $\mathbb{P}(\mathrm{Sym}_d(\mathbb{C}^n))$ and comprises rank r symmetric tensors (4.2.3). The same ambient space contains the fradeco variety $\mathcal{T}_{r,n,d}$ and all its secant varieties $\sigma_s\mathcal{T}_{r,n,d}$.

Theorem 4.2.25. *For any $r > n \geq d \geq 2$, we have*

$$\sigma_{r-n}\nu_d\mathbb{P}^{n-1} \;\subset\; \mathcal{T}_{r,n,d} \;\subset\; \sigma_r\nu_d\mathbb{P}^{n-1}, \tag{4.2.30}$$

and hence $\mathcal{T}_{r-n,n,d} \subset \mathcal{T}_{r,n,d}$ whenever $r \geq 2n$. Also, if $r = r_1r_2$ with $r_1 \geq 2$ and $r_2 \geq n$, then

$$\sigma_{r_1}\mathcal{T}_{r_2,n,d} \;\subseteq\; \mathcal{T}_{r,n,d}. \tag{4.2.31}$$

Proof. We fix d. The right inclusion in (4.2.30) is immediate from the definition. For the left inclusion we use the parametrization of $\mathcal{F}_{r,n}$ given in (4.2.15) and (4.2.16). The point is that the $(r-n)\times n$-matrix W can be chosen freely. Equivalently, the projection of $\mathcal{G}_{r,n} \subset (\mathbb{P}^{n-1})^r$ to any coordinate subspace $(\mathbb{P}^{n-1})^{r-n}$ is dominant. This means that the first $r-n$ summands in (4.2.1) are arbitrary powers of linear forms, and this establishes the left inclusion in (4.2.30).

To show the inclusion (4.2.31), we consider arbitrary frames $V_1, V_2, \ldots, V_{r_1} \in \mathcal{F}_{r_2,n}$. Then the $n \times r$-matrix $V = (V_1, V_2, \ldots, V_{r_1})$ is a frame in $\mathcal{F}_{r,n}$. Each V_i together with a choice of $\lambda_i \in \mathbb{R}^{r_2}$ determines a point on $\mathcal{T}_{r_2,n,d}$. Thus we have r_1 points in $\mathcal{T}_{r_2,n,d}$, and any point on the $\mathbb{P}^{r_1-1}$ spanned by these lies in $\mathcal{T}_{r,n,d}$, where it is represented by V with $\lambda = (\lambda_1, \ldots, \lambda_{r_1}) \in \mathbb{R}^r$. □

Example 4.2.26. *Let $n = 2$ and write $H = (t_{i+j})$ for a Hankel matrix of unknowns with $r+1$ rows and sufficiently many columns. The secant variety $\sigma_r\nu_d\mathbb{P}^1$ is defined by the ideal $I_{r+1}(H)$ of $(r+1)\times(r+1)$-minors of H. The ideal-theoretic version of (4.2.30) states that*

$$I_{r-1}(H) \;\supset\; I_{r-1}(\mathcal{M}_r) \;\supset\; I_{r+1}(H).$$

It is instructive to check this. The left inclusion follows from the Cauchy-Binet Theorem applied to $\mathcal{M}_r = A \cdot H$ where A is the $(r-1)\times(r+1)$ integer matrix underlying M_r.

To see more precisely how the matrix A is defined, consider the formula in Conjecture 4.2.15. Rewriting it, we obtain the following matrix product where the matrix on the left is A:

$$\begin{pmatrix}
r-4 & 0 & r & & & & & & \\
 & r-6 & 0 & r-2 & & & & & \\
 & & & \ddots & & & & & \\
 & & & -1 & 0 & 3 & & & \\
 & & & & 3 & 0 & -1 & & \\
 & & & & & 5 & 0 & 1 & \\
 & & & & & & & \ddots & \\
 & & & & & & & r & 0 & r-4
\end{pmatrix}
\begin{pmatrix}
t_0 & t_1 & \cdots & t_{d-r} \\
t_1 & t_2 & \cdots & t_{d-r+1} \\
\vdots & \vdots & \ddots & \vdots \\
t_{r-2} & t_{r-1} & \cdots & t_{d-2} \\
t_{r-1} & t_r & \cdots & t_{d-1} \\
t_r & t_{r+1} & \cdots & t_d
\end{pmatrix}$$

Remark 4.2.27. *(a) Since concatenations of frames in $\mathbb{R}^n$ are always frames, (4.2.31) generalizes from secant varieties to joins. Namely, if $r = r_1 + r_2$, then $\mathcal{T}_{r_1,n,d} \star \mathcal{T}_{r_2,n,d} \subset \mathcal{T}_{r,n,d}$.*

(b) The inclusion in (4.2.31) is always strict, with one notable exception: $\sigma_2\mathcal{T}_{2,2,d} = \mathcal{T}_{4,2,d}$.

Theorem 4.2.25 implies that the Veronese variety $\nu_d\mathbb{P}^{n-1}$ is contained in the fradeco variety $\mathcal{T}_{r,n,d}$ with $r > n$. This is illustrated in Example 4.2.23 where we wrote the quadric that vanishes on $\mathcal{T}_{4,3,4}$ as a linear combination of the binomials that define $\nu_4\mathbb{P}^2 \subset \mathbb{P}^{14}$. The formula (4.2.29) shows that this cubic vanishes on $\sigma_2\nu_3\mathbb{P}^2$. Similarly, we can verify that the cubic in Example 4.2.24 vanishes on $\sigma_2\nu_4\mathbb{P}^2$ by writing it as a linear combination of the 3×3-minors $C_{ijk,lmm}$ of the 6×6-catalecticant C matrix in (4.2.34). One such expression is

$$\begin{gathered}50C_{012,012} - 30C_{012,123} + 50C_{012,034} - 30C_{012,125} + 50C_{012,045} + 63C_{012,345} - 10C_{013,024} + 10C_{013,234} \\ +5C_{013,015} + 35C_{013,135} + 34C_{013,245} + 5C_{023,023} - 80C_{023,134} + 5C_{023,025} - 26C_{023,235} - 19C_{023,145} \\ -30C_{123,123} + 29C_{123,125} - 10C_{123,345} - 10C_{014,025} + 19C_{014,235} - 53C_{014,145} - 30C_{024,245} + 5C_{034,034} \\ +26C_{034,045} + 5C_{034,345} + 50C_{134,134} + 50C_{134,235} + 30C_{134,145} + 30C_{234,245} + 5C_{015,015} + 26C_{015,135} \\ +50C_{015,245} - 5C_{025,235} - 10C_{025,145} - 10C_{125,345} - 4C_{035,035} + 5C_{135,135} + 50C_{135,245} + 5C_{235,235} \\ +5C_{045,045} + 5C_{045,345} + 50C_{245,245}.\end{gathered}$$

Theorem 4.2.25 gives lower bounds on the degrees of the equations defining fradeco varieties:

Corollary 4.2.28. *All non-zero polynomials in the ideal of $\mathcal{T}_{r,n,d}$ must have degree at least $r - n + 1$.*

Proof. The ideal of the Veronese variety $\nu_d\mathbb{P}^{n-1}$ contains no linear forms. It is generated by 2×2 minors of catalecticants. A general result on secant varieties [139, Thm. 1.2] implies that the ideal of $\sigma_{r-n}\nu_d\mathbb{P}^{n-1}$ is zero in degree $\leq r - n$. The inclusion $\sigma_{r-n}\nu_d\mathbb{P}^{n-1} \subset \mathcal{T}_{r,n,d}$ yields the claim. □

In Table 4.2 we see that $\mathcal{T}_{4,3,4}$, $\mathcal{T}_{4,3,5}$, $\mathcal{T}_{5,3,4}$, $\mathcal{T}_{5,3,5}$ and $\mathcal{T}_{6,3,5}$ have their first minimal generators in the lowest possible degrees. However this is not always the case, as shown dramatically by $\mathcal{T}_{7,4,3}$.

4.2.5 Numerical Recipes

Methods from Numerical Algebraic Geometry (NAG) are useful for studying the decomposition of tensors into frames. Many of the results on fradeco varieties $\mathcal{T}_{r,n,d}$ reported in subsections 4.2.3 and 4.2.4 were discovered using NAG. In this subsection we discuss the relevant methodologies. Our experiments involve a mixture of using `Bertini` [15], `Macaulay2` [81], `Maple`, and `Matlab`.

All algebraic varieties have an *implicit representation*, as the solution set to a system of polynomial equations. Some special varieties admit a *parametric representation*, as the (closure of the) image of a map whose coordinates are rational functions. Having to pass back and forth between these two representations is a ubiquitous task in computational algebra.

The fradeco variety studied in this section is given by a mixture of implicit and parametric. Our point of departure is the implicit representation (4.2.4) of the funtf variety $\mathcal{F}_{r,n}$, or its homogenization $\mathcal{G}_{r,n}$. Built on top of that is the parametrization (4.2.1) of rank r tensors:

$$\begin{array}{ccc} \mathbb{C}^{n\times r}\times\mathbb{C}^r & & \mathrm{Sym}_d(\mathbb{C}^n) \\ \cup & & \cup \\ \mathcal{F}_{r,n}\times\mathbb{C}^r & \xrightarrow{\Sigma_d} & \widehat{\mathcal{T}}_{r,n,d} \\ (V,\lambda) & \longmapsto & \lambda_1 v_1^{\otimes d}+\lambda_2 v_2^{\otimes d}+\cdots+\lambda_r v_r^{\otimes d} \end{array} \tag{4.2.32}$$

Here, $\widehat{\mathcal{T}}_{r,n,d}$ denotes the affine cone over the projective variety $\mathcal{T}_{r,n,d}$. The input to our *decomposition problem* is an arbitrary symmetric $n\times n\times\cdots\times n$-tensor T and a positive integer r. The task is to decide whether T lies in $\widehat{\mathcal{T}}_{r,n,d}$, and, if yes, to compute a preimage (V,λ) under the map Σ_d in (4.2.32). Any preimage must satisfy the non-trivial constraint $V\in\mathcal{F}_{r,n}$.

4.2.5.1 Decomposing fradeco tensors

We discuss three approaches to finding frame decompositions of symmetric tensors.

4.2.5.2 Tensor power method

Our original motivation for this project came from the case $r=n$ of odeco tensors (see Section 3.1). If $T\in\widehat{\mathcal{T}}_{n,n,d}$, then the *tensor power method* of [8] reliably reconstructs the decomposition (4.2.1) where $\{v_1,\ldots,v_n\}$ is an orthonormal basis of $\mathbb{R}^n$. The algorithm is to iterate the rational map $\nabla T:\mathbb{P}^{n-1}\dashrightarrow\mathbb{P}^{n-1}$ given by the gradient vector $\nabla T=(\partial T/\partial x_1,\ldots,\partial T/\partial x_n)$. This map is regular when the hypersurface $\{T=0\}$ is smooth. The fixed points of ∇T are the *eigenvectors* of the tensor T. Their number was given in [37]. The punchline is this: if the multipliers $\lambda_1,\ldots,\lambda_n$ in (4.2.1) are positive, then $v_1,\ldots,v_n$ are precisely the *robust eigenvectors*, i.e. the attracting fixed points of the gradient map ∇T.

This raises the question whether the tensor power method also works for fradeco tensors. The answer is "no" in general, but it is "yes" in some special cases.

Example 4.2.29. *Let $n=2, r=4, d=5$ and consider the fradeco quintic*

$$T \;=\; \alpha x^5+y^5+(x+y)^5+(x-y)^5 \quad\in\; \mathcal{T}_{4,2,5},$$

where $\alpha>6$ is a parameter. The eigenvectors of T are the zeros in $\mathbb{P}^1$ of the binary quintic

$$y\frac{\partial T}{\partial x}-x\frac{\partial T}{\partial y} \;=\; 5y\cdot\left((\alpha x-6)x^4+\left(2xy-\frac{1}{4}y^2\right)^2+\frac{31}{16}y^4\right).$$

The point $(1:0)$ is an eigenvector, but there are no other real eigenvectors, as the expression is a sum of squares. Hence the frame decomposition of T cannot be recovered from its eigenvectors. ♢

Example 4.2.30. *For any reals $\lambda_1, \lambda_2, \lambda_3, \lambda_4 > 0$ and any integer $d \geq 5$, we consider the tensor*

$$T \,=\, \lambda_1(-5x_1+x_2+x_3)^d+\lambda_2(x_1-5x_2+x_3)^d+\lambda_3(x_1+x_2-5x_3)^d+\lambda_4(3x_1+3x_2+3x_3)^d. \quad (4.2.33)$$

This tensor has precisely four robust eigenvectors, namely the columns of the matrix V in (4.2.7). Hence the frame decomposition of T can be recovered by the tensor power method. ♢

The following conjecture generalizes this example.

Conjecture 4.2.31. *Let $r = n + 1 < d$ and $T \in \mathcal{T}_{n+1,n,d}$ with $\lambda_1, \ldots, \lambda_{n+1} > 0$ in (4.2.1). Then $v_1, \ldots, v_{n+1}$ are the robust eigenvectors of T, so they are found by the tensor power method.*

Example 4.2.29 shows that Conjecture 4.2.31 is false for $r \geq n+2$, and it suggests that the Tensor Power Method will not work in general. We next discuss two alternative approaches.

4.2.5.3 Catalecticant method for frames

The matrices in Theorem 4.2.13 furnish a practical algorithm for the frame decomposition problem when $n = 2$. This is a variant of Sylvester's Catalecticant Algorithm, and it works as follows.

Our input is a binary form $T \in \widehat{\mathcal{T}}_{r,2,n}$. We seek to recover the tight frame into which T decomposes. Since we do not know the fradeco rank r in advance, we start with $\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$, etc. and plug in the coordinates t_i of T. The fradeco rank is the first index r with $\mathcal{M}_r$ rank deficient.

If the matrix $\mathcal{M}_r$ is rank deficient, then its rank is at most $r-2$. Let us assume that the rank equals exactly $r-2$. Otherwise T is a singular point (cf. Remark 4.2.16). Then, up to scaling, we find the unique row vector $w \in \mathbb{R}^{r-1}$ in the left kernel of $\mathcal{M}_r$. By Theorem 4.2.13 we know that $\mathcal{M}_r$ is the product of the matrix M_r and an $(r-1) \times (d-r-1)$ matrix with entries $v_{i1}^{d-r-j+1} v_{i2}^{j-1}$, where $V = (v_{ij}) \in \mathcal{G}_{r,2}$ is the desired frame. Moreover, the matrix M_r has rank $r-2$, so the vector w also lies in the left kernel of M_r, i.e. $w \cdot M_r = 0$. Thus,

$$0 = w \cdot M_r \;=\; \big(f(v_{11}, v_{21}), f(v_{12}, v_{22}), \ldots, f(v_{1r}, v_{2r})\big),$$

where $f(x, y)$ is a binary form of degree r. The r roots of $f(x, y)$ in $\mathbb{P}^1$ are the columns of the desired $V = (v_{ij}) \in \mathcal{G}_{r,2}$. Using these v_{ij}, the given binary form has the decomposition

$$T(x, y) \;=\; \sum_{j=1}^{r} \lambda_j (v_{1j}x + v_{2j}y)^d,$$

where the multipliers $\lambda_1, \ldots, \lambda_r$ are recovered by solving a linear system of equations.

Example 4.2.32. *Let $r = 5$ and $d = 8$. We illustrate this method for the binary octic*

$$\begin{aligned} T = {} & (-237-896\alpha)x^8 + 8(65+241\alpha)x^7y - 28(16+68\alpha)x^6y^2 + 56(5+31\alpha)x^5y^3 + 70(2-56\alpha)x^4y^4 \\ & +56(-7+193\alpha)x^3y^5 + 28(32-716\alpha)x^2y^6 + 8(-115+2671\alpha)xy^7 + (435-9968\alpha)y^8, \end{aligned}$$

where $\alpha = \sqrt{3}-2$. The parenthesized expressions are the coordinates $t_0, \ldots, t_8$. We find

$$\mathcal{M}_5 \;=\; \begin{pmatrix} -13548\alpha+595 & 3636\alpha-150 & -996\alpha+42 & 348\alpha+18 \\ 2092\alpha-94 & -548\alpha+26 & 100\alpha-22 & 148\alpha+50 \\ -2092\alpha+94 & 548\alpha-26 & -100\alpha+22 & -148\alpha-50 \\ 996\alpha-30 & -348\alpha-6 & 396\alpha+90 & -1236\alpha-317 \end{pmatrix}.$$

This matrix has rank 3 and its left kernel is the span of the vector $\mathbf{w} = (0,1,1,0)$. Therefore,

$$0 \,=\, \mathbf{w}M_5 \,=\, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}^T \begin{pmatrix} v_{21}^5+5v_{11}^5 & v_{22}^5+5v_{12}^5 & v_{23}^5+5v_{13}^5 & v_{24}^5+5v_{14}^5 & v_{25}^5+5v_{15}^5 \\ v_{11}v_{21}^4-3v_{11}^3v_{21}^2 & v_{12}v_{22}^4-3v_{12}^3v_{22}^2 & v_{13}v_{23}^4-3v_{13}^3v_{23}^2 & v_{14}v_{24}^4-3v_{14}^3v_{24}^2 & v_{15}v_{25}^4-3v_{15}^3v_{25}^2 \\ 3v_{11}^2v_{21}^3-v_{11}^4v_{21} & 3v_{12}^2v_{22}^3-v_{12}^4v_{22} & 3v_{13}^2v_{23}^3-v_{13}^4v_{23} & 3v_{14}^2v_{24}^3-v_{14}^4v_{24} & 3v_{15}^2v_{25}^3-v_{15}^4v_{25} \\ 5v_{11}^3v_{21}^2+v_{11}^5 & 5v_{12}^3v_{22}^2+v_{12}^5 & 5v_{13}^3v_{23}^2+v_{13}^5 & 5v_{14}^3v_{24}^2+v_{14}^5 & 5v_{15}^3v_{25}^2+v_{15}^5 \end{pmatrix}.$$

Hence the five columns of the desired tight frame $V = (v_{ij})$ are the distinct zeros in $\mathbb{P}^1$ of

$$f(v_{1i}, v_{2i}) \;=\; v_{1i}v_{2i}^4 - 3v_{1i}^3v_{2i}^2 + 3v_{1i}^2v_{2i}^3 - v_{1i}^4v_{2i} \qquad \text{for } i = 1, \ldots, 5.$$

We find

$$V \;=\; \begin{pmatrix} 1 & 0 & 1 & \alpha & 1 \\ 0 & 1 & 1 & 1 & \alpha \end{pmatrix} \;\in\; \mathcal{G}_{5,2}.$$

It remains to solve the linear system of nine equations in $\lambda = (\lambda_1, \ldots, \lambda_5)$ given by

$$T \;=\; \lambda_1 x^8 + \lambda_2 y^8 + \lambda_3 (x+y)^8 + \lambda_4 (\alpha x + y)^8 + \lambda_5 (x + \alpha y)^8.$$

The unique solution to this system is $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_5 = 1$ and $\lambda_4 = 1552 + 896\sqrt{3}$. ♢

4.2.5.4 Waring-enhanced frame decomposition

We now examine the decomposition problem for $n \geq 3$. Since no determinantal representation of $\mathcal{T}_{r,n,d}$ is known, a system of equations must be solved to recover (V, λ) from a given tensor in $\widehat{\mathcal{T}}_{r,n,d}$. In some special situations, we can approach this by taking advantage of known results on Waring decompositions. For instance, in Example 4.2.1 the Waring decomposition is already the frame decomposition. Example 4.2.17 shows that this is an exceptional situation.

We demonstrate the "Waring-enhanced" frame decomposition for the ternary quartic

$$\begin{aligned} \textstyle\sum_{i+j+k=4} \frac{24}{i!j!k!} t_{ijk} x^i y^j z^k \;=\; & 467x^4+152x^3y+1448x^3z+660x^2y^2-1488x^2yz+4020x^2z^2+536xy^3 \\ & -1992xy^2z+2352xyz^2+944xz^3+227y^4-1000y^3z+2148y^2z^2-1960yz^3+1267z^4. \end{aligned}$$

Ternary quartics of rank ≤ 5 form a hypersurface of degree 6 in $\mathbb{P}^{14}$. The equation of this hypersurface is the determinant of the 6×6 catalecticant matrix C. Here the dimension is

one less than expected; this is the first entry in the Alexander-Hirschowitz list (4.2.27). For the given quartic,

$$C = \begin{bmatrix} t_{400} & t_{310} & t_{301} & t_{220} & t_{211} & t_{202} \\ t_{310} & t_{220} & t_{211} & t_{130} & t_{121} & t_{112} \\ t_{301} & t_{211} & t_{202} & t_{121} & t_{112} & t_{103} \\ t_{220} & t_{130} & t_{121} & t_{040} & t_{031} & t_{022} \\ t_{211} & t_{121} & t_{112} & t_{031} & t_{022} & t_{013} \\ t_{202} & t_{112} & t_{103} & t_{022} & t_{013} & t_{004} \end{bmatrix} = \begin{bmatrix} 467 & 38 & 362 & 110 & -124 & 670 \\ 38 & 110 & -124 & 134 & -166 & 196 \\ 362 & -124 & 670 & -166 & 196 & 236 \\ 110 & 134 & -166 & 227 & -250 & 358 \\ -124 & -166 & 196 & -250 & 358 & -490 \\ 670 & 196 & 236 & 358 & -490 & 1267 \end{bmatrix}. \tag{4.2.34}$$

This matrix has rank 5 and its kernel is spanned by the vector corresponding to the quadric $q = 14u^2 - uv - 2uw - 4v^2 - 11vw - 10w^2$. The points $(u : v : w)$ in $\mathbb{P}^2$ that lie on the conic $\{q = 0\}$ represent all the linear forms $ux + vy + wz$ that may appear in a rank 5 decomposition.

Our task is to find five points on the conic $\{q = 0\}$ that form a frame $V \in \mathcal{G}_{5,3}$. This translates into solving a rather challenging system of polynomial equations. One of the solutions is

$$V = (v_1, v_2, v_3, v_4, v_5) = \begin{pmatrix} -1 & 2 & 2 & 1+2\sqrt{3} & -1+2\sqrt{3} \\ 2 & 2 & -1 & -2+\sqrt{3} & 2+\sqrt{3} \\ 0 & 1 & -2 & 5 & -5 \end{pmatrix}.$$

The given ternary quartic has the frame decomposition $v_1^{\otimes 4} + v_2^{\otimes 4} + v_3^{\otimes 4} + v_4^{\otimes 4} + v_5^{\otimes 4}$.

4.2.5.5 Exploring the fradeco variety

The following tasks make sense for any variety $X \subset \mathbb{P}^N$ arising in an applied context: (i) sample points on X, (ii) compute the dimension and degree of X, (iii) compute an irreducible decomposition of X, (iv) find a parametrization of X, (v) find some polynomials that vanish on X, (vi) determine polynomials that cut out X, (vii) find generators for the ideal of X. Numerical algebraic geometry (NAG) furnishes tools for addressing these points. In our study, X is the fradeco variety $\mathcal{T}_{r,n,d}$. We used NAG to find answers in some cases. In what follows, we explain our computations. Particular emphasis is placed on the results reported in subsection 4.2.4 for the degree and Hilbert function of $\mathcal{T}_{r,n,d}$. All computations are carried out by working on the affine cone $\widehat{\mathcal{T}}_{r,n,d} \subset \mathrm{Sym}_d(\mathbb{C}^n)$.

4.2.5.6 Dimension and degree

The dimension and degree of the affine variety $\widehat{\mathcal{T}}_{r,n,d}$ can be computed directly from the mixed parametric-implicit representation in (4.2.32). The dimension can be found by selecting a random point on $\mathcal{F}_{r,n} \times \mathbb{R}^r$, determining its tangent space via [145], and then taking the image of this tangent space via the derivative of the map Σ_d. The image is a linear subspace in $\mathrm{Sym}_d(\mathbb{R}^n)$, and its dimension is found via the rank of its defining matrix. These matrices are

usually given numerically, in terms of points sampled from $\mathcal{F}_{r,n}$, so we need to use singular value decompositions.

The computation of the degree is carried out using *monodromy*. We obtained the results of Theorem 4.2.21 by applying essentially the same technique as in [86, 88], adapted to our situation where the mapping is from an implicitly defined source. Here are some highlights of this method for $\widehat{\mathcal{T}}_{r,n,d}$. We performed these computations using `Bertini` and `MatLab`.

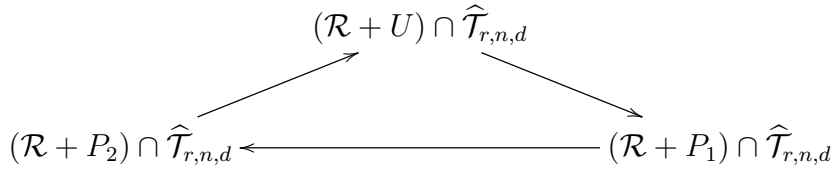
Let c denote the codimension of $\widehat{\mathcal{T}}_{r,n,d}$, as given by the formula in Conjecture 4.2.20. The degree of $\widehat{\mathcal{T}}_{r,n,d}$ is the number of points in the intersection with a random c-dimensional affine subspace of $\mathrm{Sym}_d(\mathbb{C}^n)$. Here we represent the fradeco variety purely numerically, namely as the set of images of points (V, λ) under the parametrization Σ_d shown in (4.2.32). This method verifies the dimension of $\widehat{\mathcal{T}}_{r,n,d}$ because the intersection would be empty if the dimension were lower than expected.

As a first step, we compute a numerical irreducible decomposition of the funtf variety $\mathcal{F}_{r,n}$. This also gives its degree and dimension, as shown in Table 4.1. In particular, we obtain degree-many points of $\mathcal{F}_{r,n}$ that lie in a random linear space of dimension equal to $\mathrm{codim}(\mathcal{F}_{r,n})$.

We take V to be one of these generic points in $\mathcal{F}_{r,n}$, we select a random vector $\lambda \in \mathbb{C}^r$, and we compute the fradeco tensor $\Sigma_d(V, \lambda)$. We also fix a random c-dimensional linear subspace $\mathcal{R}$ of $\mathrm{Sym}_d(\mathbb{C}^n)$ and a random point U in the c-dimensional affine space $\mathcal{R} + U$.

By construction, the affine cone $\widehat{\mathcal{T}}_{r,n,d}$ and the affine space $\mathcal{R} + U$ intersect in $\deg(\widehat{\mathcal{T}}_{r,n,d})$ many points in $\mathrm{Sym}_d(\mathbb{C}^n)$. One of these points is $\Sigma_d(V, \lambda)$. Our goal is to discover all the other intersection points by sequences of parameter homotopies that form monodromy loops. Geometrically, the base space for these monodromies is the vector space quotient $\mathrm{Sym}_d(\mathbb{C}^n)/R$.

We fix two further random points P_1 and P_2 in $\mathrm{Sym}_d(\mathbb{C}^n)$. These represent residue classes modulo the linear subspace $\mathcal{R}$. The data we fixed now define a (triangular) monodromy loop



We use `Bertini` to perform each linear parameter homotopy. This constructs a path (V_t, λ_t) in the parameter space. Here t runs from 0 to 3. We start at $(V_0, \lambda_0) = (V, \lambda)$, the point $\Sigma_d(V_i, \lambda_i)$ lies in $(\mathcal{R} + P_i) \cap \widehat{\mathcal{T}}_{r,n,d}$ for $i = 1, 2$, and $\Sigma_d(V_3, \lambda_3)$ is back in $(\mathcal{R}+U) \cap \widehat{\mathcal{T}}_{r,n,d}$. With high probability, $\Sigma_d(V_3, \lambda_3) \neq \Sigma_d(V, \lambda)$ holds, and we have discovered a new point. Then we iterate the process. Let $S_k := \{\Sigma_d(V, \lambda), \ldots, \Sigma_d(V', \lambda')\}$ denote the subset of $(\mathcal{R}+U) \cap \widehat{\mathcal{T}}_{r,n,d}$ that has been found after k steps. In the next monodromy loop we trace the paths of S_k to produce $\tilde{S}_{k+1}$, the endpoints of monodromy loops starting from S_k. Using `MatLab`, we then merge the point sets to form $S_{k+1} = S_k \cup \tilde{S}_{k+1}$. We repeat this process until no new points are found after 20 consecutive monodromy loops. The number of points in S_k is very strong

numerical evidence for the degree of $\mathcal{T}_{r,n,d}$. At this point, one can also use the trace test [141] with pseudowitness sets [87] to confirm that degree.

4.2.5.7 Numerical Hilbert Function

We wish to learn the polynomial equations that vanish on $\mathcal{T}_{r,n,d}$. The set I of all such polynomials is a homogeneous prime ideal in the polynomial ring over $\mathbb{Q}$ whose variables are the entries $t_{i_1 i_2 \cdots i_d}$ of an indeterminate tensor T. We write this polynomial ring as

$$\mathbb{Q}[T] \,=\, \bigoplus_{e\geq 0} \mathbb{Q}[T]_e \quad \simeq \quad \bigoplus_{e\geq 0} \mathrm{Sym}_e(\mathrm{Sym}_d(\mathbb{Q}^n)) \,=\, \mathrm{Sym}_*(\mathrm{Sym}_d(\mathbb{Q}^n)).$$

The space of all polynomials of degree e in the ideal I is the subspace

$$I_e \,=\, I \cap \mathbb{Q}[T]_e \quad \subset \quad \mathbb{Q}[T]_e \,\simeq\, \mathrm{Sym}_e(\mathrm{Sym}_d(\mathbb{Q}^n)).$$

A natural approach is to fix some small degree e and to ask for a $\mathbb{Q}$-linear basis of I_e.

The dimensions of these vector spaces are organized into the *Hilbert function*

$$\mathbb{N} \,\to\, \mathbb{N},\; e \,\mapsto\, \dim_{\mathbb{Q}}(I_e).$$

We used `Bertini` and `Matlab` to determine specific values of the Hilbert function. In some cases, an independent `Maple` computation was used to construct a basis for the $\mathbb{Q}$-vector space I_e.

Fix values for r, n, d. As discussed above, we can use the parametrization (4.2.32) to produce many sample points $T = \Sigma_d(V, \lambda)$ on $\mathcal{T}_{r,n,d}$. The condition $f(T) = 0$ translates into a linear equation in the coefficients of a given polynomial $f \in \mathbb{Q}[T]_e$, and I_e is the solution space to these equations as T ranges over $\mathcal{T}_{r,n,d}$. We write these linear equations as a matrix whose number of columns is $\dim(\mathbb{Q}[T]_e) = \binom{\binom{n+d-1}{d}+e-1}{e}$, and with one row per sample point T. In practice we take enough sample points so that I_e is sure to equal the kernel of that matrix.

This procedure may be carried out in exact arithmetic over $\mathbb{Q}$ when sufficiently many exact points can be found on $\mathcal{F}_{r,n}$. When floating point approximations are used, some care is required in choosing the appropriate number of points and a sufficient degree of precision. This numerical test can become inconclusive in high dimension due to these issues. Using floating point arithmetic and 30,000 points of $\mathcal{F}_{r,n}$ we obtained the values listed in Table 4.3. The blanks indicate that we did not find conclusive evidence for the exact value of $\dim(I_e)$ in that case. For $\mathcal{T}_{5,4,3}$, $\mathcal{T}_{6,3,4}$, $\mathcal{T}_{6,4,3}$, $\mathcal{T}_{7,3,5}$, and $\mathcal{T}_{8,3,5}$ we also found no conclusive numerical evidence for equations in degrees less than 5.

The calculation of $\dim(I_e)$ is a numerical rank computation via singular value decomposition, so at least in principle it is possible to also extract a basis of I_e. However, in practice, round-off errors yield imprecise values for the coefficients of the basis elements of I_e. This makes it difficult to reliably determine an exact $\mathbb{Q}$-basis of I_e by numerical methods.

ideal \ deg e	2	3	4	5	6
$\dim \mathcal{I}(\mathcal{T}_{5,2,9})_e$	0	0	5	46	235
$\dim \mathcal{I}(\mathcal{T}_{4,3,4})_e$	6	127	1093	5986	
$\dim \mathcal{I}(\mathcal{T}_{4,3,5})_e$	27	651	6370		
$\dim \mathcal{I}(\mathcal{T}_{5,3,4})_e$	0	1	21		
$\dim \mathcal{I}(\mathcal{T}_{5,3,5})_e$	0	20	633		
$\dim \mathcal{I}(\mathcal{T}_{6,3,5})_e$	0	0	1		

Table 4.3: Numerical computation of the Hilbert functions of fradeco varieties

To discover the explicit ideal generators displayed in Subsections 4.2.3 and 4.2.4, we instead used exact arithmetic in `Maple`. A key step was to produce points in the funtf variety $\mathcal{F}_{r,n}$ that are defined over low-degree extension of $\mathbb{Q}$, and to map them carefully via Σ_d. To accomplish this, we used the representation of $\mathcal{G}_{r,n}$ discussed in Subsection 4.2.2. In our experiments, we found that the `solve` command in `Maple` was able to handle dense linear systems with up to $3,500$ unknowns.

4.3 Conclusion

In the first section of this chapter we studied the varieties of orthogonally decomposable tensors in two different cases: symmetric and ordinary tensors. We showed that these varieties are defined set-theoretically by quadratic equations that arise from the associativity of a certain algebra defined by the given tensor. Motivated by these results and by the fact that odeco tensors constitute a very low-dimensional variety, in the second section of this chapter we extended the definition of odeco tensors to that of frame decomposable tensors. In small cases we described the variety of such tensors and we showed how one can find their frame decomposition.

Chapter 5

Superresolution without Separation

This chapter provides a theoretical analysis of diffraction-limited superresolution, demonstrating that arbitrarily close point sources can be resolved in ideal situations. Precisely, we assume that the incoming signal is a linear combination of M shifted copies of a known waveform with unknown shifts and amplitudes, and one only observes a finite collection of evaluations of this signal. We characterize properties of the base waveform such that the exact translations and amplitudes can be recovered from $2M+1$ observations. This recovery can be achieved by solving a weighted version of basis pursuit over a continuous dictionary. Our analysis shows that ℓ_1-based methods enjoy the same separation-free recovery guarantees as polynomial root finding techniques such as Prony's method or Vetterli's method for signals of finite rate of innovation. Our proof techniques combine classical polynomial interpolation techniques with contemporary tools from compressed sensing. This chapter is based on joint work with Geoffrey Schiebinger and Benjamin Recht titled *Superresolution without separation* [137].

5.1 Introduction

Imaging below the diffraction limit remains one of the most practically important yet theoretically challenging problems in signal processing. Recent advances in superresolution imaging techniques have made substantial progress towards overcoming these limits in practice [62, 119], but theoretical analysis of these powerful methods remains elusive. Building on polynomial interpolation techniques and tools from compressed sensing, this chapter provides a theoretical analysis of diffraction-limited superresolution, demonstrating that arbitrarily close point sources can be resolved in ideal situations.

We assume that the measured signal takes the form

$$x(s) = \sum_{i=1}^{M} c_i \psi(s, t_i), \tag{5.1.1}$$

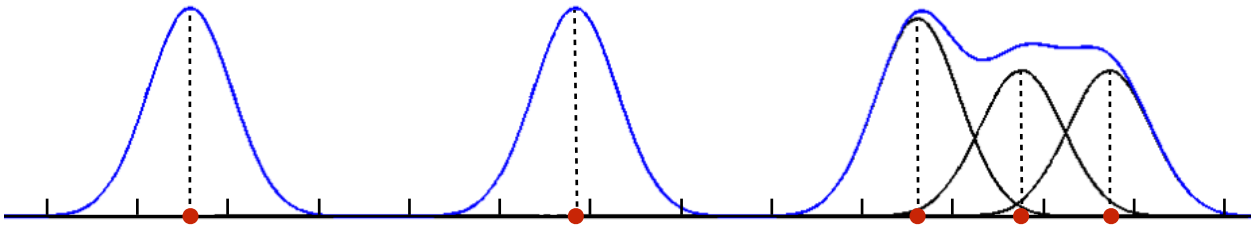


Figure 5.1: An illustrative example of (5.1.1) with the Gaussian point spread function $\psi(s,t) = e^{-(s-t)^2}$. The t_i are denoted by red dots, and the true intensities c_i are illustrated by vertical, dashed black lines. The super position resulting in the signal x is plotted in blue. The samples $\mathcal{S}$ would be observed at the tick marks on the horizontal axis.

Here $\psi(s,t)$ is a differentiable function that describes the image at spatial location s of a point source of light localized at t. The function ψ is called the *point spread function*, and we assume its particular form is known beforehand. In (5.1.1), $t_1, \ldots, t_M$ are the locations of the point sources and $c_1, ..., c_M > 0$ are their intensities. Throughout we assume that these quantities together with the number of point sources M, are fixed but unknown. The primary goal of superresolution is to recover the locations and intensities from a set of noiseless observations

$$\{x(s) \mid s \in \mathcal{S}\}\,.$$

Here $\mathcal{S}$ is the set of points at which we observe x; we denote the elements of $\mathcal{S}$ by $s_1, \ldots, s_n$. A mock-up of such a signal x is displayed in Figure 5.1.

In this section, building on the work of Candès and Fernadez-Granda [31, 32, 66] and Tang *et al* [20, 150, 149], we aim to show that we can recover the tuple (t_i, c_i, M) by solving a convex optimization problem. We formulate the superresolution imaging problem as an infinite dimensional optimization over measures. Precisely, note that the observed signal can be rewritten as

$$x(s) = \sum_{i=1}^{M} c_i \psi(s, t_i) = \int \psi(s,t) d\mu_\star(t)\,. \tag{5.1.2}$$

Here, $\mu_\star$ is the positive discrete measure $\sum_{i=1}^{M} c_i \delta_{t_i}$, where δ_t denotes the Dirac measure centered at t. We aim to show that we can recover $\mu_\star$ by solving the following optimization problem:

$$\begin{aligned} \underset{\mu}{\text{minimize}} \quad & \int w(t)\mu(dt) \\ \text{subject to} \quad & x(s) = \int \psi(s,t) d\mu(t), \quad s \in \mathcal{S} \\ & \operatorname{supp}\mu \subset B \\ & \mu \geq 0\,. \end{aligned} \tag{5.1.3}$$

Here, B is a fixed compact subset of the real line and $w(t)$ is a weighting function that weights the measure at different locations. The optimization problem (5.1.3) is over the set of all positive finite measures μ supported on B.

The optimization problem (5.1.2) is an analog of ℓ_1 minimization over the continuous domain B. Indeed, if we know a priori that the t_i are elements of a finite discrete set Ω, then optimizing over all measures subject to $\operatorname{supp}\mu \subset \Omega$ is precisely equivalent to weighted ℓ_1 minimization. This infinite dimensional analog with uniform weights has proven useful for compressed sensing over continuous domains [150], resolving diffraction-limited images from low-pass signals [31, 66, 149], system identification [138], and many other applications [41]. We will see below that the weighting function essentially ensures that all of the candidate locations are given equal influence in the optimization problem.

Our main result, Theorem 5.1.4, establishes that for one-dimensional signals, under rather mild conditions, we can recover $\mu_\star$ from the optimal solution of (5.1.3). Our conditions, described in full-detail below, essentially require the observation of at least $2M+1$ samples, and that the set of translates of the point spread function forms a linearly independent set. In Theorem 5.1.1 we verify that these conditions are satisfied by the Gaussian point spread function for any M source locations with no minimum separation condition. This is the first analysis of an ℓ_1 based method that matches the separation-free performance of polynomial root finding techniques [157, 53, 124]. Our motivation for such an analysis is that ℓ_1 based methods generalize to higher dimensions and are empirically stable in the presence of noise.

Boyd, Schiebinger and Recht [25] show that the problem (5.1.3) can be optimized to precision ϵ in polynomial time using a greedy algorithm. In our experiments in Section 5.3, we use this algorithm to demonstrate that our theory applies, and show that even in multiple dimensions with noise, we can recover closely spaced point sources.

5.1.1 Main Result

We restrict our theoretical attention to the one-dimensional case, leaving the higher-dimensional cases to future work. Let $\psi : \mathbb{R}^2 \to \mathbb{R}$ be our one dimensional point spread function, with the first argument denoting the position where we are observing the image of a point source located at the second argument. We assume that ψ is differentiable in both arguments.

For convenience, we will assume that $B = [-T, T]$ for some large scalar T. However, our proof will extend to more restricted subsets of the real line. Moreover, we will state our results for the special case where $\mathcal{S} = \{s_1, \ldots, s_n\}$, although our proof is written for possibly infinite measurement sets. We define the weighting function in the objective of our optimization problem via

$$w(t) = \frac{1}{n}\sum_{i=1}^{n} \psi(s_i, t)\,.$$

Our main result establishes conditions on ψ such that the true measure $\mu_\star$ is the unique optimal solution of (5.1.3). Importantly, we show that these conditions are satisfied by the Gaussian point spread function with no separation condition.

Theorem 5.1.1. *Suppose $|\mathcal{S}| > 2M$, and $\psi(s,t) = e^{-(s-t)^2}$. Then for any $t_1 < \ldots < t_M$, the true measure $\mu_\star$ is the unique optimal solution of* (5.1.3).

Before we proceed to state the main result, we need to introduce a bit of notation and define the notion of a Tchebycheff system. Let $K(t,\tau) = \frac{1}{n}\sum_{i=1}^{n} \psi(s_i,t)\psi(s_i,\tau)$, and define the vector valued function $v : \mathbb{R} \to \mathbb{R}^{2M}$ via

$$v(s) = \begin{bmatrix} \psi(s,t_1) & \ldots & \psi(s,t_M) & \frac{d}{dt_1}\psi(s,t_1) & \ldots & \frac{d}{dt_M}\psi(s,t_M) \end{bmatrix}^T . \tag{5.1.4}$$

Definition 5.1.2. *A set of functions $u_1, \ldots, u_n$ is called a* Tchebycheff system *(or* T-system*) if for any points $\tau_1 < \ldots < \tau_n$, the matrix*

$$\begin{pmatrix} u_1(\tau_1) & \ldots & u_1(\tau_n) \\ \vdots & & \\ u_n(\tau_1) & \ldots & u_n(\tau_n) \end{pmatrix}$$

is invertible.

Conditions 5.1.3. *We impose the following three conditions on the point spread function ψ:*

Positivity *For all $t \in B$ we have $w(t) > 0$.*

Independence *The matrix $\frac{1}{n}\sum_{i=1}^{n} v(s_i)v(s_i)^T$ is nonsingular.*

T-system *$\{K(\cdot,t_1), \ldots, K(\cdot,t_M), \frac{d}{dt_1}K(\cdot,t_1), \ldots, \frac{d}{dt_M}K(\cdot,t_M), w(\cdot)\}$ form a T-system.*

Theorem 5.1.4. *If ψ satisfies Conditions 5.1.3 and $|\mathcal{S}| > 2M$, then the true measure $\mu_\star$ is the unique optimal solution of* (5.1.3).

Note that the first two parts of Conditions 5.1.3 are easy to verify. Positivity eliminates the possibility that a candidate point spread function could equal zero at all locations—obviously we would not be able to recover the source in such a setting! Independence is satisfied if

$$\{\psi(\cdot,t_1), \ldots, \psi(\cdot,t_M), \frac{d}{dt_1}\psi(\cdot,t_1), \ldots, \frac{d}{dt_M}\psi(\cdot,t_M)\} \quad \text{is a T-system.}$$

This condition allows us to recover the amplitudes uniquely assuming we knew the true t_i locations *a priori*, but it is also useful for constructing a *dual certificate* as we discuss below.

We remark that we actually prove the theorem under a weaker condition than T-system. Define the matrix-valued function $\Lambda : \mathbb{R}^{2M+1} \to \mathbb{R}^{(2M+1)\times(2M+1)}$ by

$$\Lambda(p_1, \ldots, p_{2M+1}) := \begin{bmatrix} \kappa(p_1) & \ldots & \kappa(p_{2M+1}) \\ 1 & \ldots & 1 \end{bmatrix}, \tag{5.1.5}$$

where $\kappa : \mathbb{R} \to \mathbb{R}^{2M}$ is defined as

$$\kappa(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\psi(s_i,t)}{w(t)} v(s_i)\,. \tag{5.1.6}$$

Our proof of Theorem 5.1.4 replaces condition T-SYSTEM with the following:

DETERMINANTAL There exists $\rho > 0$ such that for any $t_i^-, t_i^+ \in (t_i - \rho, t_i + \rho)$, and $t \in [-T, T]$, the matrix $\Lambda\big(t_1^-, t_1^+, \ldots, t_M^-, t_M^+, t\big)$ is nonsingular whenever t, t_i^-, t_i^+ are distinct.

This condition looks more complicated than T-SYSTEM and is indeed nontrivial to verify. It is essentially a *local T-system* condition in the sense that the points τ_i in Definition 5.1.2 are restricted to lie in a small neighborhood about the t_i. It is clear that T-SYSTEM implies DETERMINANTAL. The advantage of the more general condition is that it can hold for finitely supported ψ, while this is not true for T-SYSTEM. In fact, it is easy to see that if T-SYSTEM holds for any point spread function ψ, then DETERMINANTAL holds for the truncated version $\psi(s,t)\mathbf{1}\{|s-t| \le 3T\}$, where $\mathbf{1}\{x \le y\}$ is the indicator variable equal to 1 when $x \le y$ and zero otherwise. We suspect that DETERMINANTAL may hold for significantly tighter truncations.

As we will see below, T-SYSTEM and INDEPENDENCE are related to the existence of a canonical *dual certificate* that is used ubiquitously in sparse approximation [33, 71]. In compressed sensing, this construction is due to Fuchs [71], but its origins lie in the theory of polynomial interpolation developed by Markov and Tchebycheff, and extended by Gantmacher, Krein, Karlin and others (see the survey in Section 5.1.2).

In the continuous setting of superresolution, the dual certificate becomes a *dual polynomial*: a function of the form $Q(t) = \frac{1}{n}\sum_{j=1}^n \psi(s_j, t)q(s_j)$ satisfying

$$\begin{aligned} Q(t) &\le w(t) \\ |Q(t_i)| &= w(t), \quad i = 1, \ldots, M. \end{aligned} \tag{5.1.7}$$

To see how T-SYSTEM might be useful for constructing a dual polynomial, note that as $t_1^+ \downarrow t_1$ and $t_1^- \uparrow t_1$, the first two columns of $\Lambda(t_1^+, t_1^-, \ldots, t)$ converge to the same column, namely $\kappa(t_1)$. However, if we divide by the difference $t_1^+ - t_1^-$, and take a limit then we obtain the derivative of the second column. In particular, some calculation shows that T-SYSTEM implies

$$\det \begin{bmatrix} A & \kappa(t) \\ \omega & w(t) \end{bmatrix} \neq 0 \qquad \forall t \neq t_i,$$

where $A = \frac{1}{n}\sum_{i=1}^n v(s_i)v(s_i)^T$ is the matrix from INDEPENDENCE, and

$$\omega = [w(t_1), \ldots, w(t_M), w'(t_1), \ldots, w'(t_M)].$$

Taking the Schur complement in $w(t)$, we find

$$\det \begin{bmatrix} A & \kappa(t) \\ \omega & w(t) \end{bmatrix} = \det A \left[\omega^T A^{-1}\kappa(t) - w(t)\right].$$

Hence it seems like the function $\omega^T A^{-1}\kappa(t)$ might serve well as our dual polynomial. However, it remains unclear from this short calculation that this function is bounded above

by $w(t)$. The proof of Theorem 5.1.4 makes this construction rigorous using the theory of T-systems.

Before turning to the proofs of these theorems (cf. Sections 5.2.1 and 5.2.4), we survey the mathematical theory of superresolution imaging.

5.1.2 Foundations: Tchebycheff Systems

Our proofs rely on the machinery of Tchebycheff[1] systems. This line of work originated in the 1884 doctoral thesis of A. A. Markov on approximating the value of an integral $\int_a^b f(x)dx$ from the moments $\int_a^b xf(x)dx$, ..., $\int_a^b x^n f(x)dx$. His work formed the basis of the proof by Tchebycheff (who was Markov's doctoral advisor) of the central limit theorem in 1887 [151].

Recall that we defined a T-system in Definition 5.1.2. An equivalent definition of a T-system is: *the functions $u_1, ..., u_n$ form a T-system if and only if every linear combination $U(t) = a_1u_1(t) + \cdots + a_nu_n(t)$ has at most $n-1$ zeros.* One natural example of a T-system is given by the functions $1, t, \ldots, t^{n-1}$. Indeed, a polynomial of degree $n-1$ can have at most $n-1$ zeros. Equivalently, the Vandermonde determinant does not vanish,

$$\begin{vmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_n \\ t_1^2 & t_2^2 & \ldots & t_n^2 \\ \vdots & & & \\ t_1^{n-1} & t_2^{n-1} & \ldots & t_n^{n-1} \end{vmatrix} \neq 0,$$

for any $t_1 < \ldots < t_n$. Just as Vandermonde systems are used to solve polynomial interpolation problems, T-systems allows the generalization of the tools from polynomial fitting to a broader class of nonlinear function-fitting problems. Indeed, given a T-system $u_1, ..., u_n$, a *generalized polynomial* is a linear combination $U(t) = a_1u_1(t) + \cdots + a_nu_n(t)$. The machinery of T-systems provides a basis for understanding the properties of these generalized polynomials. For a survey of T-systems and their applications in statistics and approximation theory, see [72, 94, 95]. In particular, many of our proofs are adapted from [95], and we call out the parallel theorems whenever this is the case.

5.1.3 Prior art and related work

Broadly speaking, superresolution techniques enhance the resolution of a sensing system, optical or otherwise; *resolution* is the distance at which distinct sources appear indistinguishable. The mathematical problem of localizing point sources from a blurred signal has applications in a wide array of empirical sciences: astronomers deconvolve images of stars to angular resolution beyond the Rayleigh limit [125], and biologists capture nanometer resolution images of fluorescent proteins [22, 91, 133, 162]. Detecting neural action potentials from

[1]Tchebycheff is one among many transliterations from Cyrillic. Others include Chebyshev, Chebychev, and Cebysev.

extracellular electrode measurements is fundamental to experimental neuroscience [61], and resolving the poles of a transfer function is fundamental to system identification [138]. To understand a radar signal, one must decompose it into reflections from different sources [89]; and to understand an NMR spectrum, one must decompose it into signatures from different chemicals [149].

The mathematical analysis of point source recovery has a long history going back to the work of Prony [124] who pioneered techniques for estimating sinusoidal frequencies. Prony's method, a multivariate version of which was discussed in Subsection 1.3.2, is based on algebraically solving for the roots of polynomials, and can recover arbitrarily closely spaced frequencies. The annihilation filter technique introduced by Vetterli [157] can perfectly recover any signal of *finite rate of innovation* with minimal samples. In particular the theory of signals with finite rate of innovation shows that given a superposition of pulses of the form $\sum a_k \psi(t - t_k)$, one can reconstruct the shifts t_k and coefficients a_k from a minimal number of samples [53, 157]. This holds without any separation condition on the t_k and as long as the base function ψ can reproduce polynomials of a certain degree (see [53, Section A.1] for more details). The algorithm used for this reconstruction is however based on polynomial rooting techniques that do not easily extend to higher dimensions. Moreover, this algebraic technique is not robust to noise (see the discussion in [148, Section IV.A] for example).

In contrast we study sparse recovery techniques. This line of thought goes back at least to Carathéodory [36, 35]. Our contribution is an analysis of ℓ_1 based methods that matches the performance of the algebraic techniques of Vetterli in the one dimensional and noiseless setting. Our primary motivation is that ℓ_1 based methods may be more stable to noise and generalize to higher dimensions (although our analysis currently does not).

It is tempting to apply the theory of compressed sensing [11, 33, 34, 50] to problem (5.1.3). If one assumes the point sources are located on a finite grid and are well separated, then some of the standard models for recovery are valid (e.g. incoherency, restricted isometry property, or restricted eigenvalue property). With this motivation, many authors solve the gridded form of the superresolution problem in practice [10, 12, 55, 56, 63, 90, 113, 128, 143, 144, 162]. However, this approach has some significant drawbacks. The theoretical requirements imposed by the classical models of compressed sensing become more stringent as the grid becomes finer. Furthermore, making the grid finer can also lead to numerical instabilities and computational bottlenecks in practice.

Despite recent successes in many empirical disciplines, the theory of superresolution imaging remains limited. Candès and Fernandes-Granada [32] recently made an important contribution to the mathematical analysis of superresolution, demonstrating that semi-infinite optimization could be used to solve the classical Prony problem. Their proof technique has formed the basis of several other analyses including that of Bendory *et al* [19] and that of Tang *et al* [149]. To better compare with our approach, we briefly describe the approach of [19, 32, 149] here.

They construct the vector q of a dual polynomial $Q(t) = \frac{1}{n} \sum_{j=1}^{n} \psi(s_j, t) q_j$ as a linear combination of $\psi(s, t_i)$ and $\frac{d}{dt_i} \psi(s, t_i)$. In particular, they define the coefficients of this linear

combination as the least squares solution to the system of equations

$$\begin{aligned} Q(t_i) &= \text{sign}(c_i), \qquad i = 1, \ldots, M; \\ \frac{d}{dt} Q(t)\Big|_{t=t_i} &= 0, \qquad\qquad i = 1, \ldots, M. \end{aligned} \tag{5.1.8}$$

They prove that, under a minimum separation condition on the t_i, the system has a unique solution because the matrix for the system is a perturbation of the identity, hence invertible.

Much of the mathematical analysis on superresolution has relied heavily on the assumption that the point sources are separated by more than some minimum amount [14, 19, 32, 52, 58, 117, 51]. We note that in practical situations with noisy observations, some form of minimum separation may be necessary. One can expect, however, that the required minimum separation should go to zero as the noise level decreases: a property that is not manifest in previous results. Our approach, by contrast, does away with the minimum separation condition by observing that the matrix for the system (5.1.8) need not be close to the identity to be invertible. Instead, we impose Conditions 5.1.3 to guarantee invertibility directly. Not surprisingly, we use techniques from T-systems to construct an analog of the polynomial Q in (5.1.8) for our specific problem.

Another key difference is that we consider the weighted objective $\int w(t) d\mu(t)$, while prior work [19, 32, 149] has analyzed the unweighted objective $\int d\mu(t)$. We, too, could not remove the separation condition without reweighing by $w(t)$. In Section 5.3 we provide evidence that this mathematically motivated reweighing step actually improves performance in practice. Weighting has proven to be a powerful tool in compressed sensing, and many works have shown that weighting an ℓ_1-like cost function can yield improved performance over standard ℓ_1 minimization [70, 96, 155, 24]. To our knowledge, the closest analogy to our use of weights comes from Rauhut and Ward, who use weights to balance the influence of dynamic range of bases in polynomial interpolation problems [129]. In the setting of this section, weights will serve to lessen the influence of sources that have low overlap with the observed samples.

We are not the first to bring the theory of Tchebycheff systems to bear on the problem of recovering finitely supported measures. De Castro and Gamboa [40] prove that a finitely supported positive measure μ can be recovered exactly from measurements of the form

$$\Big\{ \int u_0 d\mu, \ldots, \int u_n d\mu \Big\}$$

whenever $\{u_0, \ldots, u_n\}$ form a T-system containing the constant function $u_0 = 1$. These measurements are almost identical to ours; if we set $u_k(t) = \psi(s_k, t)$ for $k = 1, \ldots, n$, where $\{s_1, \ldots, s_n\} = \mathcal{S}$ is our measurement set, then our measurements are of the form

$$\{x(s) \mid s \in \mathcal{S}\} = \Big\{ \int u_1 d\mu, \ldots, \int u_n d\mu \Big\}.$$

However, in practice it is often impossible to directly measure the mass $\int u_0 d\mu = \int d\mu$ as required by (5.1.3). Moreover, the requirement that $\{1, \psi(s_1, t), \ldots, \psi(s_n, t)\}$ form a T-system

does not hold for the Gaussian point spread function $\psi(s,t) = e^{-(s-t)^2}$ (see Remark 5.2.6). Therefore the theory of [40] is not readily applicable to superresolution imaging.

We conclude our review of the literature by discussing some prior literature on ℓ_1-based superresolution without a minimum separation condition. We would like to mention the work of Fuchs [71] in the case that the point spread function is band-limited and the samples are on a regularly-spaced grid. This result also does not require a minimum separation condition. However, our results hold for considerably more general point spread functions and sampling patterns. Finally, in a recent paper Bendory [18] presents an analysis of ℓ_1 minimization in a discrete setup by imposing a Rayleigh regularity condition which, in the absence of noise, requires no minimum separation. Our results are of a different flavor, as our setup is continuous. Furthermore we require linear sample complexity while the theory of Bendory [18] requires infinitely many samples.

5.2 Proofs

In this section we prove Theorem 5.1.4 and Theorem 5.1.1. We start by giving a short list of notation to be used throughout the proofs. We write our proofs for an arbitrary measurement $\mathcal{S}$ which need not be finite for the sake of the proof. Let P denote a fixed positive measure on $\mathcal{S}$, and set

$$w(t) = \int \psi(s,t) dP(s).$$

For concreteness, the reader might think of P as the uniform measure over $\mathcal{S}$, where if $\mathcal{S}$ is finite then $w(t) = \frac{1}{n}\sum_{j=1}^{n} \psi(s_j, t)$. Just note that the particular choice of P does not affect the proof.

5.2.0.0.1 Notation Glossary

- We denote the inner product of functions $f, g \in L^2_P$ by $\langle f, g \rangle_P := \int f(t)g(t)dP(t)$.
- For any differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$, we denote the derivative in its first argument by $\partial_1 f$ and in its second argument by $\partial_2 f$.
- For $t \in \mathbb{R}$, let $\psi_t(\cdot) = \psi(\cdot, t)$.

5.2.1 Proof of Theorem 5.1.4

We prove Theorem 5.1.4 in two steps. We first reduce the proof to constructing a function q such that $\langle q, \psi_t \rangle_P$ possesses some specific properties.

Proposition 5.2.1. *If the first three items of Conditions 5.1.3 hold, and if there exists a function q such that $Q(t) := \langle q, \psi_t \rangle_P$ satisfies*

$$\begin{aligned} Q(t_j) &= w(t_j), \quad j = 1, \ldots, M \\ Q(t) &< w(t_j), \quad \text{for } t \in [-T, T] \text{ and } t \neq t_j, \end{aligned} \tag{5.2.1}$$

then the true measure $\mu_\star := \sum_{j=1}^{M} c_j \delta_{t_j}$ is the unique optimal solution of the program 5.1.3.

This proof technique is somewhat standard [33, 71]: the function $Q(t)$ is called a *dual certificate* of optimality. However, introducing the function $w(t)$ is a novel aspect of our proof. The majority of arguments have $w(t) = 1$. Note that when $\int \psi(s,t) dP(s)$ is independent of t, then $w(t)$ is a constant and we recover the usual method of proof.

In the second step we construct $q(s)$ as a linear combination of the t_i-centered point spread functions $\psi(s, t_i)$ and their derivatives $\partial_2 \psi(s, t_i)$.

Theorem 5.2.2. *Under the Conditions 5.1.3, there exist $\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_M, c \in \mathbb{R}$ such that $Q(t) = \langle q, \psi_t \rangle_P$ satisfies* (5.2.1)*, where*

$$q(s) = \sum_{i=1}^{M} (\alpha_i \psi(s, t_i) + \beta_i \frac{d}{dt_i} \psi(s, t_i)) + c.$$

To complete the proof of Theorem 5.1.4, it remains to prove Proposition 5.2.1 and Theorem 5.2.2. Their proofs can be found in Sections 5.2.2 and 5.2.3 respectively.

5.2.2 Proof of Proposition 5.2.1

We show that $\mu_\star$ is the optimal solution of problem (5.1.3) through strong duality. The dual of problem (5.1.3) is

$$\begin{aligned} \text{maximize}_q \quad & \langle q, x \rangle_P \\ \text{subject to} \quad & \langle q, \psi_t \rangle_P \leq w(t) \quad \text{for } t \in [-T, T]. \end{aligned} \tag{5.2.2}$$

Since the primal (5.1.3) is equality constrained, Slater's condition naturally holds, implying strong duality. As a consequence, we have

$$\langle q, x \rangle_P = \int w(t) d\mu(t) \iff q \text{ is dual optimal and } \mu \text{ is primal optimal.}$$

Suppose q satisfies (5.2.1). Hence q is dual feasible and we have

$$\begin{aligned} \langle q, x \rangle_P &= \sum_{j=1}^{M} c_j \langle q, \psi_{t_j} \rangle_P = \sum_{j=1}^{M} c_j Q(t_j) \\ &= \int w(t) d\mu_\star(t). \end{aligned}$$

Therefore, q is dual optimal and $\mu_\star$ is primal optimal.

Next we show uniqueness. Suppose the primal (5.1.3) has another optimal solution

$$\hat{\mu} = \sum_{j=1}^{\hat{M}} \hat{c}_j \delta_{\hat{t}_j}$$

such that $\{\hat{t}_1, \ldots, \hat{t}_{\hat{M}}\} \neq \{t_1, \ldots, t_M\} := \mathcal{T}$. Then we have

$$\begin{aligned}
\langle q, x \rangle_P &= \sum_j \hat{c}_j \Big\langle q, \psi_{\hat{t}_j} \Big\rangle_P \\
&= \sum_{\hat{t}_j \in \mathcal{T}} \hat{c}_j Q(\hat{t}_j) + \sum_{\hat{t}_j \notin \mathcal{T}} \hat{c}_j Q(\hat{t}_j) \\
&< \sum_{\hat{t}_j \in \mathcal{T}} \hat{c}_j w(\hat{t}_j) + \sum_{\hat{t}_j \notin \mathcal{T}} \hat{c}_j w(\hat{t}_j) = \int w(t) d\hat{\mu}(t).
\end{aligned}$$

Therefore, all optimal solutions must be supported on $\{t_1, \ldots, t_M\}$.

We now show that the coefficients of any optimal $\hat{\mu}$ are uniquely determined. By condition INDEPENDENCE, the matrix $\int v(s)v(s)^T dP(s)$ is invertible. Since it is also positive semidefinite, it is positive definite, so, in particular its upper $M \times M$ block is also positive definite.

$$\det \int \begin{bmatrix} \psi(s,t_1) \\ \vdots \\ \psi(s,t_M) \end{bmatrix} \begin{bmatrix} \psi(s,t_1) & \ldots & \psi(s,t_M) \end{bmatrix} dP(s) \neq 0.$$

Hence there must be $s_1, \ldots, s_M \in \mathcal{S}$ such that the matrix with entries $\psi(s_i, t_j)$ is nonsingular.

Now consider some optimal $\hat{\mu} = \sum_{i=1}^M \hat{c}_i t_i$. Since $\hat{\mu}$ is feasible we have

$$x(s_j) = \sum_{i=1}^M \hat{c}_i \psi(s_j, t_i) = \sum_{i=1}^M c_i \psi(s_j, t_i) \quad \text{for} \quad j = 1, \ldots, M.$$

Since $\psi(s_i, t_j)$ is invertible, we conclude that the coefficients $c_1, \ldots, c_M$ are unique. Hence $\mu_\star$ is the unique optimal solution of (5.1.3).

5.2.3 Proof of Theorem 5.2.2

We construct $Q(t)$ via a limiting interpolation argument due to Krein [103]. We have adapted some of our proofs (with nontrivial modifications) from the aforementioned text by Karlin and Studden [95]. We give reference to the specific places where we borrow from classical arguments.

In the sequel, we make frequent use of the following elementary manipulation of determinants:

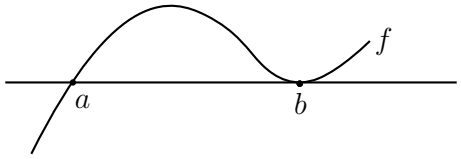


Figure 5.2: The point a is a *nodal* zero of f, and the point b is a *non-nodal* zero of f.

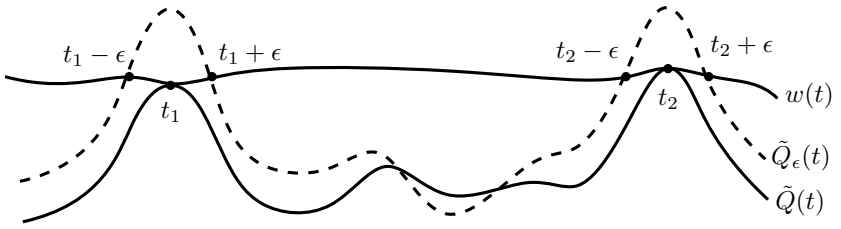


Figure 5.3: The relationship between the functions $w(t)$, $\tilde{Q}_\epsilon(t)$ and $\tilde{Q}(t)$. The function $\tilde{Q}_\epsilon(t)$ touches $w(t)$ only at $t_i \pm \epsilon$, and these are nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$. The function $\tilde{Q}(t)$ touches $w(t)$ only at t_i and these are non-nodal zeros of $\tilde{Q}(t) - w(t)$.

Lemma 5.2.3. *If $v_0, \ldots, v_n$ are vectors in $\mathbb{R}^n$, and n is even, then*

$$\begin{vmatrix} v_1 - v_0 & \ldots & v_n - v_0 \end{vmatrix} = \begin{vmatrix} v_1 & \ldots & v_n & v_0 \\ 1 & \ldots & 1 & 1 \end{vmatrix}.$$

We leave the proof of this lemma to the reader.

In what follows, we consider $\epsilon > 0$ such that

$$t_1 - \epsilon < t_1 + \epsilon < t_2 - \epsilon < t_2 + \epsilon < \cdots < t_M - \epsilon < t_M + \epsilon.$$

Definition 5.2.4. *A point t is a* nodal *zero of a continuous function $f : \mathbb{R} \to \mathbb{R}$ if $f(t) = 0$ and f changes sign at t. A point t is a* non-nodal *zero if $f(t) = 0$ but f does not change sign at t. This distinction is illustrated in Figure 5.2.*

Our proof of Theorem 5.2.2 proceeds as follows. With ϵ fixed, we construct a function

$$\tilde{Q}_\epsilon(t) = \sum_{i=1}^{M} \alpha_\epsilon^{[i]} K_P(t, t_i) + \beta_\epsilon^{[i]} \partial_2 K_P(t, t_i)$$

such that $\tilde{Q}_\epsilon(t) = w(t)$ only at the points $t = t_j \pm \epsilon$ for all $j = 1, 2, \ldots, M$ and the points $t_j \pm \epsilon$ are nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$ for all $j = 1, 2, \ldots, M$. We then consider the limiting function $\tilde{Q}(t) = \lim_{\epsilon \downarrow 0} \tilde{Q}_\epsilon(t)$, and prove that either $\tilde{Q}(t)$ satisfies (5.2.1) or $2w(t) - \tilde{Q}(t)$ satisfies (5.2.1). An illustration of this construction is pictured in Figure 5.3.

We begin with the construction of $\tilde{Q}_\epsilon$. We aim to find the coefficients $\alpha_\epsilon, \beta_\epsilon$ to satisfy

$$\tilde{Q}_\epsilon(t_i - \epsilon) = w(t_i - \epsilon) \quad \text{and} \quad \tilde{Q}_\epsilon(t_i + \epsilon) = w(t_i + \epsilon) \quad \text{for} \quad i = 1, \ldots, M.$$

This system of equations is equivalent to the system

$$\begin{aligned} \tilde{Q}_\epsilon(t_i - \epsilon) &= w(t_i - \epsilon) \quad \text{for} \quad i = 1, \dots, M \\ \frac{\tilde{Q}_\epsilon(t_i + \epsilon) - \tilde{Q}_\epsilon(t_i - \epsilon)}{2\epsilon} &= \frac{w(t_i + \epsilon) - w(t_i - \epsilon)}{2\epsilon} \quad \text{for} \quad i = 1, \dots, M. \end{aligned} \tag{5.2.3}$$

Note that this is a linear system of equations in $\alpha_\epsilon, \beta_\epsilon$ with coefficient matrix given by

$$\mathbf{K}_\epsilon := \left[\begin{array}{c|c} K_P(t_j - \epsilon, t_i) & \partial_2 K_P(t_j - \epsilon, t_i) \\ \hline \frac{1}{2\epsilon}\big(K_P(t_j + \epsilon, t_i) - K_P(t_j - \epsilon, t_i)\big) & \frac{1}{2\epsilon}\big(\partial_2 K_P(t_j + \epsilon, t_i) - \partial_2 K_P(t_j - \epsilon, t_i)\big) \end{array}\right].$$

That is, the equations (5.2.3) can be written as

$$\mathbf{K}_\epsilon \begin{bmatrix} | \\ \alpha_\epsilon \\ | \\ | \\ \beta_\epsilon \\ | \end{bmatrix} = \begin{bmatrix} w(t_1 - \epsilon) \\ \vdots \\ w(t_M - \epsilon) \\ \frac{1}{2\epsilon}(w(t_1 + \epsilon) - w(t_1 - \epsilon)) \\ \vdots \\ \frac{1}{2\epsilon}(w(t_M + \epsilon) - w(t_M - \epsilon)) \end{bmatrix}.$$

We first show that the matrix $\mathbf{K}_\epsilon$ is invertible for all ϵ sufficiently small. Note that as $\epsilon \to 0$ the matrix $\mathbf{K}_\epsilon$ converges to

$$\mathbf{K} := \left[\begin{array}{c|c} K_P(t_j, t_i) & \partial_2 K_P(t_j, t_i) \\ \hline \partial_1 K_P(t_j, t_i) & \partial_1 \partial_2 K_P(t_j, t_i) \end{array}\right] = \int v(s) v(s)^T dP(s),$$

which is positive definite by INDEPENDENCE. Since the entries of $\mathbf{K}_\epsilon$ converge to the entries of $\mathbf{K}$, there is a $\Delta > 0$ such that $\mathbf{K}_\epsilon$ is invertible for all $\epsilon \in (0, \Delta)$. Moreover, $\mathbf{K}_\epsilon^{-1}$ converges to $\mathbf{K}^{-1}$ as $\epsilon \to 0$ and for all $\epsilon < \Delta$, the coefficients are uniquely defined as

$$\begin{bmatrix} | \\ \alpha_\epsilon \\ | \\ | \\ \beta_\epsilon \\ | \end{bmatrix} = \mathbf{K}_\epsilon^{-1} \begin{bmatrix} w(t_1 - \epsilon) \\ \vdots \\ w(t_M - \epsilon) \\ \frac{1}{2\epsilon}(w(t_1 + \epsilon) - w(t_1 - \epsilon)) \\ \vdots \\ \frac{1}{2\epsilon}(w(t_M + \epsilon) - w(t_M - \epsilon)) \end{bmatrix}. \tag{5.2.4}$$

We denote the corresponding function by

$$\tilde{Q}_\epsilon(t) := \sum_{i=1}^{M} \alpha_\epsilon^{[i]} K_P(t, t_i) + \beta_\epsilon^{[i]} \partial_2 K_P(t, t_i).$$

Before we construct $\tilde{Q}(t)$, we take a moment to establish the following remarkable consequences of the DETERMINANTAL condition. For all $\epsilon > 0$ sufficiently small the following hold:

(a). $\tilde{Q}_\epsilon(t) = w(t)$ only at the points $t_1 - \epsilon, t_1 + \epsilon, \ldots, t_M - \epsilon, t_M + \epsilon$.

(b). These points $t_1 - \epsilon, t_1 + \epsilon, \ldots, t_M - \epsilon, t_M + \epsilon$ are nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$.

We adapted the proofs of (a) and (b) (with nontrivial modification) from the proofs of Theorem 1.6.1 and Theorem 1.6.2 of [95].

Proof of (a). Suppose for the sake of contradiction that there is a $\tau \in [-T, T]$ such that $\tilde{Q}_\epsilon(\tau) = w(\tau)$ and $\tau \notin \{t_1 - \epsilon, t_1 + \epsilon, \ldots, t_M - \epsilon, t_M + \epsilon\}$. Then we have the system of $2M$ linear equations

$$\frac{\tilde{Q}_\epsilon(t_j - \epsilon)}{w(t_j - \epsilon)} - \frac{\tilde{Q}_\epsilon(\tau)}{w(\tau)} = 0 \quad j = 1, \ldots, M$$
$$\frac{\tilde{Q}_\epsilon(t_j + \epsilon)}{w(t_j + \epsilon)} - \frac{\tilde{Q}_\epsilon(\tau)}{w(\tau)} = 0 \quad j = 1, \ldots, M.$$

Rewriting this in matrix form, the coefficient vector $\begin{bmatrix}\alpha_\epsilon & \beta_\epsilon\end{bmatrix} = \begin{bmatrix}\alpha_\epsilon^{[1]} & \cdots & \alpha_\epsilon^{[M]} & \beta_\epsilon^{[1]} & \cdots & \beta_\epsilon^{[M]}\end{bmatrix}$ of $\tilde{Q}_\epsilon$ satisfies

$$\begin{bmatrix}\alpha_\epsilon & \beta_\epsilon\end{bmatrix} \begin{pmatrix}\kappa(t_1 - \epsilon) - \kappa(\tau) & \kappa(t_1 + \epsilon) - \kappa(\tau) & \ldots & \kappa(t_M + \epsilon) - \kappa(\tau)\end{pmatrix} = \begin{bmatrix}0 & \ldots & 0\end{bmatrix}. \tag{5.2.5}$$

By Lemma 5.2.3 applied to the $2M + 1$ vectors $v_1 = \kappa(t_1 - \epsilon), \ldots, v_{2M} = \kappa(t_M + \epsilon)$, and $v_0 = \kappa(\tau)$, the matrix for the system of equations (5.2.5) is nonsingular if and only if the following matrix is nonsingular:

$$\begin{bmatrix}\kappa(t_1 - \epsilon) & \ldots & \kappa(t_M + \epsilon) & \kappa(\tau) \\ 1 & \ldots & 1 & 1\end{bmatrix} = \Lambda(t_1 - \epsilon, \ldots, t_M + \epsilon, \tau).$$

However, this is nonsingular by the DETERMINANTAL condition. This gives us the contradiction that completes the proof of part (a).

Proof of (b). Suppose for the sake of contradiction that $\tilde{Q}_\epsilon(t) - w(t)$ has $N_1 < 2M$ nodal zeros and $N_0 = 2M - N_1$ non-nodal zeros. Denote the nodal zeros by $\{\tau_1, ..., \tau_{N_1}\}$, and denote the non-nodal zeros by $z_1, \ldots, z_{N_0}$. In what follows, we obtain a contradiction by doubling the non-nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$. We do this by constructing a certain generalized polynomial $u(t)$ and adding a small multiple of it to $\tilde{Q}_\epsilon(t) - w(t)$.

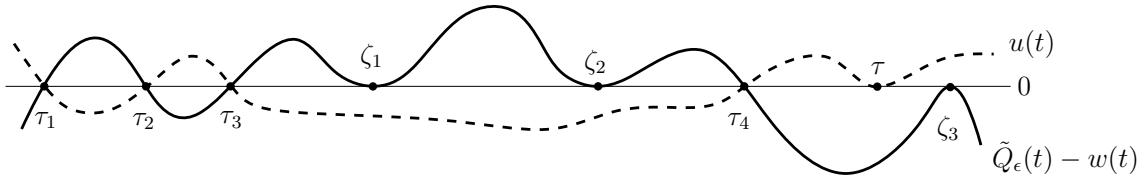


Figure 5.4: The points $\{\tau_1, \tau_2, \tau_3, \tau_4\}$ are nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$, and the points $\{\zeta_1, \zeta_2, \zeta_3\}$ are non-nodal zeros. The function $u(t)$ has the appropriate sign so that $\tilde{Q}_\epsilon(t) - w(t) + \delta u(t)$ retains nodal zeros at τ_i, and obtains two zeros in the vicinity of each ζ_i.

We divide the non-nodal zeros into groups according to whether $\tilde{Q}_\epsilon(t) - w(t)$ is positive or negative in a small neighborhood around the zero; define

$$\mathcal{I}^- := \{i \mid \tilde{Q}_\epsilon \leq w \text{ near } z_i\} \quad \text{and} \quad \mathcal{I}^+ := \{i \mid \tilde{Q}_\epsilon \geq w \text{ near } z_i\}.$$

We first show that there are coefficients $a_0, \ldots, a_M$, and $b_1, \ldots, b_M$ such that the polynomial

$$u(t) = \sum_{i=1}^{M} a_i K_P(t, t_i) + \sum_{i=1}^{M} b_i \partial_2 K_P(t, t_i) + a_0 w(t)$$

satisfies the system of equations

$$\begin{aligned} u(z_j) &= +1 \quad j \in \mathcal{I}^- \\ u(z_j) &= -1 \quad j \in \mathcal{I}^+ \\ u(\tau_i) &= 0 \quad i = 1, \ldots, N_1 \\ u(\tau) &= 0, \end{aligned} \tag{5.2.6}$$

where τ is some arbitrary additional point. The matrix for this system is

$$\mathbf{W} \begin{pmatrix} \kappa(z_1)^T & 1 \\ \vdots & \\ \kappa(z_{N_0})^T & 1 \\ \kappa(\tau_1)^T & 1 \\ \vdots & \\ \kappa(\tau_{N_1})^T & 1 \\ \kappa(\tau) & 1 \end{pmatrix}$$

where $\mathbf{W} = \operatorname{diag}\big(w(z_1), \ldots, w(z_{N_0}), w(\tau_1), \ldots, w(\tau_{N_1}), w(\tau)\big)$. This matrix is invertible by Determinantal since the nodal and non-nodal zeros of $\tilde{Q}_\epsilon(t) - w(t)$ are given by $t_1 - \epsilon, \ldots, t_M + \epsilon$. Hence there is a solution to the system (5.2.6).

Now consider the function

$$U^\delta(t) = \tilde{Q}_\epsilon(t) + \delta u(t) = \sum_{i=1}^{M} [\alpha_\epsilon^{[i]} + \delta a_i] K_P(t, t_i) + \sum_{i=1}^{M} [\beta_\epsilon^{[i]} + \delta b_i] \partial_2 K_P(t, t_i) + \delta a_0 w(t)$$

where $\delta > 0$. By construction, $u(\tau_i) = 0$, so $U^\delta(t) - w(t)$ has nodal zeros at $\tau_1, \ldots, \tau_{N_1}$. We can choose δ small enough so that $U^\delta(t) - w(t)$ vanishes twice in the vicinity of each z_i. This means that $U^\delta(t) - w(t)$ has $2M + N_0$ zeros. Assuming $N_0 > 0$, select a subset of these zeros $p_1 < \ldots < p_{2M+1}$ such that there are two in each interval $[t_i - \rho, t_i + \rho]$. This is possible if $\epsilon < \rho$ and δ is sufficiently small. We have the system of $2M + 1$ equations

$$\sum_{i=1}^{M} [\alpha_\epsilon^{[i]} + \delta a_i] K_P(p_1, t_i) + \sum_{i=1}^{M} [\beta_\epsilon^{[i]} + \delta b_i] \partial_2 K_P(p_1, t_i) = (1 - \delta a_0) w(\tau)$$

$$\vdots$$

$$\sum_{i=1}^{M} [\alpha_\epsilon^{[i]} + \delta a_i] K_P(p_{2M+1}, t_i) + \sum_{i=1}^{M} [\beta_\epsilon^{[i]} + \delta b_i] \partial_2 K_P(p_{2M+1}, t_i) = (1 - \delta a_0) w(\tau).$$

Subtracting the last equation from each of the first $2M$ equations, we find that

$$(\alpha_\epsilon^{[1]} + \delta a_1, \ldots, \beta_\epsilon^{[M]} + \delta b_M) \begin{pmatrix} \kappa(p_1) - \kappa(p_{2M+1}) & \ldots & \kappa(p_{2M}) - \kappa(p_{2M+1}) \end{pmatrix} = (0, \ldots, 0).$$

This matrix is nonsingular by Lemma 5.2.3 combined with the DETERMINANTAL condition. This contradiction implies that $N_0 = 0$. This completes the proof of (b).

We now complete the proof by constructing $\tilde{Q}(t)$ from $\tilde{Q}_\epsilon(t)$ by sending $\epsilon \to 0$. Note that the coefficients $\alpha_\epsilon, \beta_\epsilon$ converge as $\epsilon \to 0$ since the right hand side of equation (5.2.4) converges to

$$\mathbf{K}^{-1} \begin{bmatrix} w(t_1) \\ \vdots \\ w(t_M) \\ w'(t_1) \\ \vdots \\ w'(t_M) \end{bmatrix} = \begin{bmatrix} | \\ \alpha \\ | \\ | \\ \beta \\ | \end{bmatrix}.$$

We denote the limiting function by

$$\tilde{Q}(t) = \sum_{i=1}^{M} \alpha_i K_P(t, t_i) + \sum_{i=1}^{M} \beta_i \partial_2 K_P(t, t_i). \tag{5.2.7}$$

We conclude that $w(t) - \tilde{Q}(t)$ does not change sign at t_i since $w(t) - \tilde{Q}_\epsilon(t)$ changes sign only at $t_i \pm \epsilon$.

We now show that the limiting process does not introduce any additional zeros of $w(t) - \tilde{Q}(t)$. Suppose $\tilde{Q}(t)$ does touch $w(t)$ at some $\tau_1 \in [-T, T]$ with $\tau_1 \neq t_i$ for any $i = 1, ..., M$. Since $w(t) - \tilde{Q}(t)$ does not change sign, the points $t_1, \ldots, t_M, \tau_1$ are non-nodal zeros of $w(t) - \tilde{Q}(t)$. We find a contradiction by constructing a polynomial with two nodal zeros in the vicinity of each of these $M + 1$ points (but possibly only one nodal zero in the vicinity of τ_1 if $\tau_1 = T$ or $\tau_1 = -T$).

For sufficiently small $\gamma > 0$, the polynomial

$$W_\gamma(t) = \tilde{Q}(t) + \gamma w(t)$$

attains the value $w(t)$ twice in the vicinity of each t_i and twice in the vicinity of τ_1. In other words there exist $p_1 < \ldots < p_{2M+2}$ such that $W_\gamma(p_i) = w(p_i)$. Therefore

$$\tilde{Q}(p_i) = (1-\gamma)w(p_i) \quad \text{for} \quad i = 1, \ldots, 2M+2,$$

and so $\frac{\tilde{Q}(p_i)}{w(p_i)} - \frac{\tilde{Q}(p_{2M+1})}{w(p_{2M+1})} = 0$ for $i = 1, 2, ..., 2M$. Thus, the coefficient vector for the polynomial $\tilde{Q}(t)$ lies in the left nullspace of the matrix

$$\begin{pmatrix} \kappa(p_1) - \kappa(p_{2M+1}) & \ldots & \kappa(p_{2M}) - \kappa(p_{2M+1}) \end{pmatrix}.$$

However, this matrix is nonsingular by Lemma 5.2.3 and the DETERMINANTAL condition.

Collecting our results, we have proven that $\tilde{Q}(t) - w(t) = 0$ if and only if $t = t_i$ and that $\tilde{Q}(t) - w(t)$ does not change sign when t passes through t_i. Therefore one of the following is true

$$w(t) \geq \tilde{Q}(t) \quad \text{or} \quad \tilde{Q}(t) \geq w(t)$$

with equality iff $t = t_i$. In the first case, $Q(t) = \tilde{Q}(t)$ fulfills the prescriptions (5.2.1) with

$$q(t) = \sum_{i=1}^{M} \alpha_i \psi(s, t_i) + \beta_i \frac{d}{dt_i} \psi(s, t_i).$$

In the second case, $Q(t) = 2w(t) - \tilde{Q}(t)$ satisfies (5.2.1) with

$$q(t) = 2 - \sum_{i=1}^{M} \alpha_i \psi(s, t_i) + \beta_i \frac{d}{dt_i} \psi(s, t_i).$$

5.2.4 Proof of Theorem 5.1.1

INTEGRABILITY and POSITIVITY naturally hold for the Gaussian point spread function $\psi(s,t) = e^{-(s-t)^2}$. INDEPENDENCE holds because $\psi(s,t_1), \ldots, \psi(s,t_M)$ together with their derivatives $\partial_2\psi(s,t_1), \ldots, \partial_2\psi(s,t_M)$ form a T-system (see for example [95]). This means that for any $s_1 < \ldots < s_{2M} \in \mathbb{R}$,

$$\left| v(s_1) \ldots v(s_{2M}) \right| \neq 0,$$

and the determinant always takes the same sign. Therefore, by an integral version of the Cauchy-Binet formula for the determinant (cf. [94]),

$$\left| \int v(s)v(s)^T dP(s) \right| = (2M)! \int\limits_{s_1 < \ldots < s_{2M}} \left| v(s_1) \ldots v(s_{2M}) \right| \begin{vmatrix} v(s_1)^T \\ \vdots \\ v(s_{2M})^T \end{vmatrix} dP(s_1) \ldots dP(s_{2M}) \neq 0.$$

To establish the DETERMINANTAL condition, we prove the slightly stronger statement:

$$|\Lambda(p_1,\ldots,p_{2M+1})| = \left|\int \begin{bmatrix} v(s) \\ 1 \end{bmatrix} \begin{bmatrix} \frac{\psi(s,p_1)}{w(p_1)} & \ldots & \frac{\psi(s,p_{2M+1})}{w(p_{2M+1})} \end{bmatrix} dP(s)\right| \neq 0 \tag{5.2.8}$$

for any distinct $p_1,\ldots,p_{2M+1}$. When $p_1,\ldots,p_{2M+1}$ are restricted so that two points p_i, p_j lie in each ball $(t_k-\rho, t_k+\rho)$, we recover the statement of DETERMINANTAL.

We prove (5.2.8) with the following key lemma.

Lemma 5.2.5. *For any $s_1 < \ldots < s_{2M+1}$ and $t_1 < \ldots < t_M$,*

$$\begin{vmatrix} e^{-(s_1-t_1)^2} & \ldots & e^{-(s_{2M+1}-t_1)^2} \\ -(s_1-t_1)e^{-(s_1-t_1)^2} & \ldots & -(s_{2M+1}-t_1)e^{-(s_{2M+1}-t_1)^2} \\ \vdots & & \vdots \\ e^{-(s_1-t_M)^2} & \ldots & e^{-(s_{2M+1}-t_M)^2} \\ -(s_1-t_M)e^{-(s_1-t_M)^2} & \ldots & -(s_{2M+1}-t_M)e^{-(s_{2M+1}-t_M)^2} \\ 1 & \ldots & 1 \end{vmatrix} \neq 0.$$

Before proving this lemma, we show how it can be used to prove (5.2.8). By Lemma 5.2.5, we know in particular that for any $s_1 < \cdots < s_{2M+1}$,

$$\det \begin{bmatrix} v(s_1) & \cdots & v(s_{2M+1}) \\ 1 & \cdots & 1 \end{bmatrix} \neq 0$$

and is always the same sign. Moreover, for any $s_1 < \cdots < s_{2M+1}$, and any $p_1 < \ldots < p_{2M+1}$,

$$\det \begin{bmatrix} \psi(s_1,p_1) & \ldots & \psi(s_1,p_{2M+1}) \\ \vdots & & \\ \psi(s_{2M+1},p_1) & \ldots & \psi(s_{2M+1},p_{2M+1}) \end{bmatrix} > 0.$$

Any function with this property is called *totally positive* and it is well known that the Gaussian kernel is totally positive [95]. Now, to show that DETERMINANTAL holds for the finite sampling measure P, we use an integral version of the Cauchy-Binet formula for the determinant:

$$\left|\int \begin{bmatrix} v(s) \\ 1 \end{bmatrix} \begin{bmatrix} \frac{\psi(s,p_1)}{w(p_1)} & \ldots & \frac{\psi(s,p_{2M+1})}{w(p_{2M+1})} \end{bmatrix} dP(s)\right| =$$

$$= (2M+1)! \int_{s_1<\cdots<s_{2M+1}} \begin{vmatrix} v(s_1) & \cdots & v(s_{2M+1}) \\ 1 & \cdots & 1 \end{vmatrix} \begin{vmatrix} \frac{\psi(s_1,p_1)}{w(p_1)} & \ldots & \frac{\psi(s_1,p_{2M+1})}{w(p_{2M+1})} \\ \vdots & & \\ \frac{\psi(s_{2M+1},p_1)}{w(p_1)} & \ldots & \frac{\psi(s_{2M+1},p_{2M+1})}{w(p_{2M+1})} \end{vmatrix} dP(s_1)\ldots dP(s_{2M+1}).$$

The integral is nonzero since all integrands are nonzero and have the same sign. This proves (5.2.8).

Proof of Lemma 5.2.5. Multiplying the $2i-1$ and $2i$-th row by $e^{t_i^2}$ and the i-th column by $e^{s_i^2}$, and subtracting t_i times the $2i-1$-th row from the $2i$-th row, we obtain that we equivalently have to show that

$$\begin{vmatrix} e^{s_1t_1} & e^{s_2t_1} & \dots & e^{s_{2M+1}t_1} \\ s_1e^{s_1t_1} & s_2e^{s_2t_1} & \dots & s_{2M+1}e^{s_kt_1} \\ e^{s_1t_2} & e^{s_2t_2} & \dots & e^{s_{2M+1}t_2} \\ \vdots & & & \\ e^{s_1t_M} & e^{s_2t_M} & \dots & e^{s_{2M+1}t_M} \\ s_1e^{s_1t_M} & s_2e^{s_2t_M} & \dots & s_{2M+1}e^{s_{2M+1}t_M} \\ e^{s_1^2} & e^{s_2^2} & \dots & e^{s_{2M+1}^2} \end{vmatrix} \neq 0.$$

The above matrix has a vanishing determinant if and only if there exists a nonzero vector

$$(a_1, b_1, ..., a_M, b_M, a_{M+1})$$

in its left null space. This vector has to have nonzero last coordinate since by Example 1.1.5. in [95], the Gaussian kernel is extended totally positive and therefore the upper $2M \times 2M$ submatrix has a nonzero determinant. Therefore, we assume that $a_{M+1} = 1$. Thus, the matrix above has a vanishing determinant if and only if the function

$$\sum_{i=1}^{M}(a_i + b_is)e^{t_is} + e^{s^2} \tag{5.2.9}$$

has at least the $2M+1$ zeros $s_1 < s_2 < ... < s_{2M+1}$. Lemma 5.2.7, applied to $r = M$ and $d_1 = \cdots = d_M = 1$, establishes that this is impossible. To complete the proof of Lemma 5.2.5, it remains to state and prove Lemma 5.2.7. □

Remark 5.2.6. *The inclusion of the derivatives is essential for the shifted Gaussians to form a T-system together with the constant function* 1*. In particular, following the same logic as in the proof of Lemma 5.2.5, we find that* $\{1, e^{(s-t_1)^2}, \dots, e^{(s-t_M)^2}\}$ *form a T-system if and only if the function*

$$\sum_{i=1}^{M} a_ie^{t_is} + e^{s^2}$$

has at most M *zeros. However, for* $M = 3$ *the function has* 4 *zeros if we select* $a_1 = -3$*,* $t_1 = 1$*,* $a_2 = 7$*,* $t_2 = 0$*,* $a_3 = -5$*,* $t_3 = -1$*.*

Lemma 5.2.7. *Let* $d_1, ..., d_r \in \mathbb{N}$*. The function*

$$\phi_{d_1,...,d_r}(s) = \sum_{i=1}^{r}(a_{i0} + a_{i1}s + \cdots + a_{i(2d_i-1)}s^{2d_i-1})e^{t_is} + e^{s^2}$$

has at most $2(d_1 + \cdots + d_r)$ *zeros.*

Proof. We are going to show that $\phi_{d_1,...,d_r}(s)$ has at most $2(d_1+\cdots+d_r)$ zeros as follows. Let

$$g_0(s) = \phi_{d_1,...,d_r}(s).$$

For $k = 1, ..., d_1+\cdots+d_r$, let

$$g_k(s) = \begin{cases} \frac{d^2}{ds^2}\left[g_{k-1}(s)e^{(-t_j+t_1+\cdots+t_{j-1})s}\right], & \text{if } k = d_1+\cdots+d_{j-1}+1 \text{ for some } j, \\ \frac{d^2}{ds^2}\left[g_{k-1}(s)\right], & \text{otherwise.} \end{cases} \tag{5.2.10}$$

If we show that $g_{d_1+\cdots+d_r}(s)$ has no zeros, then, $g_{d_1+\cdots+d_r-1}(s)$ has at most two zeros, counting with multiplicity. By induction, it will follow that $g_0(s)$ has at most $2(d_1+\cdots+d_r)$ zeros, counting with multiplicity. Note that if $d_1+\cdots+d_{j-1} \leq k < d_1+\cdots+d_{j-1}+d_j$, then

$$\begin{aligned} g_k(s) &= (\tilde{a}_{j,2(k-d_1+\cdots+d_{j-1})}+\cdots+\tilde{a}_{j,(2d_j-1)}s^{2d_j-1-2(k-d_1+\cdots+d_{j-1})})+ \\ &\quad + \sum_{i=j+1}^{r}(\tilde{a}_{i0}+\cdots+\tilde{a}_{i(2d_i-1)}s^{2d_i-1})e^{(t_i-(t_1+\cdots+t_{j-1}))r} + cf_i(r)e^{r^2} \end{aligned}$$

where $c > 0$ is a constant and $r := s - c_i$. We are going to show that $f_i(r)$ is a sum of squares polynomial such that one of the squares is a positive constant. This would mean that $g_k(s) = f_k(s)e^{s^2}$ has no zeros.

Denote

$$\begin{aligned} p_0(s) &= 1 \\ p_1(s) &= 2s \\ &\vdots \\ p_i(s) &= 2sp_{i-1}(s-c_i) + p'_{i-1}(s-c_i), \end{aligned}$$

where $c_1, ..., c_i$ are constants. It follows by induction that the degree of $p_i(s)$ is $\deg(p_i) = i$ and the leading coefficient of $p_i(s)$ is 2^i.

We will show by induction that

$$\begin{aligned} f_i(s) &= p_i(s)^2 + \frac{1}{2}p'_i(s)^2 + \cdots + \frac{1}{2^i i!}p_i^{(i)}(s)^2 \\ &= \sum_{j=0}^{i}\frac{1}{2^j j!}p_i^{(j)}(s)^2. \end{aligned}$$

When $i = 0$, we have that $f_0(s) = 1$ and $\sum_{j=0}^{0}\frac{1}{2^j j!}p_0^{(j)}(s)^2 = 1$. We are going to prove the general statement by induction. Suppose the statement is true for $i-1$. By the rela-

tionship (5.2.10), we have

$$\begin{aligned} f_i(s)e^{s^2} &= \frac{d^2}{ds^2}\big[e^{s^2} f_{i-1}(s-c_i)\big] = \frac{d^2}{ds^2}\big[e^{s^2}\sum_{j=0}^{i-1}\frac{1}{2^j j!}p_{i-1}^{(j)}(s-c_i)^2\big] \qquad (5.2.11)\\ &= \sum_{j=0}^{i-1}\frac{e^{s^2}}{2^j j!}\Big\{2p_{i-1}^{(j+2)}(s-c_i)p_{i-1}^{(j)}(s-c_i) + 2p_{i-1}^{(j+1)}(s-c_i)^2 \\ &\quad + (4s^2+2)p_{i-1}^{(j)}(s-c_i)^2 + 8sp_{i-1}^{(j)}(s-c_i)p_{i-1}^{(j+1)}(s-c_i)\Big\} \end{aligned}$$

We need to show that this expression is equal to $e^{s^2}(\sum_{j=0}^{i}\frac{p_i^{(j)}(s)^2}{2^j j!})$. Since

$$p_i(s) = 2sp_{i-1}(s-c_i) + p'_{i-1}(s-c_i),$$

it follows by induction that $p_i^{(j)}(s) = 2jp_{i-1}^{(j-1)}(s-c_i)+2sp_{i-1}^{(j)}(s-c_i)+p_{i-1}^{(j+1)}(s-c_i)$. Therefore we obtain

$$\begin{aligned} e^{s^2}(\sum_{j=0}^{i}\frac{p_i^{(j)}(s)^2}{2^j j!}) &= e^{s^2}\sum_{j=0}^{i}\frac{1}{2^j j!}\Big[2jp_{i-1}^{(j-1)}(s-c_i) + 2sp_{i-1}^{(j)}(s-c_i) + p_{i-1}^{(j+1)}(s-c_i)\Big]^2. \\ &= e^{s^2}\sum_{j=0}^{i}\frac{1}{2^j j!}\Big[4j^2p_{i-1}^{(j-1)}(s-c_i)^2 + 4s^2p_{i-1}^{(j)}(s-c_I)^2 + p_{i-1}^{(j+1)}(s-c_i)^2 + \\ &\quad + 8jsp_{i-1}^{(j-1)}(s-c_i)p_{i-1}^{(j)}(s-c_i) + \\ &\quad + 4sp_{i-1}^{(j)}(s-c_i)p_{i-1}^{(j+1)} + 4jp_{i-1}^{(j-1)}(s-c_i)p_{i-1}^{(j+1)}(s-c_i)\Big] \end{aligned} \qquad (5.2.12)$$

There are four types of terms in the sums (5.2.11) and (5.2.12):

$$p_{i-1}^{(j)}(s-c_i)^2, \quad s^2p_{i-1}^{(j)}(s-c_i)^2, \quad p_{i-1}^{(j-1)}(s-c_i)p_{i-1}^{(j)}(s-c_i), \quad \text{and} \quad sp_{i-1}^{(j-1)}(s-c_i)p_{i-1}^{(j)}(s-c_i).$$

For a fixed $j \in \{0,1,...,i+1\}$, it is easy to check that the coefficients in front of each of these terms in (5.2.11) and (5.2.12) are equal. Therefore,

$$\begin{aligned} f_i(s) &= p_i(s)^2 + \frac{1}{2}p_i'(s)^2 + \cdots + \frac{1}{2^i i!}p_i^{(i)}(s)^2 \\ &= \sum_{j=0}^{i}\frac{1}{2^j j!}p_i^{(j)}(s)^2 \end{aligned}$$

Note that since $\deg(p_i) = i$, the i-th derivation $p_i^{(i)}(s)$ equals the leading coefficient of $p_i(s)$, which, as we discussed above, equals 2^i. Therefore, the term $\frac{1}{2^i i!}p_i^{(i)}(s)^2$ equals $2^i i!$. Thus, one of the squares in $f_i(s)$ is a positive number, so $f_i(s) > 0$ for all s. □

5.3 Numerical Experiments

In this section we present the results of several numerical experiments to complement our theoretical results. To allow for potentially noisy observations, we solve the constrained least squares problem

$$\begin{aligned} \underset{\mu \geq 0}{\text{minimize}} \quad & \sum_{i=1}^{n} \left(\int \psi(s_i, t) d\mu(t) - x(s_i) \right)^2 \\ \text{subject to} \quad & \int w(t) \mu(dt) \leq \tau \end{aligned} \tag{5.3.1}$$

using the conditional gradient method proposed in [25].

5.3.1 Reweighing matters for source localization

Our first numerical experiment provides evidence that weighting by $w(t)$ helps recover point sources near the border of the image. This matches our intuition: near the border, the mass of an observed point-source is smaller than if it were measured in the center of the image. Hence, if we didn't weight the candidate locations, sources that are close to the edge of the image would be beneficial to add to the representation.

We simulate two populations of images, one with point sources located away from the image boundary, and one with point sources located near the image boundary. For each population of images, we solve (5.3.1) with $w(t) = \int \psi(s,t) dP(s)$ (weighted) and with $w(t) = 1$ (unweighted). We find that the solutions to (5.3.1) recover the true point sources more accurately with $w(t) = \int \psi(s,t) dP(s)$.

We use the same procedure for computing accuracy as in [134]. Namely we match true point sources to estimated point courses and compute the *F-score* of the match. To describe this procedure in detail, we compute the F-score by solving a bipartite graph matching problem. In particular, we form the bipartite graph with an edge between t_i and $\hat{t}_j$ for all i, j such that $\|t_i - \hat{t}_j\| < r$, where $r > 0$ is a tolerance parameter, and $\hat{t}_1, \ldots, \hat{t}_N$ are the estimated point sources. Then we greedily select edges from this graph under the constraint that no two selected edges can share the same vertex; that is, no t_i can be paired with two $\hat{t}_j, \hat{t}_k$ or vice versa. Finally, the $\hat{t}_i$ successfully paired with some t_j are categorized as true positives, and we denote their number by T_P. The number of false negatives is $F_N = M - T_P$, and the number of false positives is $N - T_P$. The precision and recall are then $P = \frac{T_P}{T_P + F_N}$, and $R = \frac{T_P}{T_P + F_P}$ respectively, and the F-score is the harmonic mean:

$$F = \frac{2PR}{P + R}.$$

We find a match by greedily pairing points of $\{\tau_1, \ldots, \tau_N\}$ to elements of $\{t_1, \ldots, t_M\}$, and a tolerance radius $r > 0$ upper bounds the allow distance between any potential pairs. To emphasize the dependence on r, we sometimes write $F(r)$ for the F-score.

Both populations contain 100 images simulated using the Gaussian point spread function

$$\psi(s,t) = e^{-\frac{(s-t)^2}{\sigma^2}}$$

with $\sigma = 0.1$, and in both cases, the measurement set $\mathcal{S}$ is a dense uniform grid of $n = 100$ points covering $[0, 1]$. The populations differ in how the point sources for each image are chosen. Each image in the first population has five points drawn uniformly in the interval $(.1, .9)$, while each image in the second population has a total of four point sources with two point sources in each of the two boundary regions $(0, .1)$ and $(.9, 1)$. In both cases we assign intensity of 1 to all point sources, and solve (5.3.1) using an optimal value of τ (chosen with a preliminary simulation).

The results are displayed in Figure 5.5. The left subplot shows that the F-scores are essentially the same for the weighted and unweighted problems when the point sources are away from the boundary. This is not surprising because when t is away from the border of the image, then $\int \psi(s,t)dP(s)$ is essentially a constant, independent of t. But when the point sources are near the boundary, the weighting matters and the F-scores are dramatically better as shown in the right subplot.

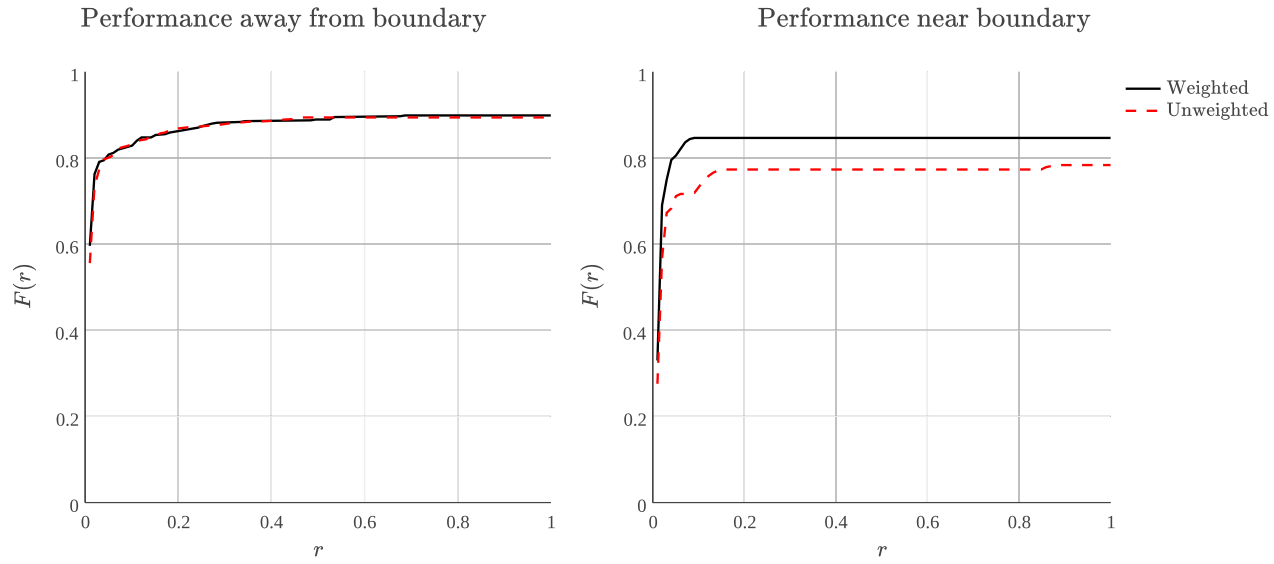
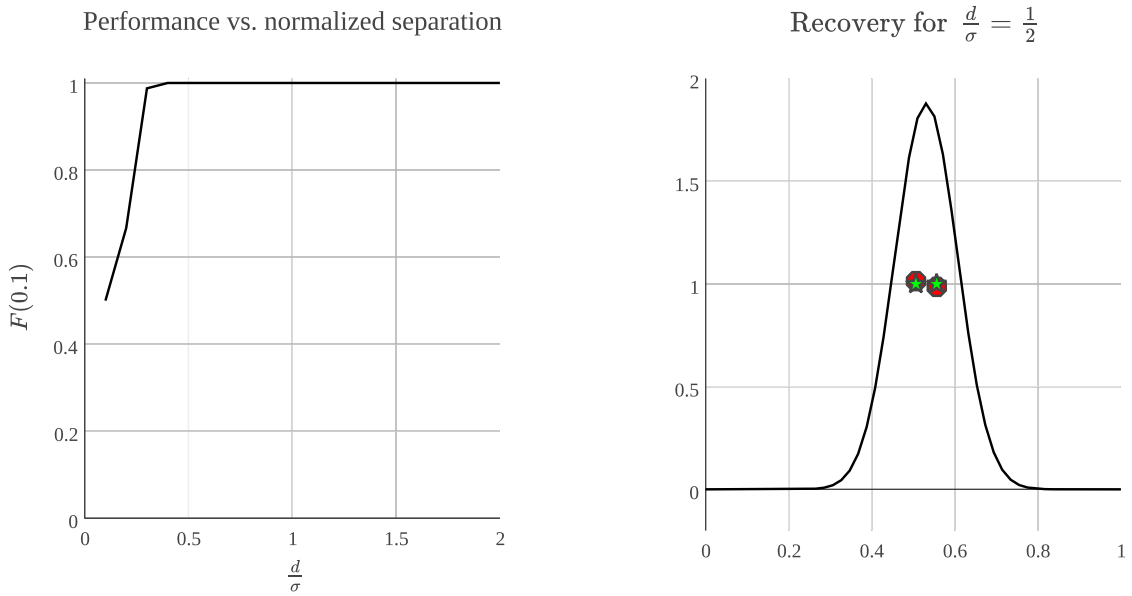


Figure 5.5: Reweighing matters for source localization. The two plots above compare the quality of solutions to the weighted problem (with $w(t) = \int \psi(s,t)dP(s)$) and the unweighted problem (with $w(t) = 1$). When point sources are away from the boundary (left plot), the performance is nearly identical. But when the point sources are near the boundary (right plot), the weighted method performs significantly better.



(a) (b)

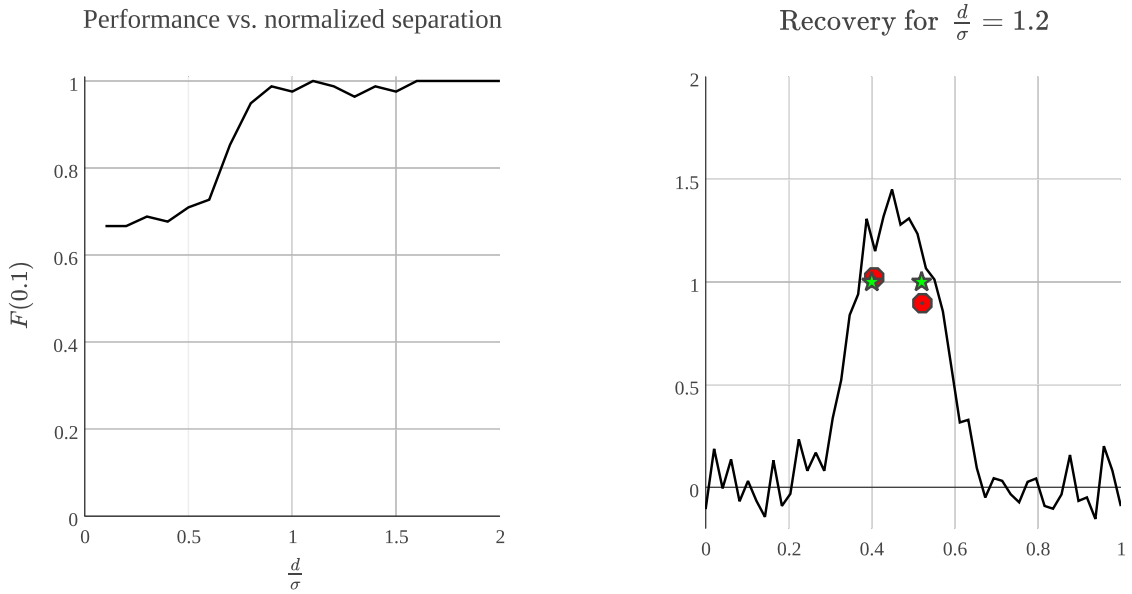
Figure 5.6: Sensitivity to point-source separation. (a) The F-score at tolerance radius $r = 0.1$ as a function of normalized separation $\frac{d}{\sigma}$. (b) The black trace shows an image for $\frac{d}{\sigma} = \frac{1}{2}$. The green stars show the locations (x-coordinate) and weights (y-coordinate) of the true point sources. The red dots show the recovered locations and weights.

5.3.2 Sensitivity to point-source separation

Our theoretical results assert that in the absence of noise the optimal solution of (5.1.3) recovers point sources with no minimum bound on the separation. In the following experiment, we explore the ability of (5.3.1) to recover pairs of points as a function of their separation. The setup is similar to the first numerical experiment. We use the Gaussian point spread function with $\sigma = 0.1$ as before, but here we observe only $n = 50$ samples. For each separation $d \in \{.1\sigma, .2\sigma, \ldots, 1.9\sigma, 2\sigma\}$, we simulate a population of 20 images containing two point sources separated by d. The point sources are chosen by picking a random point x away from the border of the image and placing two point sources at $x \pm \frac{d}{2}$. Again, each point source is assigned an intensity of 1, and we attempt to recover the locations of the point sources by solving (5.3.1).

In the left subplot of Figure 5.6 we plot F-score versus separation for the value of τ that produces the best F-scores. Note that we achieve near perfect recovery for separations greater than $\frac{\sigma}{4}$. The right subplot of Figure 5.6 shows the observations, true point sources, and estimated point sources for a separation of $\frac{d}{\sigma} = \frac{1}{2}$. Note the near perfect recovery in spite of the small separation.

Due to numerical issues, we cannot localize point sources with arbitrarily small $d > 0$. Indeed, the F-score for $\frac{d}{\sigma} < \frac{1}{4}$ is quite poor. This does not contradict our theory because



(a) (b)

Figure 5.7: Sensitivity to noise. (a) The F-score at tolerance radius $r = 0.1$ as a function of normalized separation $\frac{d}{\sigma}$. (b) The black trace is the 50 pixel image we observe. The green stars show the locations (x-coordinate) and weights (y-coordinate) of the true point sources. The red dots show the recovered locations and weights.

numerical ill-conditioning is in effect adding noise to the recovery problem, and we expect that a separation condition will be necessary in the presence of noise.

5.3.3 Sensitivity to noise

Next, we investigate the performance of (5.3.1) in the presence of additive noise. The setup is identical to the previous numerical experiment, except that we add Gaussian noise to the observations. In particular, our noisy observations are

$$\{x(s_i) + \eta_i \mid s_i \in \mathcal{S}\}$$

where $\eta_i \sim \mathcal{N}(0, 0.1)$.

We measure the performance of (5.3.1) in Figure 5.7. Note that we achieve near-perfect recovery when $d > \sigma$. However, if $d < \sigma$ the F-scores are clearly worse than the noiseless case. Unsurprisingly, we observe that sources must be separated in order to recover their locations to reasonable precision. We defer an investigation of the dependence of the signal separation as a function of the signal-to-noise ratio to future work.

5.3.4 Extension to two-dimensions

Though our proof does not extend as is, we do expect generalizations of our recovery result to higher dimensional settings. The optimization problem (5.3.1) extends immediately to arbitrary dimensions, and we have observed that it performs quite well in practice. We demonstrate in Figure 5.8 the power of applying (5.3.1) to a high density fluorescence image in simulation. Figure 5.8 shows an image simulated with parameters specified by the Single Molecule Localization Microscopy challenge [83]. In this challenge, point sources are blurred by a Gaussian point-spread function and then corrupted by noise. The green stars show the true locations of a simulated collection of point sources, and the red dots show the support of the measure output by (5.3.1) applied to the greyscale image forming the background of Figure 5.8. The overlap between the true locations and estimated locations is near perfect with an F-score of 0.98 for a tolerance radius corresponding to one third of a pixel.

5.4 Conclusions and Future Work

In this section we have demonstrated that one can recover the centers of a nonnegative sum of Gaussians from a few samples by solving a convex optimization problem. This recovery is theoretically possible no matter how close the true centers are to one-another. We remark that similar results are true for recovering measures from their moments. Indeed, the atoms of a positive atomic measure can be recovered no matter how close together the atoms are, provided one observes twice the number of moments as there are atoms. Our work can be seen as a generalization of this result, applying generalized polynomials and the theory of Tchebycheff systems in place of properties of Vandermonde systems.

As we discussed in our numerical experiments, this work opens up several theoretical problems that would benefit from future investigation. We close with a very brief discussion of some of the possible extensions.

5.4.1 Noise

Motivated by the fact that there is no separation condition in the absence of noise, it would be interesting to study how the required separation decays to zero as the noise level decreases. One of the key-advantages of using convex optimization for signal processing is that dual certificates generically give stability results, in the same way that Lagrange multipliers measure sensitivity in linear programming. Previous work on estimating line-spectra has shown that dual polynomials constructed for noiseless recovery extend to certify properties of estimation and localization in the presence of noise [31, 66, 149]. We believe that these methods should be directly applicable to our problem set-up.

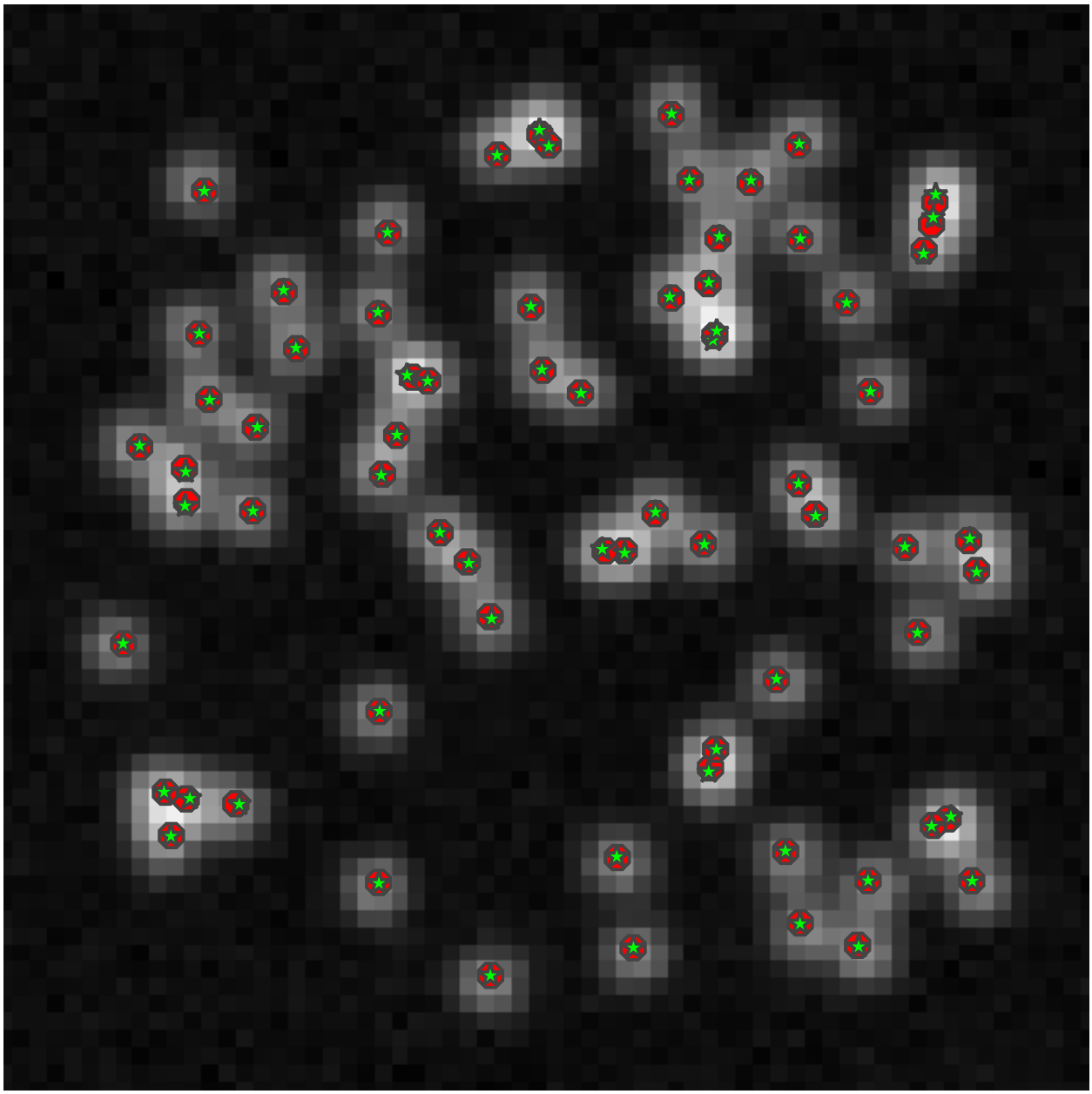


Figure 5.8: High density single molecule imaging. The green stars show the locations of a simulated collection point sources, and the greyscale background shows the noisy, pixelated point spread image. The red dots show the support of the measure-valued solution of (5.3.1).

5.4.2 Higher dimensions

One logical extension is proving that the same results hold in higher dimensions. Most scientific and engineering applications of interest have point sources arising from one to four dimensions, and we expect that some version of our results should hold in higher dimensions. Indeed, we believe a guarantee for recovery with no separation condition can be proven in higher dimensions with noiseless observations. However, it is not straightforward to extend our results to higher dimensions because the theory of Tchebycheff systems is only developed in one dimension. In particular, our approach using limits of polynomials does not directly generalize to higher dimensions.

5.4.3 Other point spread functions

We have shown that our Conditions 5.1.3 hold for the Gaussian point spread function, which is commonly used in microscopy as an approximation to an Airy function. It will be very useful to show that they also hold for other point spread functions such as the Airy function and other common physical models. Our proof relied heavily on algebraic properties of the Gaussian, but there is a long, rich history of determinantal systems that may apply to generalize our result. In particular, works on properties of totally positive systems may be fruitful for such generalizations [9, 123].

5.4.4 Model mismatch in the point spread function

Our analysis relies on perfect knowledge of the point spread function. In practice one never has an exact analytic expression for the point spread function. Aberrations in manufacturing and scattering media can lead to distortions in the image not properly captured by a forward model. It would be interesting to derive guarantees on recovery that assume only partial knowledge of the point spread function. Note that the optimization problem of searching both for the locations of the sources and for the associated wave-function is a blind deconvolution problem, and techniques from this well-studied problem could likely be extended to the super-resolution setting. If successful, such methods could have immediate practical impact when applied to denoising images in molecular, cellular, and astronomical imaging.

Bibliography

[1] A. Aggrawal et al. "Finding minimal convex nested polygons". In: *Information and Computation* 83 (1989), pp. 98–110.

[2] J. Alexander and A. Hirschowitz. "Polynomial interpolation in several variables". In: *Journal of Algebraic Geometry* 4.2 (1995), pp. 201–222.

[3] E. Allman, C. Matias, and J. Rhodes. "Identifiability of parameters in latent structure models with many observed variables". In: *Annals of Statistics* 37 (2009), pp. 3099–3132.

[4] E. Allman, J. Rhodes, and A. Taylor. "A semialgebraic description of the general Markov model on phylogenetic trees". In: *SIAM Journal on Discrete Mathematics* 28 (2014), pp. 736–755.

[5] E. Allman et al. "Tensors of nonnegative rank two". In: *Linear Algebra and its Applications* 473 (2015), pp. 37–53.

[6] A. Anandkumar, R. Ge, and M. Janzamin. "Guaranteed non-orthogonal tensor decomposition via altenating rank-1 updates". In: *JMLR: Workshop and Conference Proceedings* 40 (2015), pp. 1–77.

[7] A. Anandkumar, R. Ge, and M. Janzamin. "Sample Complexity Analysis for Learning Overcomplete Latent Variable Models through Tensor Methods". In: *Preprint:arxiv:1408.0553+* (2014).

[8] A. Anandkumar et al. "Tensor Decompositions for Learning Latent Variable Models". In: *Journal of Machine Learning Research* (2012).

[9] T. Ando. "Totally positive matrices". In: *Linear algebra and its applications* 90 (1987), pp. 165–219.

[10] W.U. Bajwa et al. "Compressed channel sensing: A new approach to estimating sparse multipath channels". In: *Proc. IEEE* 98.6 (2010), pp. 1058–1076.

[11] R. Baraniuk. "Compressive sensing [lecture notes]". In: *IEEE Signal Process Mag* 24.4 (2007), pp. 118–121.

[12] R. Baraniuk and P. Steeghs. "Compressive radar imaging". In: *In IEEE Radar Conf., Waltham, MA* (2007), pp. 128–133.

[13] S. Basu, R. Pollack, and M. F. Roy. *Algorithms in real algebraic geometry*. Vol. 10. Algorithms and Computation in Mathematics. Berlin: Springer Verlag, 2003.

[14] D. Batenkov and Y. Yomdin. "Algebraic fourier reconstruction of piecewise smooth functions". In: *Math. Comput.* 81 (2012).

[15] D.J. Bates et al. *Numerically Solving Polynomial Systems with Bertini. Software, Environments, and Tools*. SIAM, 2013.

[16] K. Batselier, H. Liu, and N. Wong. "A constructive algorithm for decomposing a tensor into a finite sum of orthonormal rank-1 terms." In: *SIAM Journal on Matrix Analysis and Applications* 36.3 (2015), pp. 1315–1337.

[17] E. Beale. "Discussion of "Maximum likelihood from incomplete data via the EM algorithm" by A. Dempster, N. Laird and D. Rubin". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 22–23.

[18] T. Bendory. "Robust recovery of positive stream of pulses". In: *Preprint: arXiv:1503.08782* (2015).

[19] T. Bendory, S. Dekel, and A. Feuer. "Robust recovery of stream of pulses using convex optimization". In: *Preprint: arXiv:1412.3262* (2014).

[20] B.N. Bhaskar, G. Tang, and B. Recht. "Atomic norm denoising with applications to line spectral estimation". In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5987–5999.

[21] C. Bocci, E. Carlini, and F. Rapallo. "Perturbation of matrices and nonnegative rank with a view toward statistical models". In: *SIAM Journal on Matrix Analysis and Applications* 32 (2011), pp. 1500–1512.

[22] J.S. Bonifacino et al. "Imaging intracellular fluorescent proteins at nanometer resolution". In: *Science* 313 (2006), pp. 1642–1645.

[23] A. Boralevi et al. "Orthogonal and unitary tensor decomposition from an algebraic perspective". In: *Preprint arXiv:1512.08031* (2015).

[24] R. Von Borries, C.J. Miosso, and C. Potes. "Compressed sensing using prior information". In: *Computational Advances in Multi-Sensor Adaptive Processing, 2007. CAMPSAP 2007. 2nd IEEE International Workshop on*. IEEE. 2007, pp. 121–124.

[25] N. Boyd, G. Schiebinger, and B. Recht. "The alternating descent conditional gradient method for sparse inverse problems". In: *Preprint.* (2015).

[26] J. Brachat et al. "Symmetric Tensor Decomposition". In: *Linear Algebra and its Applications* 433.11-12 (2010), pp. 851–872.

[27] M. Brambilla and G. Ottaviani. "On the Alexander-Hirschowitz theorem". In: *Journal of Pure and Applied Algebra* 212 (2008), pp. 1229–1251.

[28] M. Brion and S. Kumar. *Frobenius splitting methods in geometry and representation theory*. Vol. 231. Boston, MA: Birkhäuser, 2005.

[29] J. Cahill, D. Mixon, and N. Strawn. "Connectivity and irreducibility of algebraic varieties of finite unit norm tight frames". In: *Preprint: arXiv:1311.4748* (2013).

[30] J. Cahill and N. Strawn. *Finite Frames. Algebraic geometry and finite frames.* Applied Numerical Harmonic Analysis. Birkhäuser/Springer New York, 2013, pp. 141–170.

[31] E.J. Candès and C. Fernandez-Granda. "Super-resolution from noisy data". In: *Journal of Fourier Analysis and Applications* 19.6 (2013), pp. 1229–1254.

[32] E.J. Candès and C. Fernandez-Granda. "Towards a mathematical theory of super resolution". In: *Comm. Pure Appl. Math* (2013).

[33] E.J. Candès, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". In: *IEEE Trans. Inf. Thy.* 52.2 (2006), pp. 489–509.

[34] E.J. Candès and M. Wakin. "An introduction to compressive sampling". In: *IEEE Signal Process. Mag.* 25.2 (2008), pp. 21–30.

[35] C. Carathéodory. "Ueber den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen". In: *Rend. Circ. Mat.* 32 (1911), pp. 193–217.

[36] C. Carathéodory. "Ueber den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehman." In: *Math. Ann.* 64 (1907), pp. 95–115.

[37] D. Cartwright and B. Sturmfels. "The number of eigenvalues of a tensor". In: *Linear Algebra and its Applications* 432.2 (2013), pp. 942–952.

[38] P. Casazza, G. Kutyniok, and F. Philipp. *Finite Frames. Introduction to finite frame theory.* Applied Numerical Harmonic Analysis. Birkhäuser/Springer New York, 2013, pp. 1–53.

[39] P. Casazza et al. "Every Hilbert space frame has a Naimark complement". In: *Journal of Mathematical Analysis and Applications* 406 (2013), pp. 111–119.

[40] Y. de Castro and F. Gamboa. "Exact reconstruction using Beurling minimal extrapolation". In: *http://arxiv.org/abs/1103.4951* (2011).

[41] V. Chandrasekaran et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[42] J. Chen and Y. Saad. "On the tensor SVD and the optimal low rank orthogonal approximation of tensors." In: *SIAM Journal on Matrix Analysis and Applications* 30.4 (2009), pp. 1709–1734.

[43] J. Cohen and U. Rothblum. "Nonnegative ranks, decompositions, and factorizations of nonnegative matrices". In: *Linear Algebra and its Applications* 190 (1993), pp. 149–168.

[44] P. Comon. "Independent component analysis, a new concept?" In: *Signal Process.* 36.3 (1994), pp. 287–314.

[45] P. Comon et al. "Symmetric tensors and symmetric tensor rank". In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1254–1279.

[46] D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms. An introduction to computational algebraic geometry and commutative algebra*. Undergraduate Texts in Mathematics. New York: Springer-Verlag, 1992.

[47] A. Critch. "Binary hidden Markov models and varieties". In: *Journal of Algebraic Statistics* 4 (2013), pp. 1–30.

[48] L. De Lathauwer, B. De Moor, and J. Vandewalle. "A multilinear singular value decomposition." In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[49] A. Dempster, N. Laird, and D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 1–22.

[50] D. Donoho. "Compressed sensing". In: *IEEE Trans. Inf. Thy.* 52.4 (2006), pp. 1289–1306.

[51] D. Donoho. "Superresolution via sparsity constraints". In: *SIAM J. Math. Anal.* (1992).

[52] D. Donoho and P. Stark. "Uncertainty principles and signal recovery". In: *SIAM J. Appl. Math* 49 (1989), pp. 906–931.

[53] P. Dragotti, M. Vetterli, and T. Blu. "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang-fix". In: *IEEE Transactions on Signal Processing* 55 (2007), pp. 1741–1757.

[54] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Vol. 39. Oberwolfach Seminars. Basel: Birkhäuser, 2009.

[55] M. Duarte and R. Baraniuk. "Spectral compressive sensing". In: *Applied and Computational Harmonic Analysis* 35.1 (2013), pp. 111–129.

[56] V. Duval and G. Peyré. "Exact support recovery for sparse spikes deconvolution". In: *Foundations of Computational Mathematics* 15.5 (2015).

[57] K. Dykema and N. Strawn. "Manifold structure of spaces of spherical tight frames". In: *International Journal of Pure and Applied Mathematics* 28 (2006), pp. 217–256.

[58] K.S. Eckhoff. "Accurate reconstructions of functions of finite regularity from truncated fourier series expansions". In: *Math. Comput* 64 (1995), pp. 671–690.

[59] R. Eggermont, E. Horobet¸, and K. Kubjas. "Algebraic boundary of matrices of nonnegative rank at most three". In: *Preprint: arXiv:1412.1654* (2014).

[60] D. Eisenbud and B. Sturmfels. "Binomial Ideals". In: *Duke Mathematical Journal* 84 (1996), pp. 1–45.

[61] C. Ekanadham, D. Tranchina, and E.P. Simoncelli. "Neural spike identifcation with continuous basis pursuit". In: *Computational and Systems Neuroscience (CoSyNe), Salt Lake City, Utah* (2011).

[62] D. Evanko. "Primer: fluorescence imaging under the diffraction limit". In: *Nature Methods* 6 (2009), pp. 19–20.

[63] A.C. Fannjiang, T. Strohmer, and P. Yan. "Compressed remote sensing of sparse objects". In: *SIAM Journal of Imaging Science* 3.3 (2010), pp. 595–618.

[64] H. Fawzi and P. Parillo. "Self-scaled bounds for atomic cone ranks: applications to nonnegative rank and cp-rank". In: *To appear in Mathematical Programming Series A* (2014).

[65] H. Fawzi et al. "Positive semidefinite rank". In: *Mathematical Programming Series B, special issue on "Lifts of Convex Sets"* 153.1 (2015), pp. 133–177.

[66] C. Fernandez-Granda. "Support detection in super-resolution". In: *arXiv:1302.3921* (2013).

[67] S. Fienberg. "Discussion of "Maximum likelihood from incomplete data via the EM algorithm" by A. Dempster, N. Laird and D. RubinMaximum likelihood from incomplete data via the EM algorithm" by A. Dempster, N. Laird and D. Rubin". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 29–30.

[68] S. Fienberg et al. "Maximum likelihood estimation in latent class models for contingency table data". In: *Algebraic and Geometric Methods in Statistics, Cambridge University Press* (2010), pp. 27–62.

[69] S. Friedland and G. Ottaviani. "The number of singular vector tuples and uniqueness of best rank one approximation of tensors". In: *Foundations of Computational Mathematics* 14 (2014), pp. 1209–1242.

[70] M.P. Friedlander et al. "Recovering compressively sampled signals using partial support information". In: *Information Theory, IEEE Transactions on* 58.2 (2012), pp. 1122–1134.

[71] J.-J. Fuchs. "Sparsity and uniqueness for some specific under-determined linear systems". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* 5 (2005), pp. 729–732.

[72] F. P. Gantmacher and M.G. Krein. "Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems". In: *Revised English Ed. AMS Chelsea Pub. Providence, RI* (2002).

[73] L. Garcia, M. Stillman, and B. Sturmfels. "Algebraic geometry of Bayesian networks". In: *Journal of Symbolic Computation* 39 (2005), pp. 331–355.

[74] B. Georgi and A. Schliep. "Context-specific independence mixture modeling for positional weight matrices". In: *Bioinformatics* 22 (2006), pp. 166–173.

[75] N. Gillis and F. Glineur. "On the geometric interpretation of the nonnegative rank". In: *Linear Algebra and its Applications* 437 (2012), pp. 2685–2712.

[76] J. Gouviea, P. A. Parrilo, and R. R. Thomas. "Approximate cone factorizations and lifts of polytopes". In: *Mathematical Programming* 151 (2015), pp. 613–637.

[77] J. Gouviea, P. A. Parrilo, and R. R. Thomas. "Lifts of Convex Sets and Cone Factorizations". In: *Mathematics of Operations Research* 38 (2013), pp. 248–264.

[78] J. Gouviea, R. Z. Robinson, and R. R. Thomas. "Polytopes of minimum positive semidefinite rank". In: *Discrete and Computational Geometry* 50 (2013), pp. 679–699.

[79] J. Gouviea, R. Z. Robinson, and R. R. Thomas. "Worst-case results for positive semidefinite rank". In: *Mathematical Programming Series B* 153.1 (2015), pp. 201–212.

[80] J. Gouviea et al. "Four dimensional polytopes of minimum positive semidefinite rank". In: *Preprint arXiv:1506.00187* (2015).

[81] D. Grayson and M. Stillman. "Macaulay2, a software system for research in algebraic geometry". In: *available at http://www.math.uiuc.edu/Macaulay2/* ().

[82] E. Gross and J. Rodriguez. "Maximum likelihood geometry in the presence of data zeros". In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation* (2014), pp. 232–239.

[83] Biomedical Imaging Group. *Benchmarking of Single-Molecule Localization Microscopy Software*. 2013. URL: `http://bigwww.epfl.ch/palm/`.

[84] J. Harris. *Algebraic Geometry: A First Course*. Vol. 133. Graduate Texts in Mathematics. Springer-Verlag, New York, 1995.

[85] J. Hauenstein, J. Rodriguez, and B. Sturmfels. "Maximum likelihood for matrices with rank constraints". In: *Journal of Algebraic Statistics* 5 (2014), pp. 18–38.

[86] J.D. Hauenstein, C. Ikenmeyer, and J.M. Landsberg. "Equations for lower bounds on border rank". In: *Experimental Mathematics* 22 (2013), pp. 373–383.

[87] J.D. Hauenstein and A.J. Sommese. "Witness sets of projections". In: *Applied Mathematics and Computation* 217 (2010), pp. 3349–3354.

[88] J.D. Hauenstein et al. "Homotopy techniques for tensor decomposition and perfect identifiability". In: *Preprint: arXiv:1501.00090* (2015).

[89] R. Heckel, V. Morgenshtern, and M. Soltanolkotabi. "Super-resolution radar". In: *arXiv:1411.6272v2* (2015).

[90] M.A. Herman and T. Strohmer. "High-resolution radar via compressed sensing". In: *IEEE Trans. Signal Process.* 57.6 (2009), pp. 2275–2284.

[91] S.T. Hess, T.P. Giriajan, and M.D. Mason. "Ultra-high resolution imaging by fluorescence pho- toactivation localization microscopy". In: *Biophysical Journal* 91 (2006), pp. 4258–4272.

[92] C. Hillar and L.-H. Lim. "Most tensor problems are NP hard". In: *Journal of the ACM* 60.6 (2013), Art. 45.

[93] J. Huh and B. Sturmfels. "Likelihood Geometry". In: *Combinatorial Algebraic Geometry (eds. Aldo Conca et al.), Lecture Notes in Mathematics, Springer* 2108 (2014), pp. 63–117.

[94] S. Karlin. "Total Positivity: Volume I". In: *Stanford University Press* (1968).

[95] S. Karlin and W. Studden. "Tchebycheff Systems: with Applications in Analysis and Statistics". In: *Wiley Interscience* (1967).

[96] M.A. Khajehnejad et al. "Analyzing weighted minimization for sparse recovery with nonuniform sparse models". In: *Signal Processing, IEEE Transactions on* 59.5 (2011), pp. 1985–2001.

[97] H. Klauck et al. "Limitations of convex programming: lower bounds on extended formulations and factorization ranks". In: *(Dagstuhl Seminar 15082) Dagstuhl Reports* 5 (2015), pp. 2192–5283.

[98] T. Kolda. "A counterexample to the possibility of an extension of the Eckart-Young low-rank approximation theorem for the orthogonal rank tensor decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 24.3 (2003), pp. 762–767.

[99] T. Kolda. "Orthogonal Tensor Decompositions". In: *SIAM Journal on Matrix Analysis and Applications* 23.1 (2001), pp. 243–255.

[100] T. Kolda. "Symmetric Orthogonal Tensor Decomposition is Trivial". In: *Preprint: arXiv:1503.01375+* (2015).

[101] T. Kolda and B. Bader. "Tensor decompositions and applications". In: *SIAM Review* 51.3 (2009), pp. 455–500.

[102] H. Kraft and C. Procesi. *Classical Invariant Theory, a Primer*. https://math.unibas.ch/uploads/x4e
1996.

[103] M.G. Krein. "The ideas of P.L. Tchebycheff and A.A. Markov in the theory of limiting values of integrals and their futher development". In: *American Matematical Society Translations. Series 2.* 12 (1959).

[104] K. Kubjas, E. Robeva, and R. Z. Robinson. "Positive semidefinite rank and nested spectrahedra". In: *Preprint arXiv:1512.08766* (2015).

[105] K. Kubjas, E. Robeva, and B. Sturmfels. "Fixed Points of the EM algorithm and nonnegative rank boundaries". In: *Annals of Statistics* 43 (2015), pp. 422–461.

[106] K. Kubjas and Z. Rosen. "Matrix completion for the independence model". In: *arXiv:1407.3254* (2014).

[107] S. Kunis et al. "A multivariate generalization of Prnoy's method". In: *preprint arXiv:1506.00450* (2015).

[108] J.M. Landsberg. *Tensors: geometry and applications*. Graduate Studies in Mathematics. American Mathematical Society, 2011.

[109] J.M. Landsberg and G. Ottaviani. "Equations for secant varieties of Veronese and other varieties". In: *Annali di Matematica Pura ed Applicata* 192.4 (2013), pp. 569–606.

[110] L. De Lathauwer, B. De Moor, and J. Vandewalle. "A multilinear singular value decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[111] D. Lee and H. Seung. "Learning the parts of objects by nonnegative matrix factorization". In: *Nature* 401 (1999), pp. 788–791.

[112] L.-H. Lim. "Singular values and eigenvalues of tensors: a variational approach". In: *Computational Advances in Multi-Sensor Adaptive Processing, 1st IEEE International Workshop* (2005), pp. 129–132.

[113] D. Malioutov, M. Cetin, and A.S. Willsky. "A sparse signal reconstruction perspective for source localization with sensor arrays". In: *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 3010–3022.

[114] E. Miller and B. Sturmfels. *Combinatorial commutative algebra*. Vol. 227. Graduate Texts in Mathematics. Springer, 2005.

[115] A. Moitra. "An almost optimal algorithm for computing nonnegative rank". In: *Symposium on Discrete Algorithms* (2013), pp. 1454–1464.

[116] D. Mond, J. Smith, and D. van Straten. "Stochastic factorizations, sandwiched simplices and the topology of the space of explanations". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 459 (2003), pp. 2821–2845.

[117] V.I. Morgenshtern and E.J. Candès. "Stable Super-Resolution of Positive Sources: the Discrete Setup". In: *arXiv:1504.00717* (2015).

[118] G. Murray. "Discussion of "Maximum likelihood from incomplete data via the EM algorithm" by A. Dempster, N. Laird and D. Rubin". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 27–28.

[119] Nobelprize.org. *The Nobel Prize in Chemistry 2014*. URL: `http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2014/`.

[120] L. Oeding and G. Ottaviani. "Eigenvectors of tensors and algorithms for Waring decomposition". In: *Journal of Symbolic Computation* 54 (2013), pp. 9–35.

[121] L. Oeding, E. Robeva, and B. Sturmfels. "Decomposing tensors into frames". In: *Advances in Applied Mathematics* 76 (2016), pp. 125–153.

[122] L. Pachter and B. Sturmfels. *Algebraic statistics for computational biology*. Cambridge University Press, 2005.

[123] A. Pinkus. *Totally positive matrices*. Vol. 181. Cambridge University Press, 2010.

[124] B.G.R. de Prony. "Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, à différentes températures". In: *Journal de l'École Polytechnique* 1.22 (1795), pp. 24–76.

[125] K.G. Puschmann and F. Kneer. "On super-resolution in astronomical imaging". In: *Astronomy and Astrophysics* 436 (2005), pp. 373–378.

[126] L. Qi. "Eigenvectors of a real symmetric tensor". In: *Journal of Symbolic Computation* 40.6 (2005), pp. 1302–1324.

[127] S. Ragnarsson and C. Van Loan. "Block tensors and symmetric embeddings." In: *Linear Algebra and its Applications* 438.2 (2013), pp. 853–874.

[128] H. Rauhut. "Random sampling of sparse trigonometric polynomials". In: *Applied and Comput. Hamon. Anal.* 22.1 (2007), pp. 16–42.

[129] H. Rauhut and R. Ward. "Interpolation via weighted l_1 minimization". In: *Applied and Computational Harmonic Analysis* (2015). To Appear. Preprint available at arxiv:1308.0759.

[130] E. Robeva. "Orthogonal Decomposition of Symmetric Tensors". 2014.

[131] E. Robeva. "Orthogonal decomposition of symmetric tensors". In: *SIAM Journal on Matrix Analysis and Applications* 37 (2016), pp. 86–102.

[132] E. Robeva and A. Seigal. "Singular vectors of orthogonally decomposable tensors". In: *Preprint arXiv:1603.09004* (2016).

[133] M.J. Rust, M. Bates, and X. Zhuang. "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)". In: *Nature Methods* 3 (2006), pp. 793–796.

[134] D. Sage et al. "Quantitative evaluation of software packages for single-molecule localization microscopy". In: *Nat Meth* advance online publication (June 15, 2015), pages. URL: `http://dx.doi.org/10.1038/nmeth.3442`.

[135] J. Salmi, A. Richter, and V. Koivunen. "Sequential Unfolding SVD for Tensors With Applications in Array Signal Processing". In: *IEEE Trans. Signal Proc.* 57.12 (2009), pp. 4719–4733.

[136] R. Sanyal, F. Sottille, and B. Sturmfels. "Orbitopes". In: *Mathematika* 57 (2011), pp. 275–314.

[137] G. Schiebinger, E. Robeva, and B. Recht. "Superresolution without separation". In: *Preprint arXiv:1506.03144* (2015).

[138] P. Shah et al. "Linear System Identification via Atomic Norm Regularization". In: *arXiv:1204.0590* (2012).

[139] J. Sidman and S. Sullivant. "Prolongations and computational algebra". In: *Canadian Journal of Mathematics* 4 (2009), pp. 930–949.

[140] R. Sinn and B. Sturmfels. "Generic spectrahedral shadows". In: *SIAM Journal on Optimization* 25 (2015), pp. 1209–1220.

[141] A.J. Sommese, J. Verschelde, and C.W. Wampler. "Symmetric functions applied to decomposing solution sets of polynomial systems". In: *SIAM Journal of Numerical Analysis* 40 (2002), pp. 2026–2046.

[142] R. Stanley. *Combinatorics and Commutative Algebra*. Ed. by Second. Vol. 41. Progress in Mathematics. Birkhäuser Boston, Boston, MA, 1996.

[143] P. Stoica and P. Babu. "Spice and likes: Two hyperparameter-free methods for sparse-parameter estimation". In: *Signal Processing* (2011).

[144] P. Stoica, P. Babu, and J. Li. "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data". In: *IEEE Transactions on Signal Processing* 59.1 (2011), pp. 35–47.

[145] N. Strawn. "Finite frame varieties: nonsingular points, tangent spaces, and explicit local parametrizations". In: *Journal of Fourier Analysis and Applications* 17.5 (2011), pp. 821–853.

[146] B. Sturmfels. *Gröbner bases and convex polytopes*. University Lecture Series. American Mathematical Society, 1996.

[147] B. Sturmfels. *Solving systems of polynomial equations*. Vol. 97. CBMS Series. Providece, RI: American Mathematical Society, 2002.

[148] V. Tan and V. Goyal. "Estimating signals with finite rate of innovation from noisy samples: a stochastic algorithm". In: *IEEE Transactions on Signal Processing* 56.10 (Oct. 2008).

[149] G. Tang, B. Bhaskar, and B. Recht. "Near minimax line spectral estimation". In: *IEEE Transactions on Information Theory* (2014). To appear.

[150] G. Tang et al. "Compressed Sensing off the Grid". In: *IEEE Transactions on Information Theory* 59.11 (2013), pp. 7465–7490.

[151] P.L. Tchebycheff. "On two theorems with respect to probabilities". In: *Zap. Akad. Nauk S.-Petersburg* 55 (1887), pp. 156–168.

[152] *The Netflix Prize*. URL: `http://www.netflixprize.com/rules`.

[153] A. Vandaele et al. "Heuristics for exact nonnegative matrix factorization". In: *Journal of Global Optimization* (2015), pp. 1–32.

[154] N. Vannieuwenhoven et al. "On generic nonexistence of the Schmidt-Eckart-Young decomposition for complex tensors." In: *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), pp. 886–903.

[155] N. Vaswani and W. Lu. "Modified-CS: Modifying compressive sensing for problems with partially known support". In: *Signal Processing, IEEE Transactions on* 58.9 (2010), pp. 4595–4607.

[156] S. Vavasis. "On the complexity of nonnegative matrix factorization". In: *SIAM Journal on Optimization* 20 (2009), pp. 1364–1377.

[157] M. Vetterli, P Marziliano, and T. Blu. "Sampling signals with finite rate of innovation". In: *IEEE Transactions on Signal Processing* 50.6 (2002), pp. 1417–1428.

[158] J. Cannon W. Bosma and C. Playoust. "The Magma algebra system. I. The user language". In: *Computational Algebra and Number Theory* 24 (1997), pp. 235–265.

[159] N. White. "Geometric applications of the Grassmann-Cayley algebra". In: *in: Handbook of Discrete and Computational Geometry, CRC Press Ser. Discrete Math. Appl., CRC, Boca Raton, F* (1997), pp. 881–892.

[160] T. Zhang and G. Golub. "Rank-one approximation to high order tensors". In: *SIAM Journal on Matrix Analysis and Applications* 23 (2001), pp. 534–550.

[161] T. Zhang and G. Golub. "Rank-one approximation to high order tensors." In: *SIAM Journal on Matrix Analysis and Applications* 23.2 (2001), pp. 534–550.

[162] L. Zhu et al. "Faster storm using compressed sensing". In: *Nature Methods* 9 (2012), pp. 721–723.

[163] M. Zhu, G. Jiang, and S. Gao. "Solving the 100 Swiss Francs problem". In: *Mathematics in Computer Science* 5 (2011), pp. 195–207.

[164] P. Zwiernik and J. Smith. "Implicit inequality constraints in a binary tree model". In: *Electronic Journal of Statistics* 5 (2011), pp. 1276–1312.