Title
Numerical algebraic geometry for maximum likelihood estimation

Author
Rodriguez, Jose Israel

Publication Date
2014

Peer reviewed|Thesis/dissertation

Numerical algebraic geometry for maximum likelihood estimation

by

Jose Israel Rodriguez III

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 Lauren Williams
Professor Laurent El Ghaoui

Spring 2014

Numerical algebraic geometry for maximum likelihood estimation

Abstract

Numerical algebraic geometry for maximum likelihood estimation

by

Jose Israel Rodriguez III

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Bernd Sturmfels, Chair

Numerical algebraic geometry provides numerical descriptions of solution sets of polynomial systems of equations in several unknown. Such sets are called algebraic varieties. In algebraic statistics, a statistical model is associated to an algebraic variety to study its geometric structure. This thesis contains my work at UC Berkeley that uses numerical algebraic geometry for the algebraic statistics problem of maximum likelihood estimation.

In Chapter 2 we study the maximum likelihood estimation problem on manifolds of matrices with bounded rank. These represent mixtures of distributions of two independent discrete random variables. We determine the maximum likelihood degree for a range of determinantal varieties, and we apply numerical algebraic geometry to compute all critical points of their likelihood functions.

In Chapter 3 we prove a bijection between critical points of the likelihood function on the complex variety of matrices of rank r and critical points on the complex variety of matrices of corank $r-1$. From the perspective of statistics, we show that maximum likelihood estimation for matrices of rank r is the same problem as minimum likelihood estimation for matrices of corank $r-1$, and vice versa.

In Chapter 4, a description of the maximum likelihood estimation problem in terms of dual varieties and conormal varieties is given. With this description, we define the dual likelihood equations. We show how solving these dual likelihood equations give solutions to the maximum likelihood estimation problem without having the defining equations of the model itself.

In Chapter 5, discrete algebraic statistical models are considered and solutions to the likelihood equations when the data contain zeros are studied. Focusing on sampling and model zeros, we show that the solutions of the likelihood equations in these cases are contained in a previously studied variety, the likelihood correspondence. The number of solutions give a lower bound on the ML degree, and the problem of finding critical points to the likelihood function can be partitioned into computationally easier problems involving sampling and model zeros.

In Chapter 6 the `Macaulay2` package `Bertini.m2` is introduced. `Macaulay2` is a software system designed to support research in algebraic geometry, and `Bertini` is a popular software system for numerical algebraic geometry. The package `Bertini.m2` provides an interface to `Bertini` via `Macaulay2`.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my advisor, Bernd Sturmfels. During these past four years, he has consistently offered advice, motivation, and encouragement. No matter how busy he was or what country he was located, he could always find the time to have a meeting or suggest a new problem.

I am grateful to my collaborators Daniel J. Bates, Jan Draisma, Elizabeth Gross, Jonathan Hauenstein, Anton Leykin, Maurizio Ruggiu, and Bernd Sturmfels. It was a great opportunity to have had the insightful conversations with you all.

In addition, I owe great thanks to Laurent El Ghaoui, Vera Serganova, and Lauren Williams for taking the time to serve on my committees.

I have been very fortunate to have met great friends, colleagues, and mentors, including, Dave Anderson, Daniel Appel, Roberto Barrera, Avinash Bhardwaj, Matthias Beck, Elan Bechor, Greg Blekherman, Florian Block, Sarah Brodsky, Dustin Cartwright, Melody Chan, Andrew Critch Angelica Cueto, Alicia Dickenstein, Jason Ferguson, Alex Fink, Benjamin Gaines, Oran Gannot, Noah Giansiracusa, Pete Glaze, Christian Haase, Christian Hilaire, Jan Hofmann, Serkan Hosten, Nathan Ilten, Eric Katz, Robert Krone, Kaie Kubjas, Chris Hillar, Kim Laine, Adam Kalman, Shaowei Lin, Olya Mandelshtam, Abraham Martin Del Campo, Ralph Morrison, Luke Oeding, Martin Olsson, Lior Pachter, Colette Patt, Eric Peterson, Qingchun Ren, Felipe Rincon, Elina Robeva, Zvi Rosen, Philipp Rostalski, Steven Sam, Anne Shiu, Sean Simmons, Frank Sottile, Pierre-Jean Spaenlehauer, Seth Sullivant, Thorsten Theobald, Christian Trabandt, Ngoc Tran, Caroline Uhler, Cynthia Vinzant, Barb Waller, Luming Wang, Timo de Wolff, Kevin Wray, Rika Yatchak, Ira Young, Josephine Yu, and Piotr Zwiernik.

I would also like to thank the AVID students of Grand Prairie and my fellow McNair scholars for their Texas sized support and affirmations. Finally, I must thank my family: Dad, Mom, and Marissa. Without you, I wouldn't be me.

Chapter 1

Introduction

Studying connections between statistics and algebra is at the center of algebraic statistics [12]. A statistical model $\mathcal{M}$ for discrete data is a subset of the positive orthant of $\mathbb{R}^{n+1}$ where the coordinates sum to one. In this thesis, we consider only models that are defined by the vanishing of polynomial equations restricted to the positive orthant. A remarkable fact and motivation of algebraic statistics is that many interesting statistical models are described in this fashion. Since an algebraic variety is a solution set of polynomial equations, we can use computational algebra to study statistical models.

The focus of this thesis is to study the maximum likelihood estimation (MLE) problem. Given a statistical model $\mathcal{M} \subset \mathbb{R}^{n+1}$, the MLE problem is to maximize the likelihood function on the model for given data $u \in \mathbb{N}^{n+1}$. The point that maximizes this function is called the maximum likelihood estimate (mle). One way to determine this point, is to use local hill climbing methods. However, if there exist many local maxima, then one cannot guarantee convergence of these local methods without further analysis.

The approach we take in this thesis is to consider the algebraic variety $\overline{\mathcal{M}}$ that is the Zariski closure of our statistical model. Next we determine all complex critical points of the likelihood function restricted to the variety $\overline{\mathcal{M}}$. Usually there will be finitely many critical points, and of these points, we will determine the ones with positive coordinates. So instead of maximizing the likelihood function over $\mathcal{M}$, we maximize over the finitely many positive critical points of the likelihood function on $\overline{\mathcal{M}}$. With additional hypotheses these two results will agree.

To determine the critical points, we solve the system of likelihood equations [9, 23] for the model $\overline{\mathcal{M}}$ with respect to data u. The method that we use to solve this system involves numerical algebraic geometry and homotopies [3, 41]. The key idea is that the likelihood equations are difficult to solve, and for each choice of data u one would have to consider a new system. But with numerical algebraic geometry, once the likelihood equations have been solved for a generic choice of data u, one can quickly recover solutions to likelihood equations for another choice of data using a homotopy. More specifically, the system of equations and its solutions that have already been found are deformed to a new system and target solutions using numerical algorithms. In the remainder of this chapter, we will focus on a specific statistical model to illustrate the key concepts and motivations of this thesis.

1.1 Illustrative example

In this section, we introduce an illustrative example involving a weighted coin and weighted dice. This example is based off the classical example of DiaNA and her dice in [35].

One coin, four dice

Suppose Oscar has a coin that is weighted such that the probability of observing side 1 is c_1 and the probability of observing side 2 is c_2. Further suppose Oscar has two pair of four-sided dice. The first pair of dice consists of a red die `R1` and a blue die `B1`. The probability of observing one of the four sides of these respective dice is given by the matrices

$$[r_1, r_2, r_3, r_4]^T \text{ and } [b_1, b_2, b_3, b_4]^T.$$

Similarly, the second pair consists of a red die `R2` and a blue die `B2`. The probability of observing one of the four sides of these dice is given by the matrices

$$[r'_1, r'_2, r'_3, r'_4]^T \text{ and } [b'_1, b'_2, b'_3, b'_4]^T.$$

Although the weights of each die can be different from the others, the red dice are indistinguishable from one another and the blue dice are indistinguishable from one another.

One day Oscar meets his friend Gabriella and asks if she would like to play an estimating game consisting of 100 rounds. Each round will consist of the following: Hidden from Gabriella's view, Oscar will flip the coin. If the coin lands on side 1, Oscar will select the first pair of dice to roll. If the coin lands on side 2, Oscar will select the second pair of dice to roll. Gabriella does not get to observe the coin. She only gets to observe the outcome of the pair of dice. After repeating this process 100 times Gabriella records in a 4×4 matrix $u = [u_{ij}]$, the number of times she observed the pair of dice having the outcome

$$\text{"red die on side } i \text{ and a blue die on side } j\text{"} \tag{1.1}$$

in the (i, j)-entry of u. For example, if Oscar rolled a 1 with the red die and simultaneously a 2 with the blue die 14 times among the 100 rolls, then Gabriella would have $u_{12} = 14$. We call the matrix u the *data*. After providing this data, Oscar will then ask Gabriella to estimate the true probability of observing the event (1.1). We denote this true probability distribution as the 4×4 matrix $[\overline{p}_{ij}]$ and denote Gabriella's estimate as $[\hat{p}_{ij}]$.

So what should Gabriella's estimate be? One guess might be to simply take the data u that was provided and rescale it so that the entries sum to 1. However, this guess would almost always be incredibly wrong. We can see this because $[\overline{p}_{ij}]$ must have rank at most 2 while u will almost always have full rank.

Indeed, we have that $[\overline{p}_{ij}]$ has the following decomposition as a sum of two rank 1 matrices:

$$\begin{bmatrix} \overline{p}_{11} & \overline{p}_{12} & \overline{p}_{13} & \overline{p}_{14} \\ \overline{p}_{21} & \overline{p}_{22} & \overline{p}_{23} & \overline{p}_{24} \\ \overline{p}_{31} & \overline{p}_{32} & \overline{p}_{33} & \overline{p}_{34} \\ \overline{p}_{41} & \overline{p}_{42} & \overline{p}_{43} & \overline{p}_{44} \end{bmatrix} = c_1 \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} [b_1, b_2, b_3, b_4] + c_2 \begin{bmatrix} r'_1 \\ r'_2 \\ r'_3 \\ r'_4 \end{bmatrix} [b'_1, b'_2, b'_3, b'_4].$$

This decomposition makes sense because the probability of observing, say, a red die on side 3 and a blue die on side 4 would be the probability of observing the coin on side one times $r_3 b_4$ plus the probability of observing the coin on side two times $r'_3 b'_4$. This means $[\overline{p}_{ij}]$ has rank at most 2, and more specifically, nonnegative rank 2. So Gabriella's estimate $[\hat{p}_{ij}]$ should be in the statistical model consisting of nonnegative rank 2 matrices. Now the question is how to determine the best estimate for $[\overline{p}_{ij}]$ constrained to this statistical model.

This is done by noticing that the likelihood of observing the data u for the probability distribution $[p_{ij}]$ is given by the *likelihood function*

$$\ell_u(p) = \frac{1}{p_{++}^{u_{++}}} \prod_{ij} {p_{ij}}^{u_{ij}}.$$

Here we employ the useful notation that

$$p_{++} := \sum_{ij} p_{ij} \text{ and } u_{++} := \sum_{ij} u_{ij}.$$

So the best estimate for the true distribution $[\overline{p}_{ij}]$ is the *maximum likelihood estimate*, the probability distribution that maximizes $\ell_u(p)$ on the statistical model.

Our approach to computing maximum likelihood estimates is to consider a relaxation of the problem. We will determine every critical point of $\ell_u(p)$ restricted to the regular points of the Zariski closure of a statistical model. The answer Gabriella will give to Oscar is a critical point of $\ell_u(p)$ restricted to set of rank 2 matrices. Gabriella will determine this critical point by solving a system of polynomials called the likelihood equations. This system has finitely many solutions, and we maximize $\ell_u(p)$ over this finite set of critical point to determine the maximum likelihood estimate.

Likelihood Equations

The set of nonnegative matrices of rank 2 is semi-algebraic, involving polynomial inequalities. However, the set of matrices of rank at most 2 is defined by equalities. For Gabriella, this means she is interested in determining the critical points of $\ell_u(p)$ on the zero set of the 3×3 minors of $[p_{ij}]$. There are 16 such minors $f_1, f_2, \ldots, f_{16}$ in 16 unknowns p_{ij}. To determine the critical points of $\ell_u(p)$ we want to determine when the gradient of $\ell_u(p)$ is in the row span of the gradients of the 16 minors evaluated at the point p. Up to scaling, the gradient of the likelihood function $\ell_u(p)$ equals

$$\left[\frac{u_{00}}{p_{00}} - \frac{u_{++}}{p_{++}} : \cdots : \frac{u_{44}}{p_{44}} - \frac{u_{++}}{p_{++}} \right]. \tag{1.2}$$

To get a system of equations, we augment (1.2) to the top of the matrix

$$\begin{bmatrix} \nabla f_1(p) \\ \nabla f_2(p) \\ \vdots \\ \nabla f_{16}(p) \end{bmatrix}.$$

If we take the 5×5 minors of this augmented matrix along with the polynomials $f_1, f_2, \ldots, f_{16}$, we get a system of equations. However there are two problems. First this is not a system of polynomial equations. It is a system of rational functions. To get a polynomial system, we have to subtly and carefully clear the denominators. Second, we are only interested in regular points, which means we want to avoid degenerate cases where we have a critical point with rank strictly less than 2. To handle this problem, we must saturate by the singular locus. After carefully clearing the denominators and handling this saturation process, we have likelihood equations.

The interesting aspect of these equations is that for almost any data u that Oscar generates, Gabriella will find that there are 191 complex solutions. That is, Gabriella will find that the likelihood function has 191 complex critical points with a subset of these points having positive coordinates that sum to one. The number of complex critical points of the likelihood function is called the *maximum likelihood degree* (ML degree) [9, 23] of the statistical model. The geometry of these numbers and formulations of systems of polynomial equations to compute them are at the heart of this thesis.

1.2 Results and contributions

This thesis introduces five main results and a software package that is an interface to solve systems of polynomial equations. The first main result is Theorem 2.2.1. This theorem is a new formulation of the likelihood equations that behaves well numerically. The procedure to determine likelihood equations presented in the introduction is vastly overdetermined and horrendous for numerical computations. With this new formulation, we solve a *square system* of polynomial equations, meaning the number of unknowns in the system equals the number of equations in the system. In our illustrating example Gabriella would need to only solve 16 equations in 16 unknowns with this new formulation.

The second main result are new computations of ML degrees (in bold below) as seen in Theorem 2.1.1. In terms of our illustrating example, r is the number of "sides" of the coin and number of pairs of dice. Also in terms of our illustrating example, m corresponds to the number of sides on the red dice, while n corresponds to the number of sides of the blue dice.

$(m,n) =$	$(3,3)$	$(3,4)$	$(3,5)$	$(4,4)$	$(4,5)$	$(4,6)$	$(5,5)$
$r=1$	1	1	1	1	1	1	1
$r=2$	10	26	**58**	**191**	**843**	**3119**	**6776**
$r=3$	1	1	1	**191**	**843**	**3119**	**61326**
$r=4$				1	1	1	**6776**
$r=5$							1

The 191 complex critical points that Gabriella computes correspond to the $r=2$, $(m,n) = (4,4)$ entry of the table.

From Table (2.4), one would conjecture that the vertical symmetry of a column to hold in general. Indeed, the third main result of the thesis, Theorem 3.2.4, proves this to be true. Theorem 3.2.4 also provides an explicit bijection between critical points of the likelihood function on $(m \times n)$ matrices of rank r with critical points of the likelihood function on $(m \times n)$ matrices of corank $r-1$. In terms of the illustrating example, there is a bijection

between critical points of the likelihood function on 4×4 matrices of rank 2 and critical points of the likelihood function on 4×4 matrices of rank 3. For a generic choice of $u = [u_{ij}]$, the bijection between

$$\begin{array}{l} [p_{ij}] \text{ a critical point of } \ell_u(p) \text{ on } 4 \times 4 \text{ rank 2 matrices, and} \\ [q_{ij}] \text{ a critical point of } \ell_u(p) \text{ on } 4 \times 4 \text{ rank 3 matrices} \end{array}$$

is given by the relation below.

$$\begin{bmatrix} p_{11}q_{11} & p_{12}q_{12} & p_{13}q_{13} & p_{14}q_{14} \\ p_{21}q_{21} & p_{22}q_{22} & p_{23}q_{23} & p_{24}q_{24} \\ p_{31}q_{31} & p_{32}q_{32} & p_{33}q_{33} & p_{34}q_{34} \\ p_{41}q_{41} & p_{42}q_{42} & p_{43}q_{43} & p_{44}q_{44} \end{bmatrix} = \frac{1}{u_{++}^3} \begin{bmatrix} u_{1+}u_{11}u_{+1} & u_{1+}u_{12}u_{+2} & u_{1+}u_{13}u_{+3} & u_{1+}u_{14}u_{+4} \\ u_{2+}u_{21}u_{+1} & u_{2+}u_{22}u_{+2} & u_{2+}u_{23}u_{+3} & u_{2+}u_{24}u_{+4} \\ u_{3+}u_{31}u_{+1} & u_{3+}u_{32}u_{+2} & u_{3+}u_{33}u_{+3} & u_{3+}u_{34}u_{+4} \\ u_{4+}u_{41}u_{+1} & u_{4+}u_{42}u_{+2} & u_{4+}u_{43}u_{+3} & u_{4+}u_{44}u_{+4} \end{bmatrix}$$

In addition, this bijection takes the critical point that maximizes the likelihood on the first model to the critical point that minimizes the likelihood on the second model.

The fourth main result is Theorem 4.2.5 that recasts maximum likelihood estimation in terms of conormal varieties. This elegant formulation allows one to define the dual likelihood equations and makes computing ML degrees of hyperdeterminants tractable.

The fifth main result is Algorithm 5.4.2. This algorithm uses the structure of the *Lagrange Likelihood Equations* to give lower bounds to the ML degree of a statistical model. This is done by considering what happens when zeros are in the presence of data.

The final contribution of the thesis is an introduction to an interface for the numerical algebraic software `Bertini` through the computational algebraic geometry software `Macaulay2`. This interface allows one to use the main tool in numerical algebraic geometry, *homotopy continuation*, via `Bertini` to solve system of polynomial equations.

Throughout the thesis various algebraic formulations of the maximum likelihood estimation problem will be given. With different formulations, computational results can be pushed further or theoretical results can be made clear. However, the key technique throughout the thesis is

using homotopy continuation to degenerate data.

Solving the likelihood equations for a choice of data is an extremely difficult problem. But if we can solve the likelihood equations for a generic choice of data $u_{generic}$, then we can solve the likelihood equations for any other specific choice of data v quickly. Using numerical algorithms such as Euler's method and Newton's method, we degenerate $u_{generic}$ to specific data v thereby degenerating the solutions of the likelihood equations with respect to $u_{generic}$ to solutions of the likelihood equations with respect to v. For more information on numerical algebraic geometry, the reader can jump straight to Chapter 6 or see [41, 3].

Chapter 2

Low Rank Matrix Models

The content of this chapter will be published in the *Journal of Algebraic Statistics* as an article titled *Maximum Likelihood for Matrices with Rank Constraints*, with minor changes throughout for consistency with other chapters. It is joint work with Jonathan Hauenstein and Bernd Sturmfels.

2.1 Introduction

Maximum likelihood estimation (MLE) is a fundamental computational task in statistics. A typical problem encountered in its applications is the occurrence of multiple local maxima. In order to be certain that a global maximum of the likelihood function has been achieved, one needs to locate all solutions to a system of polynomial equations. In this chapter we study these equations for two discrete random variables, having m and n states respectively. A joint probability distribution for two such random variables is written as an $m \times n$-matrix:

$$P \quad = \quad \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix}. \tag{2.1}$$

The entry p_{ij} represents the probability that the first variable is in state i and the second is in state j. Thus, the entries of P are non-negative and their sum p_{++} is 1. By a statistical model, we mean a closed subset $\mathcal{M}$ of the probability simplex Δ_{mn-1} of all such matrices P.

If i.i.d. samples are drawn from some P then we summarize the data also in a matrix

$$U \quad = \quad \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix}. \tag{2.2}$$

The entries of U are non-negative integers whose sum is u_{++}. As is customary in algebraic statistics [12, 23, 44], we write the *likelihood function* corresponding to the data matrix U as

$$\ell_U \quad = \quad \frac{\prod_{i=1}^{m} \prod_{j=1}^{n} p_{ij}^{u_{ij}}}{\left(\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \right)^{u_{++}}}. \tag{2.3}$$

This formula defines a rational function on the complex projective space $\mathbb{P}^{mn-1}$ whose restriction to the simplex Δ_{mn-1} is the usual likelihood function divided by a multinomial coefficient. The MLE problem is to find the global maximum of ℓ_U over the model $\mathcal{M}$.

Our model of interest is the set $\mathcal{M}_r$ of matrices P of rank $\leq r$. This is the intersection of the variety $\mathcal{V}_r \subset \mathbb{P}^{mn-1}$ defined by the $(r+1) \times (r+1)$-minors of P with Δ_{mn-1}. For generic U, the rational function ℓ_U has finitely many critical points on the determinantal variety $\mathcal{V}_r$. Their number is the *ML degree* of $\mathcal{V}_r$. In this chapter, we formulate a polynomial system consisting of mn equations in mn variables defining such critical points and compute them using methods from numerical algebraic geometry. That computation enables us to reliably find all local maxima of the likelihood function ℓ_U among positive points in $\mathcal{M}_r$. Among the computational results is the determination of the bold face numbers in the following table.

Theorem 2.1.1. *The known values for the ML degrees of the determinantal varieties $\mathcal{V}_r$ are*

$(m,n) =$	$(3,3)$	$(3,4)$	$(3,5)$	$(4,4)$	$(4,5)$	$(4,6)$	$(5,5)$
$r=1$	1	1	1	1	1	1	1
$r=2$	10	26	**58**	**191**	**843**	**3119**	**6776**
$r=3$	1	1	1	**191**	**843**	**3119**	**61326**
$r=4$				1	1	1	**6776**
$r=5$							1

(2.4)

The smaller numbers 10 and 26 had already been computed in [23, §5], but the symbolic computations using `Singular` that were presented in [23] had failed beyond the size 3×4.

In 2005, the third author offered a cash prize of 100 Swiss Francs (cf. [44, §3]) for the solution of a particular 4×4-instance that was described in [35, Example 1.16]. That prize was won in 2008 by Mingfu Zhu who solved this challenge in [48]. See also [42, Example 5.2] for a solution using `Singular`, and [13] for a statistical perspective on this problem. However, none of these papers had found the number 191 of critical points for the 4×4 cases.

That the column symmetry among the ML degrees always holds is the topic of the next chapter. The following is a special case of Theorem 3.1.1.

Theorem 2.1.2. *If $m \leq n$ then the ML degrees for rank r and for rank $m-r+1$ coincide.*

This chapter might appeal also to those interested in the topology of algebraic varieties. For a variety $\mathcal{V}$ in $\mathbb{P}^{mn-1}$, let $\mathcal{V}^0$ denote the open subset given by $p_{11}p_{12}\cdots p_{mn}p_{++} \neq 0$. Huh [26] recently proved that if $\mathcal{V}^0$ is smooth then the ML degree of $\mathcal{V}$ is equal to the signed Euler characteristic of $\mathcal{V}^0$. In our case, for $r \geq 2$, the open determinantal variety $\mathcal{V}_r^0$ is singular along $\mathcal{V}_{r-1}^0$, but a suitably modified statement is expected to be true. It might be speculated that the results in Theorems 2.1.1 and 2.1.2 will ultimately have a topological explanation. For more information one can also refer to [27].

The entries "1" of the table in (2.4) have easy explanations. For $r = m$ we have $\mathcal{V}_m = \mathbb{P}^{mn-1}$ and the unique critical point of the likelihood function ℓ_U is $P = \frac{1}{u_{++}}U$. The first row of (2.4) states that the independence model $\mathcal{M}_1$ has ML degree 1. This fact is well-known to statisticians, as the rank 1 matrix with entries $(u_{i+}u_{+j})/u_{++}^2$ is the unique critical point for ℓ_U on $\mathcal{V}_1^0$. We found it instructive to derive this fact from Huh's result [26, Theorem 1.(iii)]:

Example 2.1.3. Let $r = 1$. The Segre variety $\mathcal{V}_1 = \mathbb{P}^{m-1} \times \mathbb{P}^{n-1}$ is smooth. Fix coordinates $(x_1 : \cdots : x_m)$ on $\mathbb{P}^{m-1}$ and coordinates $(y_1 : \cdots : y_n)$ on $\mathbb{P}^{n-1}$. The open subset $\mathcal{V}_1^0$ consists of all points in $\mathbb{P}^{m-1} \times \mathbb{P}^{n-1}$ with $x_1 x_2 \cdots x_m y_1 y_2 \cdots y_n (x_1 + \cdots + x_m)(y_1 + \cdots + y_n) \neq 0$. Hence

$$\mathcal{V}_1^0 \quad = \quad (\mathbb{P}^{m-1} \text{ minus } m+1 \text{ hyperplanes}) \quad \times \ (\mathbb{P}^{n-1} \text{ minus } n+1 \text{ hyperplanes}).$$

Each factor has signed Euler characteristic 1, and hence so does their product. □

This chapter is organized as follows. In Section 2.2, we formulate the constraints that characterize critical points of ℓ_U on $\mathcal{V}_r$ as a square system of polynomial equations. The specific formulation in Theorem 2.2.1 is one of our key contributions. It is used to derive upper bounds in terms of m, n, and r. Theorem 2.2.3 extends our results to the case of symmetric matrices, and hence to mixtures of two identically distributed random variables.

Section 2.3 is devoted to our computations using numerical algebraic geometry. This furnishes valuable new tools for practitioners of statistics who are interested in exploring probability one algorithms for computing the global maximum of a given likelihood function.

In Section 2.4, we introduce a refined version of Theorem 2.1.2, proven in Chapter 3, and we summarize the computational evidence gathered to support it. The Galois group computations in Proposition 2.4.5 might be of independent interest. In Theorem 2.4.4, we present a proof of [48, Conjecture 11] by means of certified numerical computations.

Section 2.5 features the statistical view on our approach, and we explain how it differs from running the EM algorithm for discrete mixture models. The determinantal variety $\mathcal{V}_r$ is the Zariski closure of the latent variable model for r-fold mixtures of independent variables. They are equal in Δ_{mn-1} if and only if $r \leq 2$. For $r \geq 3$ this takes us to the real algebraic geometry problem, pioneered in [32], of distinguishing between rank and non-negative rank.

2.2 Equations and bounds

In this section, we present several formulations of the critical equations for the likelihood function on the determinantal variety $\mathcal{V}_r = \{\text{rank}(P) \leq r\}$. We view $\mathcal{V}_r$ as an affine variety in the space of matrices $\mathbb{C}^{m \times n}$ and we assume $m \leq n$. Our main result is Theorem 2.2.1 which expresses our problem as a square system of mn polynomial equations in mn unknowns.

An $m \times n$-matrix P is a regular point in the determinantal variety $\mathcal{V}_r$ if and only if $\text{rank}(P) = r$. If this holds then the tangent space T_P is a linear subspace of dimension $rn + rm - r^2$ in $\mathbb{C}^{m \times n}$, and its orthogonal complement (with respect to the standard inner product) is a linear subspace $T_P^\perp$ of dimension $(m-r)(n-r)$ in $\mathbb{C}^{m \times n}$.

Our input is a strictly positive data matrix U. We consider the logarithm of the likelihood function ℓ_U as in (2.3). The partial derivatives of the *log-likelihood function* $\log(\ell_U)$ are then

$$\frac{\partial \log(\ell_U)}{\partial p_{ij}} \quad = \quad \frac{u_{ij}}{p_{ij}} - \frac{u_{++}}{p_{++}}. \tag{2.5}$$

By [23, Proposition 3], a matrix P of rank r is a critical point for $\log(\ell_U)$ on $\mathcal{V}_r$ if and only if the linear subspace $T_P^\perp$ contains the $m \times n$-matrix whose (i,j) entry is (2.5). Hence the system of equations we seek to solve can be expressed in the following *geometric formulation*:

$$\text{rank}(P) = r\,, \qquad p_{++} = 1\,, \quad \text{and} \quad \text{the matrix } \bigl(u_{ij}/p_{ij} - u_{++}\bigr) \text{ lies in } T_P^\perp. \tag{2.6}$$

This is saying that the gradient of the objective function must be orthogonal to the tangent space of the variety at a critical point as in the elementary Lagrange multipliers method. When translating (2.6) into polynomial equations, we need to make sure to exclude matrices P of rank strictly less than r, as these are singular points in $\mathcal{V}_r$. We also need to exclude matrices P with $p_{ij} = 0$ for some (i, j). These non-degeneracy conditions require some care.

In [23], the following formulation was used to represent our problem. Let $J(P)$ denote the Jacobian matrix of the prime ideal defining $\mathcal{V}_r$. Since that ideal is minimally generated by the $\binom{m}{r+1}\binom{n}{r+1}$ subdeterminants of format $(r+1)\times(r+1)$, the Jacobian $J(P)$ is a matrix of format $\binom{m}{r+1}\binom{n}{r+1} \times mn$ whose entries are homogeneous polynomials of degree r. Let $[U]$ denote the matrix U when written as a row vector of format $1 \times mn$, and similarly $[P]$ is the vectorization of P. We write $\mathrm{diag}[P]$ for the diagonal $mn \times mn$-matrix with entries $p_{11}, p_{12}, \ldots, p_{mn}$. The following extended Jacobian has $2 + \binom{m}{r+1}\binom{n}{r+1}$ rows and mn columns:

$$\mathcal{J}(P) \quad = \quad \begin{pmatrix} [U] \\ [P] \\ J(P) \cdot \mathrm{diag}[P] \end{pmatrix}.$$

For a matrix P of rank r, the Jacobian $J(P)$ has rank $(m-r)(n-r) = \mathrm{codim}(\mathcal{V}_r)$. The third condition in (2.6) translates into the requirement that the span of the first two rows intersects the rowspace of $J(P) \cdot \mathrm{diag}[P]$. From this we derive the *rank formulation*

$$\mathrm{rank}(P) \leq r \quad \text{and} \quad \mathrm{rank}(\mathcal{J}(P)) \leq (m-r)(n-r)+1. \tag{2.7}$$

This formulation of our problem is elegant and is adapted to projective geometry in $\mathbb{P}^{mn-1}$. In terms of equations, we simply take the minors of size $r+1$ of the matrix P, and the minors of size $(m-r)(n-r)+2$ of the matrix $\mathcal{J}(P)$. However, this has two serious disadvantages: first, the number of minors is enormous, and second, we must get rid of extraneous solutions by saturation. Namely, to get rid of solutions P with $\mathrm{rank}(P) \leq r-1$, we need to saturate by the $r \times r$-minors of P, and to get rid of solutions on the boundary, we need to saturate by the product of linear forms $p_{11}p_{12}\cdots p_{mn}p_{++}$. This was done symbolically in [23, §4].

The calculation can be sped up a little bit by taking only $(m-r)(n-r)$ of the rows of $J(P)$, while also imposing the non-homogeneous equation $p_{++} = 1$. Finally, we can replace the first two rows of $J(P)$ by a single row $[U] - u_{++}[P]$ and require that the maximal minors of the resulting $((m-r)(n-r)+1) \times mn$-matrix be zero. This leads to improvements but is still far from sufficient to get to the full range of ML degrees reported in Theorem 2.1.1.

To get to those results, we pursue the following alternatives: first, we introduce new unknowns which allow us to replace the rank conditions by bilinear equations, and, second, we represent the subspace $T_P^\perp = \mathrm{rowspace}(J(P))$ using those same new unknowns. Let L be an $(m-r) \times m$-matrix of unknowns, let R be an $n \times (n-r)$-matrix of unknowns, and $\Lambda = (\lambda_{ij})$ an $(n-r) \times (m-r)$-matrix of unknowns. Then our *general kernel formulation* is:

$$p_{++} = 1, \quad L \cdot P = 0, \quad P \cdot R = 0, \quad \text{and} \quad P \star (R \cdot \Lambda \cdot L)^T + u_{++} \cdot P = U. \tag{2.8}$$

Here $A \star B$ denotes the Hadamard (entry-wise) product of two matrices of the same format. If the rows of L are linearly independent and the columns of R are linearly independent, then either of the conditions $L \cdot P = 0$ and $P \cdot R = 0$ suffice to imply that $\mathrm{rank}(P) \leq r$.

We now explain the last condition in (2.8). The space $T_P^\perp$ is spanned by the rank 1 matrices $(\rho_i \cdot \ell_j)^T$ where ρ_i is the i-th column of R and ℓ_j is the j-th row of L. Then

$$(R \cdot \Lambda \cdot L)^T \;=\; \sum_{i=1}^{n-r} \sum_{j=1}^{m-r} \lambda_{ij} (\rho_i \cdot \ell_j)^T$$

is a general matrix in $T_P^\perp$. The matrix $\left(u_{ij}/p_{ij} - u_{++}\right)$ in (2.6) can be written as

$$P^{\star(-1)} \star U - u_{++} \cdot \mathbf{1}. \tag{2.9}$$

Hence the last condition of (2.6) is equivalent to saying (2.9) equals $(R \cdot \Lambda \cdot L)^T$ for some Λ. We write this as $(R \cdot \Lambda \cdot L)^T + u_{++} \cdot \mathbf{1} = P^{\star(-1)} \star U$. We take the Hadamard product of both sides with the matrix P to get the last equation in (2.8). This operation is invertible since all entries of U are non-zero. Indeed, that last equation is $P \star \left((R \cdot \Lambda \cdot L)^T + u_{++} \cdot \mathbf{1}\right) = U$, and if this holds then all mn entries of the matrix P must be non-zero.

We conclude that (2.8) is a correct formulation of our problem provided we can ensure

$$\operatorname{rank}(L) = m - r, \quad \operatorname{rank}(R) = n - r, \quad \text{and} \quad \operatorname{rank}(P) = r.$$

We note that (2.8) is highly redundant as far as the number of variables is concerned. There are several ways to reduce that number. For instance, we can simply set $\lambda_{ij} = 1$ for all i, j. In addition, we can either replace L by a single row or replace R by a single column. Even after these simplifications, the critical points of ℓ_U on $\mathcal{V}_r$ are still represented faithfully.

After some experimentation, we found that the following simplification steps lead to the best computational results. Recall that $m \leq n$. Let P_1 be an $r \times r$-matrix of unknowns, let R_1 be an $r \times (n-r)$-matrix of unknowns, and let L_1 be an $(m-r) \times r$-matrix of unknowns. The matrix $\Lambda = (\lambda_{ij})$ is as before. Using this notation, we take (2.8) with

$$L = \begin{pmatrix} L_1 & -I_{m-r} \end{pmatrix}, \quad P = \begin{pmatrix} P_1 & P_1 R_1 \\ L_1 P_1 & L_1 P_1 R_1 \end{pmatrix}, \quad \text{and} \quad R = \begin{pmatrix} R_1 \\ -I_{n-r} \end{pmatrix}, \tag{2.10}$$

where I_{m-r} and I_{n-r} are identity matrices. We call (2.8) with (2.10) the *local kernel formulation* of our problem. Note that the constraints $L \cdot P = 0$, $P \cdot R = 0$, $\operatorname{rank}(L) = m - r$, and $\operatorname{rank}(R) = n - r$ are automatically satisfied in this formulation. The condition $\operatorname{rank}(P) = r$ is also implied for every solution provided U is generic. Finally, the equation $p_{++} = 1$ can be removed from (2.8) in this formulation since $p_{++} = 1$ is equivalent to the sum of all mn equations given by $P \star (R \cdot \Lambda \cdot L)^T + u_{++} \cdot P = U$. By counting equations and unknowns, we now see that our system is a square system consisting of mn equations in mn unknowns.

Theorem 2.2.1. *Let U be a generic $m \times n$ data matrix with $m \leq n$. The polynomial system*

$$P \star (R \cdot \Lambda \cdot L)^T + u_{++} \cdot P \;=\; U \tag{2.11}$$

consists of mn equations in mn unknowns given by (2.10). It has finitely many complex solutions (P_1, L_1, R_1, Λ), and the corresponding $m \times n$-matrices P defined by (2.10) are precisely the critical points of the likelihood function ℓ_U on the determinantal variety $\mathcal{V}_r$.

Since the column sums of $P \star (R \cdot \Lambda \cdot L)^T$ are zero, we can further simplify the n equations. For the first m columns, we replace each entry on the diagonal with the column sum. For the last $n-m$ columns, we replace the last entry in the column with the column sum.

Example 2.2.2. To illustrate the local kernel formulation (2.11), we consider $m = n = 3$ with the two subcases $r = 1$ and $r = 2$. Both have nine equations in nine unknowns.

Subcase $r = 1$: The nine unknowns are the entries in the matrices

$$L_1 = \begin{pmatrix} l_{11} \\ l_{21} \end{pmatrix}, \quad P_1 = \begin{pmatrix} p_{11} \end{pmatrix}, \quad R_1 = \begin{pmatrix} r_{11} & r_{12} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix},$$

and the nine equations from (2.11) take the form

$$\begin{aligned}
p_{11}(1 + l_{11} + l_{21}) &= (u_{11} + u_{21} + u_{31})/u_{++} \\
p_{11}r_{11}(u_{++} - l_{11}\lambda_{11} - l_{21}\lambda_{12}) &= u_{12} \\
p_{11}r_{12}(u_{++} - l_{11}\lambda_{21} - l_{21}\lambda_{22}) &= u_{13} \\
p_{11}l_{11}(u_{++} - r_{11}\lambda_{11} - r_{12}\lambda_{21}) &= u_{21} \\
p_{11}r_{11}(1 + l_{11} + l_{21}) &= (u_{12} + u_{22} + u_{32})/u_{++} \\
p_{11}l_{11}r_{12}(\lambda_{21} + u_{++}) &= u_{23} \\
p_{11}l_{21}(u_{++} - r_{11}\lambda_{12} + r_{12}\lambda_{22}) &= u_{31} \\
p_{11}l_{21}r_{11}(\lambda_{12} + u_{++}) &= u_{32} \\
p_{11}r_{12}(1 + l_{11} + l_{21}) &= (u_{13} + u_{23} + u_{33})/u_{++}.
\end{aligned}$$

This system has a unique solution which writes the unknowns as rational functions in the u_{ij}.

Subcase $r = 2$: The nine unknowns are the entries in the matrices

$$L_1 = \begin{pmatrix} l_{11} & l_{12} \end{pmatrix}, \quad P_1 = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \quad R_1 = \begin{pmatrix} r_{11} \\ r_{21} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_{11} \end{pmatrix},$$

and the nine equations take the form

$$\begin{aligned}
p_{11}(1 + l_{11}) + p_{21}(1 + l_{12}) &= (u_{11} + u_{21} + u_{31})/u_{++} \\
p_{12}(l_{11}r_{21}\lambda_{11} + u_{++}) &= u_{12} \\
(p_{11}r_{11} + p_{12}r_{21})(u_{++} - l_{11}\lambda_{11}) &= u_{13} \\
p_{21}(l_{12}r_{11}\lambda_{11} + u_{++}) &= u_{21} \\
p_{12}(1 + l_{11}) + p_{22}(1 + l_{12}) &= (u_{12} + u_{22} + u_{32})/u_{++} \\
(p_{21}r_{11} + p_{22}r_{21})(u_{++} - l_{12}\lambda_{11}) &= u_{23} \\
(p_{11}l_{11} + p_{21}l_{12})(u_{++} - r_{11}\lambda_{11}) &= u_{31} \\
(p_{12}l_{11} + p_{22}l_{12})(u_{++} - r_{21}\lambda_{11}) &= u_{32} \\
(p_{11}r_{11} + p_{12}r_{21})(1 + l_{11}) + (p_{21}r_{11} + p_{22}r_{21})(1 + l_{12}) &= (u_{13} + u_{23} + u_{33})/u_{++}.
\end{aligned}$$

This system has ten complex solutions for a generic data matrix U. In other words, the 9 unknowns $l_{..}, p_{..}, r_{..}$ and λ_{11} are algebraic functions of degree 10 in $u_{11}, u_{12}, \ldots, u_{33}$. □

Upper bounds on the ML degree of $\mathcal{V}$ arise from our formulation. The Bézout bound is

$$2^r \cdot 3^{n-r} \cdot 4^{n(m-1)}.$$

(m,n,r)	$(3,3,1)$	$(3,3,2)$	$(3,4,1)$	$(3,4,2)$	$(3,5,1)$	$(3,5,2)$
Bézout	73728	49152	3538944	2359296	169869312	113246208
4-hom	270	1350	840	29400	2025	378000
linear product	172	1018	374	20844	650	68586
polyhedral	6	53	10	472	15	2724
ML Degree	1	10	1	26	1	58

(m,n,r)	$(4,4,1)$	$(4,4,2)$	$(4,4,3)$	$(4,5,1)$	$(4,5,2)$	$(4,5,3)$
Bézout	905969664	603979776	402653184	173946175488	115964116992	77309411328
4-hom	17600	7276500	580800	63700	323723400	115615500
linear product	5690	4791168	224598	13560	165869606	58335270
polyhedral	20	15280	2847	35	241218	145273
ML Degree	1	191	191	1	843	843

Table 2.1: Comparison of upper bounds for selected (m,n,r)

If we consider (P_1, L_1, R_1, Λ) in the product space $\mathbb{C}^{r^2} \times \mathbb{C}^{r(m-r)} \times \mathbb{C}^{r(n-r)} \times \mathbb{C}^{(n-r)(m-r)}$, our system consists of r equations of degree $(1,1,0,0)$, $n-r$ equations of degree $(1,1,1,0)$, and $n(m-1)$ equations of degree $(1,1,1,1)$. The associated 4-homogeneous Bézout bound is the coefficient of the monomial $w^{r^2} \cdot x^{r(m-r)} \cdot y^{r(n-r)} \cdot z^{(n-r)(m-r)}$ in the expression

$$(w+x)^r \cdot (w+x+y)^{n-r} \cdot (w+x+y+z)^{n(m-1)}.$$

A refinement of the 4-homogeneous bound using the fact that each polynomial only depends upon a subset of the variables yields a *linear product bound* [47]. Finally, the *polyhedral root count* exploits the sparsity of the monomials in our system. We computed the polyhedral bound for various cases using `MixedVol` [15] in `PHC` [46]. All of the aforementioned bounds are presented in Table 2.1 for selected values of m, n, and r. When solving a polynomial system using homotopies built from these bounds, one must balance the added computational cost required for the tighter bound with the computational savings arising from that bound.

We close this section by discussing rank constraints on symmetric matrices of the form

$$P \quad = \quad \begin{pmatrix} 2p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{12} & 2p_{22} & p_{23} & \cdots & p_{2n} \\ p_{13} & p_{23} & 2p_{33} & \cdots & p_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{1n} & p_{2n} & p_{3n} & \cdots & 2p_{nn} \end{pmatrix}. \tag{2.12}$$

The case $n = 3$ was treated in [23, Example 12] where its ML degree was found to be 6. It is essential that the unknowns p_{ii} on the diagonal are multiplied by 2 before imposing the rank constraints. The matrices (2.12) of rank one form a Veronese variety in $\mathbb{P}^{(n+2)(n-1)/2}$. This variety has ML degree 1 and represents the independence model for two identically distributed random variables on n states. The case $n = 2$ is the Hardy-Weinberg curve [35, Figure 3.1]. Larger ranks r correspond to the secant varieties of this Veronese variety.

Theorem 2.2.3. *The known values for the ML degrees of rank r symmetric matrices (2.12) are*

$$\begin{array}{rccccc} n = & 3 & 4 & 5 & 6 \\ r=1 & 1 & 1 & 1 & 1 \\ r=2 & 6 & \mathbf{37} & \mathbf{270} & \mathbf{2341} \\ r=3 & 1 & \mathbf{37} & \mathbf{1394} & \\ r=4 & & 1 & \mathbf{270} & \\ r=5 & & & 1 & \mathbf{2341} \\ r=5 & & & & 1 \end{array} \tag{2.13}$$

Our input is a strictly positive symmetric $n \times n$-matrix U. The likelihood function equals

$$\ell_U \;\; = \;\; \frac{\prod_{i \leq j} p_{ij}^{u_{ij}}}{\left(\sum_{i \leq j} p_{ij}\right)^{\sum_{i \leq j} u_{ij}}}. \tag{2.14}$$

In the statistical context, when the sum of the p_{ij} entries equals 1, we have

$$\frac{\partial \log(\ell_U)}{\partial p_{ij}} \;\; = \;\; \frac{u_{ij}}{p_{ij}} - \sum_{i \leq j} u_{ij}. \tag{2.15}$$

We compute the critical points on the variety of rank r matrices (2.12) by adapting the formulation in Theorem 2.2.1. Let P_1 be a symmetric $r \times r$-matrix of unknowns where the diagonal entries are multiplied by 2 similar to (2.12), let L_1 be an $(n-r) \times r$-matrix of unknowns, and Λ be a symmetric $(n-r) \times (n-r)$-matrix. Following (2.10), we define

$$L = \begin{pmatrix} L_1 & -I_{m-r} \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} P_1 & P_1 L_1^T \\ L_1 P_1 & L_1 P_1 L_1^T \end{pmatrix}. \tag{2.16}$$

To account for the p_{ii}'s not being multiplied by 2 in the likelihood function, let D be the $n \times n$-matrix whose diagonal entries are 2 and off-diagonal entries are 1. The *symmetric local kernel formulation* is the square system consisting of the upper triangular part of

$$P \star (L^T \cdot \Lambda \cdot L) + \sum_{i \leq j} u_{ij} \cdot P \;\; = \;\; D \star U. \tag{2.17}$$

This is a system of $n(n+1)/2$ equations in $n(n+1)/2$ unknowns. Similar to the local kernel formulation, the column sums of $P \star (L^T \cdot \Lambda \cdot L)$ are zero. Hence (2.17) implies $\sum_{i \leq j} p_{ij} = 1$. We use this fact to replace the diagonal entries in (2.17) with the corresponding column sum.

Example 2.2.4. We illustrate the symmetric local kernel formulation (2.17) for the two subcases $r = 1, 2$ when $n = 3$. Both have 6 equations in 6 unknowns. Here, $u_{++} = \sum_{i \leq j} u_{ij}$. *Subcase* $r = 1$: The six unknowns arise from the entries in the matrices

$$L_1 = \begin{pmatrix} l_{11} \\ l_{21} \end{pmatrix}, \quad P_1 = \begin{pmatrix} 2p_{11} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix},$$

and the six equations take the form

$$\begin{aligned} 2p_{11}(1+l_{11}+l_{21}) &= (2u_{11}+u_{12}+u_{13})/u_{++} \\ 2p_{11}l_{11}(u_{++}-l_{11}\lambda_{11}-l_{21}\lambda_{12}) &= u_{12} \\ 2p_{11}l_{21}(u_{++}-l_{11}\lambda_{12}-l_{21}\lambda_{22}) &= u_{13} \\ 2p_{11}l_{11}(1+l_{11}+l_{21}) &= (u_{12}+2u_{22}+u_{23})/u_{++} \\ 2p_{11}l_{11}l_{21}(\lambda_{12}+u_{++}) &= u_{23} \\ 2p_{11}l_{21}(1+l_{11}+l_{21}) &= (u_{13}+u_{23}+2u_{33})/u_{++}. \end{aligned}$$

This system has a unique solution which writes the unknowns as rational functions in the u_{ij}.

Subcase $r = 2$: The six unknowns arise from the entries in the matrices

$$L_1 = \begin{pmatrix} l_{11} & l_{12} \end{pmatrix}, \quad P_1 = \begin{pmatrix} 2p_{11} & p_{12} \\ p_{12} & 2p_{22} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_{11} \end{pmatrix},$$

and the six equations take the form

$$\begin{aligned} 2p_{11}(1+l_{11})+p_{12}(1+l_{12}) &= (2u_{11}+u_{12}+u_{13})/u_{++} \\ p_{12}(l_{11}l_{12}\lambda_{11}+u_{++}) &= u_{12} \\ (2p_{11}l_{11}+p_{12}l_{12})(u_{++}-l_{11}\lambda_{11}) &= u_{13} \\ p_{12}(1+l_{11})+2p_{22}(1+l_{12}) &= (u_{12}+2u_{22}+u_{23})/u_{++} \\ (p_{12}l_{11}+2p_{22}l_{12})(u_{++}-l_{12}\lambda_{11}) &= u_{23} \\ (2p_{11}l_{11}+p_{12}l_{12})(1+l_{11})+(p_{12}l_{11}+2p_{22}l_{12})(1+l_{12}) &= (u_{13}+u_{23}+2u_{33})/u_{++}. \end{aligned}$$

This system has six complex solutions for a general data matrix U. In the other words, the 6 unknowns $l_{\cdot\cdot}, p_{\cdot\cdot}$, and λ_{11} are algebraic functions of degree 6 in $u_{11}, u_{12}, \ldots, u_{33}$. □

Here is the symmetric version of Theorem 2.1.2, as suggested by Theorem 2.2.3:

Theorem 2.2.5. *The ML degree for symmetric $n \times n$-matrices (2.12) of rank r is equal to the ML degree for symmetric $n \times n$-matrices (2.12) of rank $n - r + 1$.*

The proof of this statement is given by Theorem 3.3.4 of the next chapter.

2.3 Solutions using numerical algebraic geometry

Theorems 2.1.1 and 2.2.3 document considerable advances relative to the computational results found earlier by Hoşten, Khetan, and Sturmfels [23, §5]. In this project, we used numerical algebraic geometry [3] to compute the ML degrees by solving the local kernel formulation (2.11) which we explain in this section.

The statistical problem addressed here is to find the global maximum of a likelihood function ℓ_U over a matrix model $\mathcal{M}$ given by rank constraints. For this class of problems, the use of numerical algebraic geometry has the following significant advantage over symbolic computations. After having solved the likelihood equations only once, for one generic data matrix U_0, all subsequent computations for other data matrices U are much faster. Numerical homotopy continuation will start from the critical points of ℓ_{U_0} and transform them into the

(m,n,r)	$(4,4,2)$	$(4,4,3)$	$(4,5,2)$	$(4,5,3)$	$(5,5,2)$	$(5,5,4)$
Preprocessing	257	427	1938	2902	348555	146952
Solving	4	4	20	20	83	83

Table 2.2: Running times for preprocessing and subsequent solving (in seconds)

critical points of ℓ_U. Intuitively speaking, for a fixed model $\mathcal{M}$, *the homotopy amounts to changing the data*. We believe that our methodology will be useful for a wider range of maximum likelihood problems than those treated here, and we decidedly agree with the statement of Buot and Richards [8, §5] that *"... homotopy continuation algorithms often provide substantial advantages over iterative methods commonly used in statistics"*.

We discuss below two options for the preprocessing stage of solving the local kernel formulation (2.11) for generic U_0. The first option is to use a single homotopy built from an upper bound discussed in Section 2.2, most notably a polyhedral homotopy built from the polyhedral root count. The second option is to use a sequence of homotopies that intersect the hypersurfaces corresponding to each equation, most notably via regeneration [21].

Parallel computation is an essential feature of numerical algebraic geometry. Both preprocessing, by solving a generic data set once, and each subsequent solve for given specific data can be performed in parallel. In our case, we used a 64-bit Linux cluster with 160 processors to perform the computations summarized in Table 2.2 which tracked each path on a separate processor. For instance, for $(m,n,r) = (4,5,2)$, there are 843 paths, to be distributed among the 160 processors. Using adaptive precision [4], this takes 20 seconds while the same computation performed sequentially takes about 20 minutes on a typical laptop.

Example 2.3.1. The following data matrix is attributed to the fictional character DiaNA in [35, Example 1.3]. It represents her alignment of two DNA sequences of length $u_{++} = 40$:

$$U \quad = \quad \begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix}.$$

According to Table 2.2, it took 257 seconds to solve the first instance for $(m,n,r) = (4,4,2)$, but now every subsequent run takes only 4 seconds. In that solving step, the integers u_{ij} become parameters over the complex numbers. For DiaNA's data matrix U, the 191 complex critical points degenerate to 25 real critical points, each of which is positive, and 166 nonreal critical points. See Theorem 2.4.4 for additional information regarding the critical points. □

Three advantages of the local kernel formulation (2.11) are that it is a square system with polynomials of degree at most 4, it is sparse in terms of the number of monomials appearing, and it has a natural product structure. These structures are clearly visible from the systems in Example 2.2.2, and they are used to derive the smaller upper bounds in Table 2.1. In what follows, we shall describe our preprocessing and how we can use its output to easily compute all critical points of ℓ_U for a given data matrix U. We also analyze some specific examples. An introduction to numerical algebraic geometry and homotopy continuation can

be found in [41] and more details using `Bertini` to perform these computations appear in the book [3] and Chapter 6.

For a square polynomial system F, *basic homotopy continuation* computes a finite set $\mathcal{S}$ of complex roots of F which contains all isolated roots. Here, "computes $\mathcal{S}$" means numerically computing the coordinates of each point in $\mathcal{S}$, and to be able to approximate these to arbitrary accuracy. Numerical approximations to nonsingular solutions can be certified using the software `alphaCertified` [22]. This certification can also determine if the solution is real or positive. To compute $\mathcal{S}$, we first construct a family of polynomial systems $\mathcal{F}$ containing F and then compute the isolated roots for a sufficiently general $G \in \mathcal{F}$. Finally, one tracks the solution paths starting with the isolated roots as G deforms to F inside $\mathcal{F}$.

Fix (m, n, r) and let $\mathcal{F} := \mathcal{F}_{m,n,r}$ be the family of polynomial systems (2.11) for $U \in \mathbb{C}^{m\times n}$. The generic root count on $\mathcal{F}$ is the ML degree of $\mathcal{V}_r$. In particular, for any generic $U_0 \in \mathbb{C}^{m\times n}$ the number of roots of the corresponding system $F_{U_0} \in \mathcal{F}$ is the ML degree of $\mathcal{V}_r$. Suppose further that we know the roots of F_{U_0}. Then, for any matrix $U \in \mathbb{C}^{m\times n}$, we can compute the isolated roots of the corresponding polynomial system F_U by tracking the ML degree number of solutions paths starting with the roots of F_{U_0} as U_0 and F_{U_0} deform to U and F_U.

Since the family $\mathcal{F}$ is parameterized by the linear space $\mathbb{C}^{m\times n} \cong \mathbb{R}^{2mn}$, we can connect U_0 to U along a line segment. If U_0 is not in a sufficiently general position with respect to U, e.g., both real, this segment may contain matrices for which the corresponding system has a root count that is different from the ML degree. To avoid this, we apply the *gamma trick* of [33]. For $\gamma \in \mathbb{S}^1 \subset \mathbb{C}^*$, the trick deforms from U_0 to U along the arc parameterized by

$$\frac{\gamma t}{1+(\gamma-1)t} \cdot U_0 \;+\; \frac{1-t}{1+(\gamma-1)t} \cdot U \quad \text{for} \quad t \in [0,1]. \tag{2.18}$$

For all but finitely many values $\gamma \in \mathbb{S}^1$, the root count for the corresponding polynomial system along this arc, except possibly at U when $t = 0$, is the ML degree.

We conclude our discussion on deforming from a known set of critical points with a practical issue. Due to choices of affine patches, the local kernel formulation (2.11), as written, is not suitable for a nongeneric data matrix U. Once given a data matrix U, we simply choose random affine patches as in [2]. Let $O_1, O_2 \in \mathbb{R}^{r\times r}$, $O_3 \in \mathbb{R}^{m\times m}$, and $O_4 \in \mathbb{R}^{n\times n}$ be random orthogonal matrices and L_1, P_1, R_1, and Λ be as before. Then, we use (2.11) with

$$L = O_1 \cdot \begin{pmatrix} L_1 & -I_{m-r} \end{pmatrix} \cdot O_3^T, \;\; P = O_3 \cdot \begin{pmatrix} P_1 & P_1R_1 \\ L_1P_1 & L_1P_1R_1 \end{pmatrix} \cdot O_4^T, \;\text{ and }\; R = O_4 \cdot \begin{pmatrix} R_1 \\ -I_{n-r} \end{pmatrix} \cdot O_2^T.$$

The homotopy (2.18) quickly computes the isolated critical points for any given data matrix U provided that we already know the critical points for a sufficiently general data matrix U_0.

We now discuss the two options for *preprocessing* mentioned above, namely polyhedral homotopies and regeneration. A summary of our computations with these two methods, now using serial processing with double precision, are presented in Table 2.3. The last pair of entries suggest that the two methods exhibit complementary behavior with respect to the duality of Theorem 2.1.2. In both cases, 191 roots are found as predicted by Theorem 2.4. These are essentially the same roots, by Theorem 2.4.2 below. For instance, using

(m,n,r)	$(3,3,2)$	$(3,4,2)$	$(3,5,2)$	$(4,4,2)$	$(4,4,3)$
Polyhedral using `PHC`	4	120	2017	23843	1869
Regeneration using `Bertini`	6	61	188	2348	7207

Table 2.3: Running times for preprocessing in serial using double precision (in seconds)

polyhedral homotopy, the rank 2 case can be solved in 1869 seconds and then we may read off the solutions for rank 3 using (2.19).

The first approach to solve the equations for U_0 is to use basic homotopy continuation in the family $\mathcal{P}$ of polynomial systems that arise from some relevant structure. The generic root count on $\mathcal{P}$ constructed from various structures are presented in Table 2.1. After computing the roots for a general element of $\mathcal{P}$, we return to basic homotopy continuation for computing the roots of F_{U_0}. Table 2.3 summarizes the results of using a polyhedral approach implemented in `PHC` [46] where the family $\mathcal{P}$ is constructed based on the Newton polytopes of the given equations.

The second approach is based on intersecting the given hypersurfaces iteratively. This can be advantageous when the degree of the intersection is significantly less than the product of the degrees. To be explicit, if $\mathcal{S}$ is a pure k-dimensional variety ($k > 0$) and $\mathcal{H}$ is a hypersurface, intersection approaches can be advantageous when the degree of the pure $(k-1)$-dimensional part of $\mathcal{S} \cap \mathcal{H}$ is less than $\deg \mathcal{S} \cdot \deg \mathcal{H}$. Regeneration is an intersection approach that builds from a product structure of the given system. We shall now discuss this.

We first consider the classical idea of solving polynomial systems using successive intersections and then discuss how to build from a product structure. Consider N polynomials $f_1, \ldots, f_N$ in N variables, defining hypersurfaces $\mathcal{H}_1, \ldots, \mathcal{H}_N$. One advantage of a square system is that the isolated solutions of $f_1 = \cdots = f_N = 0$ arise by computing the codimension i components of $\mathcal{H}_1 \cap \cdots \cap \mathcal{H}_i$ sequentially for $i = 1, 2, \ldots, N$. In fact, every codimension $i+1$ component of $\mathcal{H}_1 \cap \cdots \cap \mathcal{H}_i \cap \mathcal{H}_{i+1}$ arises as the intersection of a codimension i component C of $\mathcal{H}_1 \cap \cdots \cap \mathcal{H}_i$ and the hypersurface $\mathcal{H}_{i+1}$, where C is not contained in $\mathcal{H}_{i+1}$.

The use of the product structure arises from intersecting an algebraic set of pure codimension i with a linear space of dimension i yielding finitely many points. The first step is a hypersurface intersected with a line. If $\mathcal{L}_2, \ldots, \mathcal{L}_N$ are general hyperplanes, the hypersurface $\mathcal{H}_1$ is represented by the isolated points in $\mathcal{H}_1 \cap \mathcal{L}_2 \cap \cdots \cap \mathcal{L}_N$. Such points can be computed by solving a univariate polynomial, namely f_1 restricted to the line $\mathcal{L}_2 \cap \cdots \cap \mathcal{L}_N$. Let $1 \leq i < N$ and C_i be the pure one-dimensional component of $\mathcal{H}_1 \cap \cdots \cap \mathcal{H}_i \cap \mathcal{L}_{i+2} \cap \cdots \cap \mathcal{L}_N$. Now, *basic regeneration* computes $C_i \cap \mathcal{H}_{i+1}$ from $C_i \cap \mathcal{L}_{i+1}$ as follows. Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be hyperplanes defined by sufficiently general linear polynomials $\ell_1, \ldots, \ell_k$ that represent a linear product decomposition of f_{i+1}. Let $\mathcal{M} = \bigcup_{j=1}^{k} \mathcal{M}_j$. Basic homotopy continuation computes $C_i \cap \mathcal{M}_j$ from $C_i \cap \mathcal{L}_{i+1}$ for $j = 1, \ldots, k$. Their union is $C_i \cap \mathcal{M}$. Applying basic homotopy continuation once more yields $C_i \cap \mathcal{H}_{i+1}$ by deforming from $C_i \cap \mathcal{M}$.

For the preprocessing approaches above, we can certify that the set of approximations obtained correspond to distinct solutions using `alphaCertified`. At each stage of the regeneration and at the end of the computation, we can perform one additional test to confirm that we have obtained all of the solutions: the trace test [39]. During regeneration, the centroid of the solutions must move linearly as the hyperplane $\mathcal{L}_N$ is moved linearly.

Moreover, the centroid of the critical $m \times n$-matrices must move linearly as the data matrix U moves linearly. With these tests, we are able to claim, with high probability, that our initial randomly selected data matrix U_0 was sufficiently generic, and Theorems 2.1.1 and 2.2.3 hold.

After computing the positive critical points for a given data matrix U, we identify the local maximizers by analyzing the Hessian of the corresponding Lagrangian function, namely

$$L(P, \lambda) \;=\; \log \ell_U(P) + \sum_{i=1}^{k} \lambda_i g_i(P),$$

where $\mathcal{V}_r$ is defined by the vanishing of the polynomials $g_1, \ldots, g_k$. If P is a critical point of rank r, let $\lambda \in \mathbb{C}^k$ be the unique vector such that $\nabla L(P, \lambda) = 0$. Then, P is a local maximizer if the matrix $N^T \cdot HL(P, \lambda) \cdot N$ is negative semidefinite where $HL(P, \lambda)$ is the Hessian of L and the columns of N form a basis for the tangent space of $\mathcal{V}_r \times \mathbb{C}^k$ at (P, λ).

In the remainder of this section we present three concrete numerical examples.

Example 2.3.2. We consider the symmetric matrix model (2.12) for $n = 3$ with the data

$$u_{11} = 10,\ u_{12} = 9,\ u_{13} = 1,\ u_{22} = 21,\ u_{23} = 3,\ u_{33} = 7.$$

All six critical points of the likelihood function (2.14) are real and positive. They are

p_{11}	p_{12}	p_{13}	p_{22}	p_{23}	p_{33}	$\log \ell_U(p)$
0.1037	0.3623	0.0186	0.3179	0.0607	0.1368	−82.18102
0.1084	0.2092	0.1623	0.3997	0.0503	0.0702	−84.94446
0.0945	0.2554	0.1438	0.3781	0.4712	0.0810	−84.99184
0.1794	0.2152	0.0142	0.3052	0.2333	0.0528	−85.14678
0.1565	0.2627	0.0125	0.2887	0.2186	0.0609	−85.19415
0.1636	0.1517	0.1093	0.3629	0.1811	0.0312	−87.95759

The first three points are local maxima in Δ_5 and the last three points are local minima. These six points define an extension of degree 6 over $\mathbb{Q}$. For instance, via `Macaulay 2` [16], the minimal polynomial for the last coordinate is

$$\begin{gathered} 9528773052286944 p_{33}^6 - 4125267629399052 p_{33}^5 + \\ 713452955656677 p_{33}^4 - 63349419858182 p_{33}^3 + \\ 3049564842009 p_{33}^2 - 75369770028 p_{33} + 744139872. \end{gathered}$$

As we shall see in Proposition 2.4.5, the Galois group of this irreducible polynomial is solvable. So we can express each of the coordinates in radicals. For example, the last coordinate, via `RadiRoot` [10], is

$$\begin{aligned} p_{33} \;=\; & \left(\tfrac{14779904193}{211433981207339}\zeta^2 - \tfrac{14779904193}{211433981207339}\zeta\right)\omega_1\omega_2^2 \\ & - \tfrac{66004846384302}{19221271018849}\omega_2^2 + \tfrac{16427}{227664} + \tfrac{1}{12}(\zeta - \zeta^2)\,\omega_2 + \tfrac{1}{2}\omega_3, \end{aligned}$$

where ζ is a primitive third root of unity, $\omega_1^2 = 94834811/3$, and

$$\begin{aligned} \omega_2^3 \;=\; & \left(\tfrac{5992589425361}{150972770845322208}\zeta - \tfrac{5992589425361}{150972770845322208}\zeta^2\right) + \tfrac{97163}{40083040181952}\omega_1, \\ \omega_3^2 \;=\; & \tfrac{5006721709}{1248260766912} + \left(\tfrac{212309132509}{4242035935404}\zeta - \tfrac{212309132509}{4242035935404}\zeta^2\right)\omega_2 - \tfrac{2409}{20272573168}\omega_1\omega_2 \\ & - \tfrac{158808750548335}{76885084075396}\omega_2^2 + \left(\tfrac{17063004159}{422867962414678}\zeta^2 - \tfrac{17063004159}{422867962414678}\zeta\right)\omega_1\omega_2^2. \end{aligned}$$

We finally note that the six critical points can be matched into three pairs so that (2.19) holds: the Hadamard product of points 1 and 6 agree with that of points 2 and 5, and that of points 3 and 4. Thus this example illustrates the symmetric matrix version of Theorem 2.4.2. □

Example 2.3.3. Let $m = 4, n = 5$ and consider the data matrix

$$U \quad = \quad \begin{pmatrix} 2084 & 1 & 1 & 1 & 4 \\ 4 & 23587 & 5 & 3 & 1 \\ 6 & 3 & 41224 & 3 & 2 \\ 4 & 6 & 2 & 8734 & 4 \end{pmatrix}.$$

For $r = 2$ and $r = 3$, this instance has the expected number 843 of distinct complex critical points. In both cases, 555 critical points are real, and 25 of these are positive. Consider the 25 critical points in Δ_{19}. For $r = 2$ precisely seven are local maxima, and for $r = 3$ precisely six are local maxima. We shall list them explicitly in Examples 2.5.3 and 2.5.4 respectively. □

Example 2.3.4. Let $m = n = 5$, with the non-symmetric model, and consider the data

$$U \quad = \quad \begin{pmatrix} 2864 & 6 & 6 & 3 & 3 \\ 2 & 7577 & 2 & 2 & 5 \\ 4 & 1 & 7543 & 2 & 4 \\ 5 & 1 & 2 & 3809 & 4 \\ 6 & 2 & 6 & 3 & 5685 \end{pmatrix}.$$

For $r = 2$ and $r = 4$, this instance has the expected number of 6776 distinct complex critical points. In both cases, 1774 of these are real and 90 of these are real and positive. This illustrates the last statement in Theorem 2.4.2. The number of local maxima for $r = 2$ equals 15, and the number of local maxima for $r = 4$ equals 6. For $r = 3$, we have 61326 critical points, of which 15450 are real. Of these, 362 are positive and 25 are local maxima. □

2.4 Further results and computations

The numerical algebraic geometry techniques described in Section 2.3 have the advantage that they permit fast experimentation with non-trivial instances. This led us to a variety of conjectures, including those concerning ML duality. Before we come to our discussion of duality, we briefly state a conjecture regarding the ML degree of $3 \times n$-matrices of rank 2.

Conjecture 2.4.1. *For $m = 3$ and $n \geq 3$, the ML degree of the variety $\mathcal{V}_2$ equals $2^{n+1} - 6$.*

The first three values already appeared in Theorem 2.1.1. We tested this formula by solving the equations of the local kernel formulation (2.11). This was done independently in `Macaulay2` and `Bertini`. With these computations, we verified Conjecture 2.4.1 up to $n = 10$. This conjecture, if correct, would furnish a simple and natural sequence of models, namely $3 \times n$-matrices of rank 2, whose ML degree grows exponentially in the number of states. This allows for the possibility of many local maxima.

We next formulate a refined version of the duality statement in Theorem 2.1.2. Given a data matrix U of format $m \times n$, we write Ω_U for the $m \times n$-matrix whose (i,j) entry equals

$$\frac{u_{ij}u_{i+}u_{+j}}{(u_{++})^3}.$$

Theorem 2.4.2. *Fix $m \leq n$ and U an $m \times n$-matrix with strictly positive integer entries. There exists a bijection between the complex critical points $P_1, P_2, \ldots, P_s$ of the likelihood function ℓ_U on $\mathcal{V}_r$ and the complex critical points $Q_1, Q_2, \ldots, Q_s$ of ℓ_U on $\mathcal{V}_{m-r+1}$ such that*

$$P_1 \star Q_1 \;=\; P_2 \star Q_2 \;=\; \cdots \;=\; P_s \star Q_s \;=\; \Omega_U. \tag{2.19}$$

In particular, this bijection preserves reality, positivity, and rationality of the critical points.

The proof of this theorem can be found in the next chapter as Theorem 3.1.1.

From the perspective of statistics, this result implies the following striking statement: maximum likelihood estimation for matrices of rank r is exactly the same problem as minimum likelihood estimation for matrices of corank $r-1$, and vice versa. This refined formulation of the duality statement allows us to improve the speed of MLE by passing to the complementary problem, where it may be easier to solve the likelihood equations. We saw a first instance of this in Section 2.3 when we discussed the last two columns in Table 2.3: the two methods give the same set of 191 solutions but the running times are complementary.

Remark 2.4.3. *Equation (2.19) is trivially satisfied for $r = 1$, where the ML degree is $s = 1$. Here, P_1 is the rank one matrix in (2.22), and $Q_1 = \frac{1}{u_{++}}U$. Clearly, we have $P_1 \star Q_1 = \Omega_U$.* □

We illustrate Theorem 2.4.2 for a specific case that has already appeared in the literature [13, 35, 48]. The first assertion in the next theorem resolves [48, Conjecture 11] affirmatively. In their conjecture, Zhu *et al.* [48] had identified the matrix $P(a,b)$ below, and they had asserted that it is the global maximum of the likehood function for the data matrix $U(a,b)$. Note that, for $a = 4$ and $b = 2$, this is the matrix for DiaNA's data in [35, Example 1.16].

Theorem 2.4.4. *Let $m = n = 4$, $a > b > 0$, and consider the following matrices:*

$$U(a,b) \;=\; \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{bmatrix} \quad \text{and} \quad P(a,b) \;=\; \frac{1}{8(a+3b)} \begin{bmatrix} a+b & a+b & 2b & 2b \\ a+b & a+b & 2b & 2b \\ 2b & 2b & a+b & a+b \\ 2b & 2b & a+b & a+b \end{bmatrix}.$$

The distribution $P(a,b)$ maximizes the likelihood function for the data matrix $U(a,b)$ on $\mathcal{M}_2$.

Proof. This statement is invariant under scaling the vector (a,b). We normalize by taking $4a + 12b = 16$. Then $b = (4-a)/3$ and a ranges in the open interval defined by $1 < a < 4$. For each such a, the likelihood function $\ell_{U(a,b)}$ has exactly 25 positive critical points in the rank 2 model $\mathcal{M}_2$, with the maximum value occurring at $P(a,b)$. This statement was shown using the following method and its illustration in Figure 2.1.

First, we selected $a = 2$ and computed the 191 critical points using `Bertini`. From these, `alphaCertified` proved that exactly 25 are real and, using the computed error

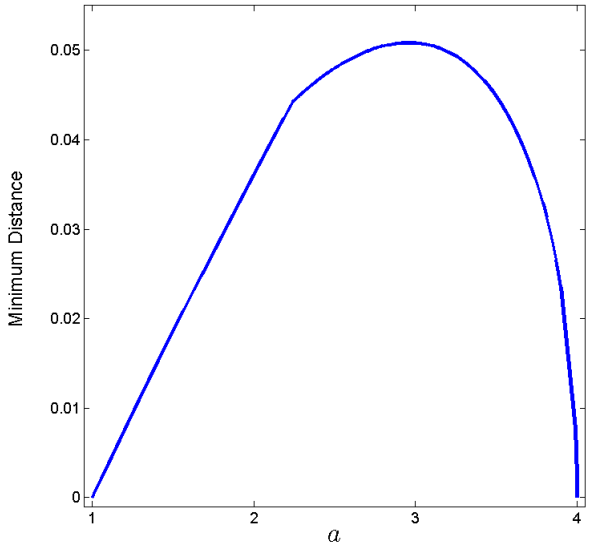

Figure 2.1: Minimum pairwise distance and lower bound (2.20) as a function of a.

bounds, it verified that all lie in Δ_{15}. We then expressed these real solutions as rational functions in a and b to show that all 25 real solutions remain positive for all $a > b > 0$. The critical points fall into four symmetry classes of size 6, 12, 4, and 3. Representatives of these classes are

$$X_1 = \frac{1}{16}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & \frac{2a}{a+b} & \frac{2b}{a+b} \\ 1 & 1 & \frac{2b}{a+b} & \frac{2a}{a+b} \end{bmatrix}, \qquad X_2 = \frac{1}{32(a+2b)}\begin{bmatrix} 2a+4b & 2a+4b & 2a+4b & 2a+4b \\ 2a+4b & 6a & 6b & 6b \\ 2a+4b & 6b & 3a+3b & 3a+3b \\ 2a+4b & 6b & 3a+3b & 3a+3b \end{bmatrix},$$

$$X_3 = \frac{1}{12(a+3b)}\begin{bmatrix} 3a & 3b & 3b & 3b \\ 3b & a+2b & a+2b & a+2b \\ 3b & a+2b & a+2b & a+2b \\ 3b & a+2b & a+2b & a+2b \end{bmatrix}, \qquad \text{and} \quad X_4 = P(a,b).$$

Using calculus, one can prove that $\log \ell_U(X_i) < \log \ell_U(X_{i+1})$ for $i = 1, 2, 3$.

All that remains is to show that the 191 solutions remain distinct on $1 < a < 4$ (with some coalesce at the boundary). The function mapping a to the minimum of the pairwise distances between the critical points is a piecewise smooth function. It is depicted in Figure 2.1. By tracking the homotopy paths as a changes from 2 to 1 and from 2 to 4, we are able to determine that this function is nowhere zero on the open interval $(1, 4)$. Additionally, by analyzing the solutions using [1], a lower bound on this minimum pairwise distance function is

$$\min\left\{ \frac{(a-1)\sqrt{a^2+17}}{12(a+8)}, \quad \frac{a+2-\sqrt{(a-1)(a-4)}}{48} - \frac{3(a^2-12a-16)+\sqrt{6(a-1)(a-4)(a^2-16a+96)}}{16(a+8)(a-10)} \right\} \tag{2.20}$$

which is also depicted in Figure 2.1. The first term of this minimum arises from X_2 and a member of the X_3 family which is equal to the minimum pairwise distances for values of a near 1. The second term arises from comparing the $(1, 1)$ entries of critical points. In short, all of the solutions remain distinct on $1 < a < 4$ and this establishes [48, Conjecture 11]. □

We checked the duality statement in Theorem 2.4.2 by performing the same computation for $m = n = 4$ and $r = 3$. We followed the 191 paths in the deformation from a general U_0 to a general $U(a, b)$. Using `Bertini`, we found that 12 endpoints had rank 2 while the other 179 had the expected rank of 3. Moving the other 179 solutions to $a = 2$ produced 179 distinct complex solutions that remain distinct and retain rank 3 on $(1, 4)$. Using the same certification process as above, precisely 25 are positive. These critical points of $\mathcal{M}_3$ form four symmetry classes having the same sizes $6, 12, 4$, and 3 as above, with representatives:

$$Y_1 = \tfrac{1}{8(a+3b)} \begin{bmatrix} 2a & 2b & 2b & 2b \\ 2b & 2a & 2b & 2b \\ 2b & 2b & a+b & a+b \\ 2b & 2b & a+b & a+b \end{bmatrix}, \qquad Y_2 = \tfrac{1}{12(a+3b)} \begin{bmatrix} 3a & 3b & 3b & 3b \\ 3b & a+2b & a+2b & a+2b \\ 3b & a+2b & \frac{2a(a+2b)}{a+b} & \frac{2b(a+2b)}{a+b} \\ 3b & a+2b & \frac{2b(a+2b)}{a+b} & \frac{2a(a+2b)}{a+b} \end{bmatrix},$$

$$Y_3 = \tfrac{1}{16(a+2b)} \begin{bmatrix} a+2b & a+2b & a+2b & a+2b \\ a+2b & 3a & 3b & 3b \\ a+2b & 3b & 3a & 3b \\ a+2b & 3b & 3b & 3a \end{bmatrix}, \quad Y_4 = \tfrac{1}{16(a+b)} \begin{bmatrix} 2a & 2b & a+b & a+b \\ 2b & 2a & a+b & a+b \\ a+b & a+b & 2a & 2b \\ a+b & a+b & 2b & 2a \end{bmatrix}.$$

The matrices are now sorted by decreasing value of $\ell_{U(a,b)}$, so the first matrix Y_1 is the MLE. Our real positive critical points satisfy the desired duality relation. Namely, we have

$$X_1 \star Y_1 \;=\; X_2 \star Y_2 \;=\; X_3 \star Y_3 \;=\; X_4 \star Y_4 \;=\; \frac{1}{64(a+3b)} U(a,b) \;=:\; \Omega_U.$$

We verified the same for the complex solutions.

When Theorem 2.4.2 was still a conjecture, we verified it for randomly selected data matrices with i.i.d. entries sampled from the uniform distribution on $[0, 1]$. After generating a random matrix, we verified equation (2.19) using the critical points computed by homotopy continuation. For $m = n = 3$ and $r = 2$, we verified (2.19) for 50000 instances. Additionally, for $m = n = 4$ and $r = 2$, we verified (2.19) for 10000 instances. We also did this for a handful of 4×5 instances (such as Example 2.3.3) and 5×5 instances (such as Example 2.3.4). The user can find `Macaulay2` code, which uses the `Bertini.m2` package (described in Chapter 6), to perform more numerical experiments at `www.math.ncsu.edu/~jdhauens/MLE`.

Theorem 2.4.2 and its analogue for symmetric matrices is particularly interesting in the special case when $m = n = 2r - 1$. Here we have an involution on the set of critical points of ℓ_U on $\mathcal{V}_r$ which has the following property. If $P_1, P_2, \ldots, P_s$ are the positive critical points in the model $\mathcal{M}_r$, ordered by increasing value of the log-likelihood function, then

$$\ell_U(P_1) + \ell_U(P_s) \;=\; \ell_U(P_2) + \ell_U(P_{s-1}) \;=\; \cdots \;=\; \ell_U(P_{\lceil s/2 \rceil}) + \ell_U(P_{\lfloor s/2 \rfloor}).$$

The identity (2.19) implies that Galois group which permutes the set of critical points is considerably smaller than the full symmetric group on these points. We shall demonstrate this for $n = 3$. What follows will explain the solutions in radicals seen in Example 2.3.2.

Let $\mathbb{Q}(U)$ denote the field of rational functions in entries of an indeterminate data matrix U, and let K denote the algebraic extension of $\mathbb{Q}(U)$ that is defined by adjoining all solutions of the likelihood equations. Thus the degree of the extension $K/\mathbb{Q}(U)$ is the ML degree. We are interested in the Galois group $G = \mathrm{Gal}(K, \mathbb{Q}(U))$ of this algebraic extension. This Galois group is a subgroup of the full symmetric group S_M where M is the ML degree.

The following result was found by explicit computations using `maple` and `Sage` [43].

Proposition 2.4.5. *The Galois group for MLE on* 3×3*-matrices (2.1) of rank* 2 *is a subgroup of order* 1920 *in* S_{10}*. As an abstract group, it is the semidirect product of* S_5 *and* $(\mathbb{Z}_2)^4$*. The Galois group for MLE on symmetric* 3×3*-matrices (2.12) of rank* 2 *is a subgroup of order* 24 *in* S_5*. As an abstract group, it is the symmetric group* S_4*. So, in the latter case, the six critical points of the likelihood function can be written in radicals in* $u_{11}, u_{12}, u_{13}, u_{22}, u_{23}, u_{33}$*.*

We close this section with an important observation that is implied by the various polynomial formulations of our problem, but which had not been explicitly stated in Section 2.2.

Remark 2.4.6. *Every complex critical point* P *of the likelihood function* ℓ_U *on* $\mathcal{V}_r$ *satisfies*

$$p_{i+} = \frac{u_{i+}}{u_{++}} \ \text{ for } i = 1, \ldots, m \qquad \text{and} \qquad p_{+j} = \frac{u_{+j}}{u_{++}} \ \text{ for } j = 1, \ldots, n.$$

A proof of this remark is given in Lemma 3.2.1. One is tempted to speculate that some version of Theorems 2.1.2, 2.2.5, and 2.4.2 might be true for other classes of toric models.

2.5 Rank versus non-negative rank

In the previous sections, we developed accurate methods for finding the global maximum of a likelihood function ℓ_U over non-negative matrices P of rank r whose entries sum to 1. Unfortunately, this is not quite the problem most practitioners and users of statistics would actually be interested in. Rather than restricting the rank of a probability table (2.1), it is the *non-negative rank* that is more relevant for applications. In this section we discuss this.

Let Mix_r denote the subset of Δ_{mn-1} that comprises all the mixtures of r independent distributions. In statistics, this is the archetype of a latent variable model, or hidden variable model. Mathematically, we can define the *mixture model* Mix_r as the set of all matrices

$$P \ = \ A \cdot \Lambda \cdot B, \tag{2.21}$$

where A is a non-negative $m\times r$-matrix whose columns sum to 1, Λ is an $r\times r$ diagonal matrix whose diagonal entries are non-negative and sum to 1, and B is a non-negative $r \times n$-matrix whose rows sum to 1. The *rank-constrained model* $\mathcal{M}_r = \mathcal{V}_r \cap \Delta_{mn-1}$ we discussed above is an algebraic relaxation of the mixture model Mix_r. This can be made precise as follows:

Proposition 2.5.1. *The rank-constrained model* $\mathcal{M}_r$ *is the Zariski closure of the mixture model* Mix_r *inside the simplex* Δ_{mn-1}*. If* $r \leq 2$ *then* $\text{Mix}_r = \mathcal{M}_r$*. If* $r \geq 3$ *then* $\text{Mix}_r \subsetneq \mathcal{M}_r$*.*

Proof. See Example 4.1.2, Example 4.1.4 and Proposition 4.1.6 in [12]. That book refers to secant varieties of Segre varieties, tensors of any format, and joint distributions of any number of random variables. Here we only need the case of matrices and two random variables. □

Our model $\mathcal{M}_r$ is the set of all distributions P of rank at most r, while Mix_r is the set of all distributions P of non-negative rank at most r. Having non-negative rank $\leq r$ means that $P = A' \cdot B'$ for some non-negative matrices where A' has r columns and B' has r rows. Any such factorization can be transformed into the particular form (2.21) which identifies the statistical parameters. For further information on these two models see [13, 32, 35].

Understanding the inclusion of Mix_r inside $\mathcal{M}_r$ becomes crucial when comparing different methodologies for maximum likelihood estimation. We used `Bertini` to compute all critical points of the likelihood function ℓ_U on $\mathcal{M}_r$, with the aim of identifying the global maximum $\widehat{P}$ of ℓ_U over $\mathcal{M}_r$. This assumes that $\widehat{P}$ is strictly positive. This is usually the case when U is strictly positive. The standard method used by statisticians is to run the *EM algorithm* in the space of model parameters (A, Λ, B). This results in a local maximum $(\widehat{A}, \widehat{\Lambda}, \widehat{B})$ of the likelihood function expressed in terms of the parameters. The fact that $\mathcal{M}_r$ is the Zariski closure of the mixture model Mix_r in the simplex Δ_{mn-1} has the following consequence:

Corollary 2.5.2. *Let $\widehat{P}_1, \ldots, \widehat{P}_s$ be the local maxima in $\mathcal{M}_r$ of the likelihood function ℓ_U. If a matrix $\widehat{P}_i$ has non-negative rank at most r then $\widehat{P}_i$ lies in Mix_r and matching parameters $(\widehat{A}_i, \widehat{\Lambda}_i, \widehat{B}_i)$ can found by solving (2.21). If all matrices $\widehat{P}_i$ have non-negative rank strictly larger than r then ℓ_U attains its maximum over Mix_r on the topological boundary ∂Mix_r.*

Proof. The second sentence holds because every matrix $P \in \Delta_{mn-1}$ of non-negative rank $\leq r$ admits a factorization of the special form (2.21). Indeed, if $P = A' \cdot B'$ is any non-negative factorization then we first scale the rows of A' to get a matrix A with row sums equal to 1, and we adjust the second matrix so that $P = A \cdot B''$. Now let Λ be the diagonal matrix whose entries are the column sums of B'' and set $B = \Lambda^{-1}B''$. Then $P = A \cdot \Lambda \cdot B$.

For the third sentence, suppose ℓ_U has its maximum over Mix_r at a point $\widehat{P}$ in $\text{Mix}_r \backslash \partial\text{Mix}_r$. Then $\widehat{P}$ is also a local maximum of ℓ_U on $\mathcal{M}_r$. Thus $\widehat{P}$ will be found by solving the critical equations for ℓ_U on $\mathcal{V}_r$. The matrix $\widehat{P}$ is an element of $\{\widehat{P}_1, \ldots, \widehat{P}_s\}$. Hence, this set contains a matrix of non-negative rank $\leq r$. This proves the contrapositive of the assertion. □

We shall now discuss the exact solution of the MLE problem for the mixture model Mix_r. Let us start with the low rank cases. The given input is a data matrix U as in (2.2).

If $r = 1$ then the likelihood function ℓ_U has a unique critical point. Let u_{*+} be the column vector of row sums of U, and let u_{+*} be the row vector of column sums of U. Then

$$\widehat{P} \quad = \quad \frac{1}{(u_{++})^2} \cdot u_{*+} \cdot u_{+*}. \tag{2.22}$$

If $r \geq 2$ then we compute the set $\{\widehat{P}_1, \ldots, \widehat{P}_s\}$ of all local maxima of the likelihood function ℓ_U on the model $\mathcal{M}_r$. This is done using the numerical algebraic geometry methods described in Section 2.3, by solving the likelihood equations (2.11) for the determinantal variety $\mathcal{V}_r$.

If $r = 2$ then every matrix $\widehat{P}_i$ has non-negative rank ≤ 2. We therefore select the matrix whose likelihood value $\ell_U(\widehat{P}_i)$ is maximal. Then $\widehat{P}_i$ solves the MLE problem for $\text{Mix}_2 = \mathcal{M}_2$.

Example 2.5.3. We experimented with the EM Algorithm for $r = 2$, as in [35, §1.3], on the 4×5 data matrix U discussed in Example 2.3.3. We ran 10000 iterations with starting points (A, Λ, B) sampled from the uniform distribution on the 15-dimensional parameter polytope

$$(\Delta_3 \times \Delta_3) \ \times \ \Delta_1 \ \times \ (\Delta_4 \times \Delta_4).$$

From these 10000 runs of the EM algorithm we obtained the following seven local maxima:

Occurrences	Critical point	$\log(\ell_U)$
2643 occurrences:	$\begin{bmatrix} 0.001678 & 0.01892 & 0.00001325 & 0.007008 & 0.00000722 \\ 0.01894 & 0.2136 & 0.00006605 & 0.07912 & 0.00008149 \\ 0.00007930 & 0.00003964 & 0.5447 & 0.00003964 & 0.00002643 \\ 0.007023 & 0.07921 & 0.00002643 & 0.02933 & 0.00003021 \end{bmatrix}$	-105973.49
2044 occurrences:	$\begin{bmatrix} 0.001332 & 0.00001777 & 0.02627 & 0.00000792 & 0.00000382 \\ 0.00007696 & 0.2274 & 0.00006503 & 0.08423 & 0.00004823 \\ 0.02628 & 0.00003913 & 0.5185 & 0.00004103 & 0.00007542 \\ 0.00002871 & 0.08432 & 0.00002762 & 0.03123 & 0.00001788 \end{bmatrix}$	-106487.35
1897 occurrences:	$\begin{bmatrix} 0.002245 & 0.02536 & 0.00001725 & 0.000006332 & 0.000005379 \\ 0.02535 & 0.2863 & 0.00006471 & 0.00004393 & 0.00006072 \\ 0.00009818 & 0.00003897 & 0.4495 & 0.09525 & 0.00006537 \\ 0.00002773 & 0.00008630 & 0.09530 & 0.02020 & 0.00001388 \end{bmatrix}$	-109697.04
1688 occurrences:	$\begin{bmatrix} 0.001111 & 0.00001327 & 0.02187 & 0.004634 & 0.000005304 \\ 0.00005289 & 0.3117 & 0.00006605 & 0.00003968 & 0.00001322 \\ 0.02191 & 0.00003963 & 0.4314 & 0.09144 & 0.0001046 \\ 0.004647 & 0.00007931 & 0.09148 & 0.01939 & 0.00002219 \end{bmatrix}$	-111172.67
1106 occurrences:	$\begin{bmatrix} 0.005321 & 0.00002006 & 0.00001106 & 0.02226 & 0.00002038 \\ 0.00005070 & 0.1135 & 0.1983 & 0.00004009 & 0.00001444 \\ 0.00008126 & 0.1983 & 0.3465 & 0.00003939 & 0.00002520 \\ 0.02227 & 0.00007333 & 0.00002771 & 0.09316 & 0.00008532 \end{bmatrix}$	-127069.50
529 occurrences:	$\begin{bmatrix} 0.0008641 & 0.009735 & 0.01701 & 0.00001350 & 0.00000289 \\ 0.009756 & 0.1099 & 0.1921 & 0.00003965 & 0.00003259 \\ 0.01705 & 0.1921 & 0.3357 & 0.00003959 & 0.00005693 \\ 0.00005301 & 0.00007930 & 0.00002642 & 0.1154 & 0.00005294 \end{bmatrix}$	-131013.73
93 occurrences:	$\begin{bmatrix} 0.02754 & 0.00001320 & 0.00001319 & 0.00001334 & 0.00005311 \\ 0.00005280 & 0.09999 & 0.1747 & 0.03704 & 0.00002957 \\ 0.00007916 & 0.1747 & 0.3053 & 0.06472 & 0.00005164 \\ 0.00005339 & 0.03706 & 0.06476 & 0.01373 & 0.00001102 \end{bmatrix}$	-148501.63

The first matrix is the global maximum, and it was the output in 2643 of our 10000 runs. Note that the ordering by objective function value agrees with the ordering by occurrence. We know from Example 2.3.3 that Δ_{19} contains 7 local maxima, and hence our EM experiment found them all. Each of the 7 matrices above has both rank and non-negative rank $r = 2$. □

If $r \geq 3$ then the situation is more challenging. To begin with, we need a method for testing whether a matrix has non-negative rank $\leq r$. Recent work by Moitra [31] shows that the computational complexity of this problem is lower than one might fear at first glance.

So, let us assume for now that this problem has been solved and we have an algorithm to decide quickly whether any of the matrices $\widehat{P}_i$ has non-negative rank r. If so, we pick among them the matrix $\widehat{P}_i$ of largest ℓ_U-value. This matrix is now a candidate for the MLE on Mix_r. But it may not actually be the MLE because the global maximum of the likelihood function ℓ_U may be attained on the boundary $\partial\mathrm{Mix}_r$. Furthermore, it is quite possible that none of the critical points in $\{\widehat{P}_1, \ldots, \widehat{P}_s\}$ lies in Mix_r. Then, according to the third sentence of Corollary 2.5.2, the MLE in the mixture model Mix_r necessarily lies in the boundary $\partial\mathrm{Mix}_r$.

Our discussion implies that, in order to perform exact maximum likelihood estimation for the mixture model, we need to have an exact algebraic description of $\partial\mathrm{Mix}_r$. Specifically, we must determine the polynomial equations that cut out the various irreducible components of the Zariski closure of $\partial\mathrm{Mix}_r$ as a subvariety of $\mathbb{P}^{mn-1}$. For each of these components, and the various strata where they intersect, we then need to compute the ML degree. That list of further ML degrees, combined with the value for $\mathcal{V}_r$ in Theorem 2.1.1, describes the true intrinsic algebraic complexity of the MLE $\widehat{P}$ as a piecewise algebraic function of the data U.

To be even more ambitious, we could ask for an exact semi-algebraic description of the set Mix_r. Namely, what we seek is a Boolean combination of polynomial inequalities in the unknowns p_{ij} that characterize Mix_r as a subset of $\mathcal{V}_r \cap \Delta_{mn-1}$. Finding such a description is an open problem in general, but was solved in the rank at most 3 case in [28].

We illustrate the first interesting case $(m, n, r) = (4, 4, 3)$ using the techniques developed by Mond, Smith, and van Straten in [32]. Components of $\partial\mathrm{Mix}_3$ correspond to different labelings of the configurations in [32, Figure 9]. Using the translations (seen in [32, §2]) between non-negative factorizations (2.21) and nested polygons, one of the labelings of [32, Figure 9 (a)] corresponds to the factorization

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \\ a_{41} & a_{42} & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & b_{12} & b_{13} & b_{14} \\ b_{21} & 0 & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & 0 \end{pmatrix}. \tag{2.23}$$

This equation parametrizes an irreducible divisor in the 14-dimensional variety $\mathcal{V}_3 \subset \mathbb{P}^{15}$. That divisor is one of the irreducible components of the algebraic boundary of $\mathcal{M}_3$. The corresponding prime ideal of height 2 in $\mathbb{Q}[p_{11}, \ldots, p_{44}]$ is obtained by eliminating the 17 unknowns a_{ij} and b_{ij} from the 16 scalar equations in (2.23). We find that this ideal is generated by the 4×4-determinant that defines $\mathcal{V}_3$ together with four sextics such as

$$\begin{aligned} &p_{11}p_{21}p_{22}p_{32}p_{33}p_{43} - p_{11}p_{21}p_{22}p_{33}^2p_{42} - p_{11}p_{21}p_{23}p_{32}^2p_{43} + p_{11}p_{21}p_{23}p_{32}p_{33}p_{42} - p_{11}p_{22}^2p_{31}p_{33}p_{43} \\ &+p_{11}p_{22}p_{23}p_{31}p_{32}p_{43} + p_{11}p_{22}p_{23}p_{31}p_{33}p_{42} - p_{11}p_{23}^2p_{31}p_{32}p_{42} + p_{12}p_{21}p_{22}p_{33}^2p_{41} - p_{12}p_{21}p_{23}p_{32}p_{33}p_{41} \\ &-p_{12}p_{22}p_{23}p_{31}p_{33}p_{41} + p_{12}p_{23}^2p_{31}p_{32}p_{41} + p_{13}p_{21}^2p_{32}^2p_{43} - p_{13}p_{21}^2p_{32}p_{33}p_{42} - 2p_{13}p_{21}p_{22}p_{31}p_{32}p_{43} \\ &+p_{13}p_{21}p_{22}p_{31}p_{33}p_{42} + p_{13}p_{21}p_{23}p_{31}p_{32}p_{42} + p_{13}p_{22}^2p_{31}^2p_{43} - p_{13}p_{22}p_{23}p_{31}^2p_{42}. \end{aligned}$$

What needs to be studied now is the ML degree of this codimension 2 subvariety of $\mathbb{P}^{15}$, and the approach of [26] would lead us to look at the topology of the associated very affine variety. In Proposition 5.3 of [28], the ML degree was determined to be 633.

Described above is the geometry of the MLE problem for the mixture model Mix_r regarded as a subset of the ambient simplex Δ_{mn-1}. Statisticians, on the other hand, are more accustomed to working in the space of model parameters, which is the product of simplices

$$(\Delta_{m-1})^r \times \Delta_{r-1} \times (\Delta_{n-1})^r. \tag{2.24}$$

Here our parameters are (A, Λ, B). The model Mix_r is the image of this parameter space in Δ_{mn-1} under the map (2.21). That parametrization is very far from identifiable. The reason is that the fibers of $(A, \Lambda, B) \mapsto P$ are semi-algebraic sets of possibly large dimension. In fact, the whole point of the paper [32] is to study the topology of these fibers as P varies.

The expectation-maximization (EM) algorithm is the local method of choice for finding the MLE on the mixture model Mix_r. Our readers might enjoy the exposition given in [35, §1.3]. We emphasize that the EM algorithm operates entirely in the parameter space (2.24). The likelihood function ℓ_U pulls back to a function on the interior of (2.24). The EM algorithm is an iterative method that converges to a critical point of that function, and, under some mild regularity hypotheses, that critical point $(\widehat{A}, \widehat{\Lambda}, \widehat{B})$ is then a local maximum. The image $\widehat{P}$ of the point in Mix_r is then a candidate for the global maximum of ℓ_U on Mix_r.

Example 2.5.4. We tried the EM Algorithm also for $r = 3$ on the 4×5 data matrix U in Examples 2.3.3 and 2.5.3. We ran 10000 iterations with starting points sampled from the uniform distribution on the 23-dimensional parameter polytope $(\Delta_3)^3 \times \Delta_2 \times (\Delta_4)^3$. From these 10000 runs of the EM algorithm, 9997 converged to one of eight local maxima. Three of the runs led to other fixed points. The following six local maxima are precisely the solutions already found in Example 2.3.3. We note that, in this particular instance, it happened that all local maxima in the rank model $\mathcal{M}_3$ actually lie in Mix_3, i.e. they have non-negative rank 3:

Occurrences	Critical point	$\log(\ell_U)$
3521 occurrences:	$\begin{bmatrix} 0.005321 & 0.00001322 & 0.00001322 & 0.02226 & 0.00002039 \\ 0.00005285 & 0.3117 & 0.00006607 & 0.00003964 & 0.00001321 \\ 0.00007929 & 0.00003964 & 0.5447 & 0.00003964 & 0.00002643 \\ 0.02227 & 0.00007927 & 0.00002642 & 0.09316 & 0.00008532 \end{bmatrix}$	-84649.67679
2293 occurrences:	$\begin{bmatrix} 0.002244 & 0.02535 & 0.00001324 & 0.00001333 & 0.0000054 \\ 0.02535 & 0.2863 & 0.00006606 & 0.00003961 & 0.00006065 \\ 0.00007929 & 0.00003964 & 0.5447 & 0.00003964 & 0.00002643 \\ 0.00005291 & 0.00007928 & 0.00002643 & 0.1154 & 0.00005289 \end{bmatrix}$	-86583.69000
1678 occurrences:	$\begin{bmatrix} 0.001332 & 0.00001326 & 0.02627 & 0.00001341 & 0.0000038 \\ 0.00005289 & 0.3117 & 0.00006607 & 0.00003964 & 0.00001322 \\ 0.02628 & 0.00003963 & 0.5185 & 0.00003961 & 0.00007538 \\ 0.00005296 & 0.00007928 & 0.00002642 & 0.1154 & 0.00005292 \end{bmatrix}$	-87698.20128
1320 occurrences:	$\begin{bmatrix} 0.02754 & 0.00001320 & 0.00001321 & 0.00001326 & 0.00005298 \\ 0.00005277 & 0.2274 & 0.00006606 & 0.08423 & 0.00004806 \\ 0.00007928 & 0.00003964 & 0.5447 & 0.00003964 & 0.00002643 \\ 0.00005310 & 0.08430 & 0.00002643 & 0.03122 & 0.00001788 \end{bmatrix}$	-98171.25551
576 occurrences:	$\begin{bmatrix} 0.02754 & 0.00001321 & 0.00001320 & 0.00001330 & 0.00005305 \\ 0.00005285 & 0.3117 & 0.00006605 & 0.00003968 & 0.00001322 \\ 0.00007916 & 0.00003964 & 0.4495 & 0.09526 & 0.00006519 \\ 0.00005324 & 0.00007932 & 0.09528 & 0.02019 & 0.00001389 \end{bmatrix}$	-102495.4349
68 occurrences:	$\begin{bmatrix} 0.02754 & 0.00001322 & 0.00001321 & 0.00001321 & 0.00005285 \\ 0.00005287 & 0.1135 & 0.1983 & 0.00003968 & 0.00001444 \\ 0.00007927 & 0.1983 & 0.3465 & 0.00003962 & 0.00002520 \\ 0.00005285 & 0.00007930 & 0.00002642 & 0.1154 & 0.00005285 \end{bmatrix}$	-121802.8945

In addition, our runs of the EM algorithm discovered the two local maxima

488 occurrences:	$\begin{bmatrix} 0.001678 & 0.01892 & 0.00001325 & 0.007008 & 0.0000072 \\ 0.01894 & 0.2136 & 0.00006605 & 0.07912 & 0.00008149 \\ 0.00007930 & 0.00003964 & 0.5447 & 0.00003964 & 0.00002643 \\ 0.007023 & 0.07921 & 0.00002643 & 0.02933 & 0.00003021 \end{bmatrix}$	-105973.4859
53 occurrences:	$\begin{bmatrix} 0.001111 & 0.00001341 & 0.02187 & 0.004634 & 0.0000053 \\ 0.00005299 & 0.3117 & 0.00006602 & 0.00003976 & 0.00001324 \\ 0.02191 & 0.00003960 & 0.4314 & 0.09144 & 0.0001046 \\ 0.004647 & 0.00007935 & 0.09148 & 0.01939 & 0.00002219 \end{bmatrix}$	-111172.6663

These do not satisfy the likelihood equations. They are located on the boundary of Mix_3. □

In the paper [28] the (algebraic) geometry of the EM algorithm is analyzed, with focus on the small cases of Theorem 2.1.1. Comparison with the methods introduced in this chapter opens up the possibility of characterizing conditions under which EM finds the global maximum, as it did in Example 2.5.4.

In this chapter, we have introduced a numerical algebraic geometry approach to maximum likelihood estimation. This approach led to new computational results and motivated surprising conjectures.

Chapter 3

Duality of Matrix Models

The content of this chapter will be published in the *International Mathematics Research Notices* as an article titled *Maximum Duality of Determinantal Varieties*, with minor changes throughout for consistency with other chapters. This is joint work with Jan Draisma.

3.1 Introduction and results

For an $m \times n$-data table $U = (u_{ij}) \in \mathbb{N}^{m \times n}$, we define the *likelihood function* $\ell_U : \mathbb{T}^{m \times n} \to \mathbb{T}$, where $\mathbb{T} = \mathbb{C}^*$ is the complex one-dimensional torus, as $\ell_U(Y) = \prod_{ij} y_{ij}^{u_{ij}}$ for $Y = (y_{ij})_{ij} \in \mathbb{T}^{m \times n}$. This terminology is motivated by the following observation. If Y is a matrix with positive real entries adding up to 1, interpreted as the joint probability distribution of two random variables taking values in $[m] := \{1, \ldots, m\}$ and $[n] := \{1, \ldots, n\}$, respectively, then up to a multinomial coefficient depending only on U, $\ell_U(Y)$ is the probability that when independently drawing $\sum_{i,j} u_{ij}$ pairs from the distribution Y, the number of pairs equal to (i, j) is u_{ij}. In other words, $\ell_U(Y)$ is the likelihood of Y, given observations recorded in the table U. A standard problem in statistics is to *maximize* $\ell_U(Y)$.

Without further constraints on Y this maximization problem is easy: it is uniquely solved by the matrix Y obtained by scaling U to lie in said probability simplex. But various meaningful statistical models require Y to lie in some *subvariety* X of $\mathbb{T}^{m \times n}$. For instance, in the model where the first and second random variable are required to be independent, one takes X equal to the intersection of the variety of matrices of rank 1 with the hyperplane $\sum_{ij} y_{ij} = 1$ supporting the probability simplex. Taking mixtures of this model, one is also led to intersect said hyperplane with the variety of rank-r matrices.

For general X, the maximum-likelihood estimate is typically much harder to find (though in the independence model it is still well-understood). One reason for this is that the restriction of ℓ_U to X may have many critical points. Under suitable assumptions, this number of critical points is finite and independent of U (for sufficiently general U), and is called the *maximum likelihood degree* or *ML degree* of X. Finiteness and independence of U holds, for instance, for smooth closed subvarieties of a torus [26], but also for all varieties X studied in [20, 23] (which are smooth but not closed, and become closed but singular if one takes the closure).

We take X to be a smooth, irreducible, locally closed, complex subvariety of a torus.

Doing so, we tacitly shift attention from the statistical motivation to complex geometry—in particular, we no longer worry whether the critical points counted by the ML degree lie in the probability simplex or are even real-valued matrices.

The set of all critical points for varying data matrices U has a beautiful geometric interpretation: Given $P \in X$ and a vector V in the tangent space $T_P X$ to X at P, the derivative of ℓ_U at P in the direction V equals $\ell_U(P) \cdot \sum_{ij} \frac{v_{ij}}{p_{ij}} u_{ij}$. This vanishes if and only if U is perpendicular, in the standard symmetric bilinear form on $\mathbb{C}^{m\times n} = \mathbb{C}^{mn}$, to the entry-wise quotient $\frac{V}{P}$ of V by P. This leads us to define

$$\mathrm{Crit}(X) := \{(P, U) \mid \frac{T_X P}{P} \perp U\} \subseteq X \times \mathbb{C}^{m\times n},$$

which is called *the variety of critical points* of X in [26], except that there U varies over projective space and the closure is taken. By construction, $\mathrm{Crit}(X)$ is smooth and irreducible, and has dimension mn; indeed, it is a vector bundle over X of rank $mn - \dim X$. The ML degree of X is well-defined if and only if the projection $\mathrm{Crit}(X) \to \mathbb{C}^{m\times n}$ is dominant, in which case the degree of this rational map is the ML degree of X.

In this chapter, we consider three choices for X, all given by rank constraints: First, in the *rectangular* case, we order m, n such that $m \le n$, fix a rank $r \in [m]$, and take X equal to

$$\mathcal{M}_r := \{P \in \mathbb{T}^{m\times n} \mid \sum_{ij} p_{ij} = 1 \text{ and } \operatorname{rk} P = r\}.$$

Second, in the *symmetric* case, we take $m = n$ and take X equal to

$$\mathcal{SM}_r := \left\{ P = \begin{bmatrix} 2p_{11} & p_{12} & \cdots & p_{1m} \\ p_{12} & 2p_{22} & & \\ \vdots & & \ddots & \\ p_{1m} & & & 2p_{mm} \end{bmatrix} \in \mathbb{T}^{m\times m} \mid \sum_{i\le j} p_{ij} = 1 \text{ and } \operatorname{rk}(P) = r \right\}.$$

Third, in the *skew-symmetric* or *alternating* case, we take $m = n$ and, for *even* $r \in [m]$, take X equal to

$$\mathcal{AM}_r := \left\{ P = \begin{bmatrix} 0 & p_{12} & \cdots & p_{1m} \\ -p_{12} & 0 & & \\ \vdots & & \ddots & \\ -p_{1m} & & & 0 \end{bmatrix} \in \mathbb{C}^{m\times m} \mid \begin{array}{l} \sum_{i<j} p_{ij} = 1, \operatorname{rk}(P) = r, \\ \text{and } \forall i < j : p_{ij} \neq 0 \end{array} \right\}.$$

Minor modifications of the likelihood function are needed in the latter two cases: we define as $\ell_U(P) := \prod_{i\le j} p_{ij}^{u_{ij}}$ in the symmetric case, and as $\ell_U(P) := \prod_{i<j} p_{ij}^{u_{ij}}$ in the alternating case.

In Chapter 2, using the numerical algebraic geometry software `Bertini` [5, 3], the ML degree of $\mathcal{M}_r$ is computed for various values of r, m, n with $r \le m \le n$. The numbers are listed in Theorem 2.4. Observe that the numbers for rank r and rank $m - r + 1$ coincide. From these computations, the natural conjecture to put forward is that this always holds, and that there is an explicit bijection between the two sets of critical points. In addition,

the computational results from in 2.13 motivate similar conjectures regarding symmetric matrices. In this chapter, we prove these results using the term *ML-duality* suggested by Sturmfels.

Theorem 3.1.1 (ML-duality for rectangular matrices)**.** *Fix a rank $r \in [m]$ and let $U \in \mathbb{N}^{m\times n}$ with $m \leq n$ be a sufficiently general data matrix. Then there is an explicit involutive bijection between the critical points of ℓ_U on $\mathcal{M}_r$ and the critical points of ℓ_U on $\mathcal{M}_{m-r+1}$. In particular, the ML degrees of $\mathcal{M}_r$ and $\mathcal{M}_{m-r+1}$ coincide.*

Moreover, the product $\ell_U(P)\ell_U(Q)$ is the same for all pairs consisting of a rank-r critical point P and the corresponding rank-$m-r+1$ point Q.

Theorem 3.1.2 (ML-duality for symmetric matrices)**.** *Fix a rank $r \in [m]$ and let $U \in \mathbb{N}^{m\times m}$ be a sufficiently general symmetric data matrix. Then there is an explicit involutive bijection between the critical points of ℓ_U on $\mathcal{SM}_r$ and the critical points of ℓ_U on $\mathcal{SM}_{m-r+1}$. In particular, the ML degrees of $\mathcal{SM}_r$ and $\mathcal{SM}_{m-r+1}$ coincide.*

Moreover, the product $\ell_U(P)\ell_U(Q)$ is the same for all pairs consisting of a rank-r critical point P and the corresponding rank-$m-r+1$ point Q.

In the alternating case, the duality of $\mathcal{AM}_r$ turns out *not* to be some $\mathcal{AM}_s$ but rather an affine translate of a determinantal variety defined as follows. Let S be the skew $m\times m$-matrix

$$S := \begin{bmatrix} 0 & 1 & \cdots & 1 \\ -1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ -1 & \cdots & -1 & 0 \end{bmatrix},$$

and for even $s \in \{0, \ldots, m-1\}$ consider the variety

$$\mathcal{AM}'_s := \{P \in \mathbb{C}^{m\times m} \mid P \text{ skew}, \forall i < j : p_{ij} \neq 0, \text{ and } \operatorname{rk}(S-P) = s\}.$$

Note that, unlike in $\mathcal{AM}_r$, the upper triangular entries of $P \in \mathcal{AM}'_s$ are not required to add up to 1. For this reason we do not say $\mathcal{AM}_r$ and $\mathcal{AM}'_s$ are ML-dual. Instead, we only say there is a duality between critical points of ℓ_U on $\mathcal{AM}_s$ and critical points of ℓ_U on $\mathcal{AM}'_s$. The difference between this notion of duality and ML-duality is explained in Section 4.4.

Theorem 3.1.3 (Duality for skew matrices)**.** *Fix an even rank $r \in \{2, \ldots, m\}$ and let $U \in \mathbb{N}^{m\times m}$ be a sufficiently general* symmetric *data matrix with zeroes on the diagonal. Let $s \in \{0, \ldots, m-2\}$ be the largest even integer less than or equal to $m-r$. Then there is an explicit involutive bijection between the critical points of ℓ_U on $\mathcal{AM}_r$ and the critical points of ℓ_U on $\mathcal{AM}'_s$. In particular, the ML degrees of $\mathcal{AM}_r$ and $\mathcal{AM}'_s$ coincide.*

Moreover, the product $\ell_U(P)\ell_U(Q)$ is the same for all pairs consisting of a rank-r critical point P on $\mathcal{AM}_r$ and the corresponding rank-s point Q on $\mathcal{AM}'_s$.

The proof is similar in each of the three cases. First, we determine the tangent space to X at a critical point P of ℓ_U for sufficiently general U. It turns out that this space is spanned by certain rank-one or rank-two matrices. Imposing that P be a critical point, i.e., that the derivative of ℓ_U vanishes in each of these low-rank directions leads to the conclusion that a

certain matrix Q, determined from P using some involution involving the fixed matrix U, has rank at most $m-r+1$ (or s in the skew case) and is itself a critical point on the variety of matrices of its rank. Letting $k \leq m-r+1$ (respectively, $k \leq s$) be generic rank of Qs thus obtained, we reverse the roles of P and Q to argue that k must equal s, thus establishing the result. In the remainder of this chapter we fill in the details in each of the three cases, in particular making the involution $P \to Q$ explicit.

3.2 Maximum likelihood duality in the rectangular case

Let $m \leq n$ be natural numbers and let $\mathcal{M}_r \subseteq \mathbb{T}^{m\times n}$ denote the variety of $m \times n$-matrices of rank r whose entries sum up to 1. Fix a sufficiently general data matrix $U = (u_{ij})_{ij} \in \mathbb{N}^{m\times n}$, which gives rise to the likelihood function $\ell_U : \mathcal{M}_r \to \mathbb{T},\ \ell_U(P) = \prod_{i,j} p_{ij}^{u_{ij}}$. Let $P \in \mathcal{M}_r$ be a critical point for ℓ_U, which means that the derivative of ℓ_U vanishes on the tangent space $T_P\mathcal{M}_r$ to $\mathcal{M}_r$ at P. This tangent space equals

$$T_P\mathcal{M}_r = \{X = (x_{ij})_{ij} \in \mathbb{C}^{m\times n} \mid X \ker P \subseteq \operatorname{im} P \text{ and } \sum_{ij} x_{ij} = 0\}. \tag{3.1}$$

Here the first condition ensures that X is tangent at P to the variety of rank-r matrices (see, e.g., [19, Example 14.6]) and the second condition ensures that X is tangent to the hyperplane where the sum of all matrix entries is 1.

Given $X \in T_P\mathcal{M}_r$, the derivative of ℓ_U in that direction equals $\ell_U(P) \cdot \sum_{ij} \frac{x_{ij}u_{ij}}{p_{ij}}$, which vanishes if and only if the second factor vanishes. We will now prove that the marginals of P are proportional to those of U (see also 2.4.6). We write $\mathbf{1}$ for the all-one vectors in both $\mathbb{C}^m$ and $\mathbb{C}^n$, and use self-explanatory notation such as $u_{i+} := \sum_j u_{ij}$ and $u_{++} := \sum_{ij} u_{ij}$.

Lemma 3.2.1. *The column vector $P\mathbf{1}$ is a non-zero scalar multiple of $U\mathbf{1}$ and the row vector $\mathbf{1}^T P$ is a non-zero scalar multiple of $\mathbf{1}^T U$.*

Proof. We prove the first statement; the second statement is proved similarly. We want to show that the 2×2-minors of the $m \times 2$-matrix $[P\mathbf{1}|U\mathbf{1}]$ vanish. We give the argument for the upper minor. Let $X = (x_{ij})$ be the $m \times n$-matrix whose first row equals p_{2+} times the first row of P, whose second row equals $-p_{1+}$ times the second row of P, and all of whose other rows are zero. Then $X \in T_P\mathcal{M}_r$, so that the derivative $\sum_{ij} x_{ij}\frac{u_{ij}}{p_{ij}}$ is zero. On the other hand, substituting X into $\sum_{ij} x_{ij}\frac{u_{ij}}{p_{ij}}$ yields $u_{1+}p_{2+} - u_{2+}p_{1+}$, hence this minor is zero as desired. The scalar multiple in both cases is $\frac{p_{++}}{u_{++}} = \frac{1}{u_{++}}$, which is non-zero. □

Define $Q = (q_{ij})_{ij}$ by $p_{ij}q_{ij} = u_{i+}u_{ij}u_{+j}$. This is going to be our dual critical point, up to a normalization factor that we determine now.

Lemma 3.2.2. *The sum $\sum_{ij} q_{ij}$ equals $(u_{++})^3$.*

Proof. By Lemma 3.2.1 the rank-one matrix Y defined by $y_{ij} = u_{i+}u_{+j}$ has image contained in $\operatorname{im} P$. Hence it satisfies the linear condition $Y \ker P \subseteq \operatorname{im} P$, but not the condition $\sum_{ij} y_{ij} = 0$. Similarly, P itself satisfies $P \ker P \subseteq \operatorname{im} P$, but not $\sum_{ij} p_{ij} = 0$. Hence, we can

decompose Y uniquely as $cP + X$ where $c \in \mathbb{C}$ and where X satisfies $X \ker P \subseteq \operatorname{im} P$ and $\sum_{ij} x_{ij} = 0$, i.e., where $X \in T_P\mathcal{M}_r$. Then we have

$$\sum_{ij} q_{ij} = \sum_{ij} \frac{y_{ij}u_{ij}}{p_{ij}} = \sum_{ij} cu_{ij} + \sum_{ij} \frac{x_{ij}u_{ij}}{p_{ij}} = \sum_{ij} cu_{ij} + 0 = cu_{++}$$

by criticality of P. The scalar c equals

$$\frac{\sum_{ij} y_{ij}}{\sum_{ij} p_{ij}} = \frac{\sum_{ij} u_{i+}u_{+j}}{1} = (u_{++})^2,$$

which proves the lemma. □

We will use rank-one matrices in the tangent space $T_P\mathcal{M}_r$. We equip both $\mathbb{C}^m$ and $\mathbb{C}^n$ with their standard symmetric bilinear forms.

Lemma 3.2.3. *The tangent space $T_P\mathcal{M}_r$ at P is spanned by all rank-one matrices vw^T satisfying the following two conditions:*

- *$v \in \operatorname{im} P$ or $w \perp \ker P$; and*
- *$v \perp \mathbf{1}$ or $w \perp \mathbf{1}$.*

In the proof we will need that $\operatorname{im} P$ is not contained in the hyperplane $\mathbf{1}^\perp$ and that, dually, $\ker P$ does not contain $\mathbf{1}$. These conditions will be satisfied by genericity of U.

Proof. The first condition ensures that the rank-one matrices in the lemma map $\ker P$ into $\operatorname{im} P$, and the second condition ensures that the sum of all entries of those rank-one matrices is zero, so that they lie in $T_P\mathcal{M}_r$, see (3.1). To show that these rank-one matrices span the tangent space $T_P\mathcal{M}_r$, decompose $\mathbb{C}^m$ as $A \oplus B \oplus C$ where $A \oplus C = \mathbf{1}^\perp$ and $A \oplus B = \operatorname{im} P$. Here we use that $\operatorname{im} P$ is not contained in the hyperplane $\mathbf{1}^\perp$.

Similarly, decompose $\mathbb{C}^n = A' \oplus B' \oplus C'$ where $A' \oplus C'$ is the hyperplane $\mathbf{1}^\perp$ and $A' \oplus B' = (\ker P)^\perp$; here we use the second genericity assumption on P. These spaces have the following dimensions:

$$\begin{array}{lll} \dim A = r-1 & \dim B = 1 & \dim C = m-r \\ \dim A' = r-1 & \dim B' = 1 & \dim C' = n-r. \end{array}$$

The space spanned by the rank-one matrices in the lemma has the space $(B \otimes B') \oplus (C \otimes C')$ as a vector space complement. The dimension of this complement is $1 + (m-r)(n-r)$, which is also the codimension of $\mathcal{M}_r$. □

Let $R = \operatorname{diag}(u_{i+})_i$ and $K = \operatorname{diag}(u_{+j})_j$ be the diagonal matrices recording the row and column sums of U on their diagonals. Then, by Lemma 3.2.1, $P\mathbf{1}$ is a scalar multiple of $R\mathbf{1}$ and $\mathbf{1}^T P$ is a scalar multiple of $\mathbf{1}^T K$. This implies that, in the decompositions in the proof of Lemma 3.2.3, we may take B spanned by $U\mathbf{1} = R\mathbf{1}$ and B' spanned by $U\mathbf{1} = K\mathbf{1}$. Note that P, Q satisfy $P * Q = RUK$, where $*$ denotes the Hadamard product.

Observe also that criticality of P is equivalent to $v^T R^{-1} Q K^{-1} w = 0$ for all rank-one matrices vw^T as in Lemma 3.2.3. This criterion will be used in the proof of our duality result for $\mathcal{M}_r$.

Theorem 3.2.4 (ML-duality for rectangular matrices)**.** *Let $U \in \mathbb{N}^{m\times n}$ be a sufficiently general data matrix and let P be a critical point of ℓ_U on $\mathcal{M}_r$. Define $Q = (q_{ij})_{ij}$ by $q_{ij}p_{ij} = u_{i+}u_{ij}u_{+j}$. Then $Q/(u_{++}^3)$ is a critical point of ℓ_U on $\mathcal{M}_{m-r+1}$.*

Before proceeding with the proof, we point out that the construction of $Q' := Q/(u_{++})^3$ from P is symmetric in P and Q. As a consequence, the map $P \mapsto Q'$ from critical points of ℓ_U on $\mathcal{M}_r$ to critical points on $\mathcal{M}_{m-r+1}$ is a bijection. Moreover, it has the property that $\ell_U(P) \cdot \ell_U(Q')$ depends only on U. In particular, if one lists the critical points $P \in \mathcal{M}_r$ with positive real entries in order of decreasing log-likelihood, then the corresponding $Q' \in \mathcal{M}_{m-r+1}$ appear in order of increasing log-likelihood, since the sum $\log \ell_U(P) + \log \ell_U(Q')$ depends only on U.

Proof. Lemma 3.2.2 takes care of the normalization factor, which we therefore ignore during most of this proof. We first show that Q has rank at most $m - r + 1$. For this we take arbitrary v in the space $A = \mathbf{1}^\perp \cap \operatorname{im} P$ from the proof of Lemma 3.2.3 and arbitrary $w \in \mathbb{C}^n$, so that $vw^T \in T_P\mathcal{M}_r$. From $v^T R^{-1}QK^{-1}w = 0$ we conclude that $R^{-1} \operatorname{im} Q \subseteq A^\perp$ because v was arbitrary in A. Equivalently, since R is diagonal and hence symmetric, we conclude that $\operatorname{im} Q \subseteq (R^{-1}A)^\perp = (R^{-1}A)^\perp$. The latter space has dimension $m - r + 1$, which is therefore an upper bound on the rank of Q.

Similarly, for $w \in A'$ and any $v \in \mathbb{C}^m$, the matrix vw^T lies in the tangent space $T_P\mathcal{M}_r$, and we find $v^T R^{-1}QK^{-1}w = 0$. Since v was arbitrary, this means that $QK^{-1}w = 0$, so $\ker Q$ contains $K^{-1}A'$, a space of dimension $r-1$. If $n > m$, however, then by the above the kernel of Q strictly contains $K^{-1}A'$.

Next we prove that for any rank-one matrix xy^T such that

- $x \perp R^{-1}A$ or $y \perp K^{-1}A'$; *and*
- $x \perp \mathbf{1}$ or $y \perp \mathbf{1}$

we have $\sum_{ij} \frac{x_i u_{ij} y_j}{q_{ij}} = 0$. Note that the conclusion can be written as $x^T R^{-1}PK^{-1}y = 0$, and observe the similarity with the characterization of $T_P\mathcal{M}_r$ in Lemma 3.2.3 that will give us conditions of criticality of Q.

Given arbitrary $y \in \mathbb{C}^n$ we can write $PK^{-1}y$ as $v + cR\mathbf{1}$ with $v \in A$. Then for $x \in (R^{-1}A)^\perp$ perpendicular to $\mathbf{1}$ we find

$$x^T R^{-1}PK^{-1}y = x^T R^{-1}(v + cR\mathbf{1}) = 0 + cx^T\mathbf{1} = 0,$$

as desired. If, on the other hand, $x \in (R^{-1}A)^\perp$ is not perpendicular to $\mathbf{1}$ but $y \in \mathbb{C}^n$ is, then writing $w := K^{-1}y$ we claim that $v := Pw$ lies in A. For this we compute the dot product

$$\mathbf{1}^T Pw = \mathbf{1}^T Uw = \mathbf{1}^T Kw = \mathbf{1}^T y = 0,$$

where the first equality is justified by Lemma 3.2.1. Hence, again, $x^T R^{-1}PK^{-1}y = x^T R^{-1}v = 0$. The checks for the case where $y \perp K^{-1}A'$ are completely analogous.

Now denote the rank of Q by k, so that $k \leq m - r + 1$. From $\operatorname{im} Q \subseteq (R^{-1}A)^\perp$ and $(\ker Q)^\perp \subseteq (K^{-1}A')^\perp$ we conclude that the derivative of ℓ_U at Q' in the direction xy^T vanishes, in particular, when xy^T lies in the tangent space at Q' to $\mathcal{M}_k$. Hence Q' is a critical point for ℓ_U on $\mathcal{M}_k$.

Finally, we need to show that the generic rank k of Q thus obtained (from a sufficiently general U and a critical point $P \in \mathcal{M}_r$ of ℓ_U) equals $m-r+1$, rather than being strictly smaller. For this, observe that we have constructed, for any $r \in [m]$, a rational map of irreducible varieties

$$\psi_r : \mathrm{Crit}(\mathcal{M}_r) \dashrightarrow \mathrm{Crit}(\mathcal{M}_{f(r)}), \quad (P,U) \mapsto (\frac{1}{(u_{++})^3} \cdot \frac{RUK}{P}, U) = (Q', U)$$

where $f : [m] \to [m]$ maps r to the generic rank of the matrix Q' as (P,U) varies over $\mathrm{Crit}(\mathcal{M}_r)$. Since ψ_r commutes with the projection on the second factor, its image has dimension mn, hence ψ_r is dominant. But it is also injective—in fact, (P,U) can be recovered from (Q',U) with the exact same formula. This shows that ψ_r is birational, and that $\psi_{f(r)}$ is its inverse as a birational map. In particular, $f(f(r)) = r$, so that f is a bijection. But the only bijection $[m] \to [m]$ with the property that $f(r) \leq m-r+1$ for all r is $r \mapsto m-r+1$. Indeed, if r were the smallest value for which $f(r) \neq m-r+1$, then $m-r+1$ would not be in the image of f. This concludes the proof of the theorem. □

Remark 3.2.5. It *can* happen that the rank of Q is strictly smaller than $m-r+1$ but the proof above shows that for sufficiently general U this does *not* happen. For example, in the rectangular case where $m=n=4$, if we have that

$$U = \frac{1}{40}\begin{bmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{bmatrix} \text{ and } P = \frac{1}{80}\begin{bmatrix} 6+2i & 5-\sqrt{5} & 5+\sqrt{5} & 4-2i \\ 5-\sqrt{5} & 6-2i & 4+2i & 5+\sqrt{5} \\ 5+\sqrt{5} & 4+2i & 6-2i & 5-\sqrt{5} \\ 4-2i & 5+\sqrt{5} & 5-\sqrt{5} & 6+2i \end{bmatrix}$$

then there exist ML degree points in $\mathrm{Crit}(\mathcal{M}_2)$ with this choice of U. It can be shown $(P,U) \in \mathrm{Crit}(\mathcal{M}_2)$ is one such point. Because u_{++}=1 we have $Q = Q'$, and

$$Q = \frac{1}{500}\begin{bmatrix} 6-2i & 5+\sqrt{5} & 5-\sqrt{5} & 4+2i \\ 5+\sqrt{5} & 6+2i & 4-2i & 5-\sqrt{5} \\ 5-\sqrt{5} & 4-2i & 6+2i & 5+\sqrt{5} \\ 4+2i & 5-\sqrt{5} & 5+\sqrt{5} & 6-2i \end{bmatrix}$$

satisfies $p_{ij}q_{ij} = \frac{u_{i+}u_{++}u_{+j}}{u_{++}^3}$. In this case, Q has rank 2 instead of rank 3. This is an important fact for numerical computations. If we were to use the homotopy methods as in Chapter 2 to find the critical points of l_U on $\mathcal{M}_3$, we would track a path from a generic point of $\mathrm{Crit}(\mathcal{M}_3)$ to the point (Q,U). Since Q has rank less than 3, this will correspond to tracking a path to a singularity leading to numerical difficulties. But by determining all critical points of l_U on $\mathcal{M}_2$, we avoid these numerical difficulties. To determine the points of $\mathrm{Crit}(\mathcal{M}_3)$ with U as above, we use the equation $p_{ij}q_{ij} = \frac{u_{i+}u_{++}u_{+j}}{u_{++}^3}$ and determine which (q_{ij}) have rank 3.

3.3 Maximum likelihood duality in the symmetric case

Let m be a natural number and let $\mathcal{SM}_r$ denote the variety of symmetric $m \times m$-matrices of rank r whose entries sum to 2. A point P of $\mathcal{SM}_r$ and data matrix U will be denoted by

$$P = \begin{bmatrix} 2p_{11} & p_{12} & \cdots & p_{1m} \\ p_{12} & 2p_{22} & & \\ \vdots & & \ddots & \\ p_{1m} & & & 2p_{mm} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 2u_{11} & u_{12} & \cdots & u_{1m} \\ u_{12} & 2u_{22} & & \\ \vdots & & \ddots & \\ u_{1m} & & & 2u_{mm} \end{bmatrix}.$$

We denote the (i,j)-entries of P and U by P_{ij} and U_{ij} to distinguish them from the p_{ij} and u_{ij}, respectively. Recall that the likelihood function in the symmetric case is defined as $\ell_U(P) := \prod_{i \le j} p_{ij}^{u_{ij}}$, which in terms of the entries of P equals $(\prod_{i<j} P_{ij}^{u_{ij}}) \cdot (\prod_i (P_{ii}/2)^{u_{ii}})$. From now on we fix a sufficiently general data matrix U and a critical point P for ℓ_U on $\mathcal{SM}_r$. The tangent space $T_P\mathcal{SM}_r$ equals

$$T_P\mathcal{SM}_r = \{X \in \mathbb{C}^{m \times m} \text{ symmetric} \mid X \ker P \subseteq \operatorname{im} P \text{ and } \sum_{ij} x_{ij} = 0\}. \tag{3.2}$$

Given a tangent vector $X \in T_P\mathcal{SM}_r$, the derivative of ℓ_U in that direction equals

$$\sum_{i<j} \frac{X_{ij}u_{ij}}{P_{ij}} + \sum_i \frac{(X_{ii}/2)u_{ii}}{P_{ii}/2} = \sum_{i \le j} \frac{X_{ij}u_{ij}}{P_{ij}}$$

(up to a factor irrelevant for its vanishing). We set

$$U_{i+} := \sum_j U_{ij} \text{ and } U_{++} := \sum_i \sum_j U_{ij},$$

and similarly for P. The symmetric analogue of Lemma 3.2.1 is the following.

Lemma 3.3.1. *The vector $P\mathbf{1}$ is a non-zero scalar multiple of $U\mathbf{1}$.*

Proof. We need to prove that the $m \times 2$-matrix $(P\mathbf{1}|U\mathbf{1})$ has 2×2-minors equal to zero. We prove this for the minor in the first two rows. Set $a := P_{1+}$ and $b := P_{2+}$, and define $v_1, v_2 \in \mathbb{C}^m$ by $v_1 = (b, 0, 0, \ldots, 0)^T, v_2 = (0, a, 0, \ldots, 0)$. Let w_1, w_2 be the first and second column of P, respectively. Then for each $i = 1, 2$ the matrix $X^{(i)} = v_i w_i^T + w_i v_i^T$ lies in the tangent space at P to the variety of symmetric rank-r matrices, and the difference $X := X^{(1)} - X^{(2)}$ has sum of entries equal to 0 and therefore lies in $T_P\mathcal{SM}_r$. The symmetric matrix X looks like

$$\begin{bmatrix} 2bP_{11} & (b-a)P_{12} & bP_{13} & \cdots & bP_{1m} \\ * & 2aP_{22} & -aP_{23} & \cdots & -aP_{2m} \\ * & * & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ * & * & 0 & \cdots & 0 \end{bmatrix}.$$

The derivative of ℓ_U at P in the direction X equals

$$\sum_{i \le j} \frac{X_{ij}u_{ij}}{P_{ij}} = bU_{1+} - aU_{2+},$$

and this derivative vanishes by criticality of P. The relevant non-zero scalar multiple is $\frac{P_{++}}{U_{++}} = \frac{2}{U_{++}}$, which is non-zero. □

The analogue of R, K from the rectangular case is $R := \operatorname{diag}(U_{1+}, \ldots, U_{m+})$ and $K := \operatorname{diag}(U_{+1}, \ldots, U_{+m})$. Note $R = K$ because U is symmetric, but we keep this notation to mirror the rectangular case. As in the rectangular case, define the symmetric matrix Q by $P * Q = RUR$, i.e., $P_{ij}Q_{ij} = U_{i+}U_{ij}U_{j+}$ for $i, j \in [m]$. This will be our dual critical point, up to a normalizing factor to be determined now.

Lemma 3.3.2. *The sum $\sum_{ij} Q_{ij}$ equals $\frac{(U_{++})^3}{2}$.*

Proof. By Lemma 3.3.1 the rank-one matrix Y with entries $Y_{ij} = U_{i+}U_{j+}$ has image contained in $\operatorname{im} P$, and so does P. So we can decompose $Y = cP + X$ with $c \in \mathbb{C}$ and $X \in T_P\mathcal{SM}_r$, and we find

$$\sum_{ij} Q_{ij} = \sum_{ij} \frac{Y_{ij}U_{ij}}{P_{ij}} = \sum_{ij} cU_{ij} + \sum_{ij} \frac{X_{ij}U_{ij}}{P_{ij}} = cU_{++} + 0 = cU_{++}.$$

Moreover, the scalar c equals $\frac{Y_{++}}{P_{++}} = \frac{(U_{++})^2}{2}$, which shows that $Q_{++} = \frac{(U_{++})^3}{2}$. □

As in the rectangular case, we will make use of low-rank elements in $T_P\mathcal{SM}_r$, where now "low rank" means rank two.

Lemma 3.3.3. *The tangent space $T_P\mathcal{SM}_r$ is spanned by all matrices of the form $vw^T + w^Tv$ with $v \in im(P)$ and $w \in \mathbb{C}^m$, with the additional constraint that the sum of all entries is zero, i.e., that one of v and w is perpendicular to $\mathbf{1}$.*

In the proof we will implicitly use that $\operatorname{im} P$ is not contained in $\mathbf{1}^\perp$, which is true by genericity of U.

Proof. The proof is similar to that of Lemma 3.2.3. First, the matrices in the lemma satisfy the conditions characterizing $T_P\mathcal{SM}_r$; see (3.2). Second, to show that they span that tangent space, split $\mathbb{C}^m$ as $A \oplus B \oplus C$ with $A \oplus B = \operatorname{im} P$ and $A \oplus C = \mathbf{1}^\perp$, so that the second symmetric power $S^2\mathbb{C}^m$ equals

$$S^2(A) \oplus S^2(B) \oplus S^2(C) \oplus (A \otimes B) \oplus (A \otimes C) \oplus (B \otimes C).$$

The matrices in the lemma span $S^2(A) + A \otimes B + (A \oplus B) \otimes C$. This space has dimension $\binom{r}{2} + (r-1) + r(n-r)$, which equals $\binom{r+1}{2} + r(n-r) - 1 = \dim \mathcal{SM}_r$. □

By Lemma 3.3.3, it suffices to understand the derivative $\sum_{i \le j} \frac{X_{ij}u_{ij}}{P_{ij}}$ for X equal to $vw^T + wv^T$, in which case it equals

$$\sum_{i \le j} \frac{X_{ij}u_{ij}}{P_{ij}} = \sum_{i \le j} (v_iw_j + w_iv_j) \frac{u_{ij}}{P_{ij}} = v^T \begin{bmatrix} \frac{2u_{11}}{P_{11}} & \frac{u_{12}}{P_{12}} & \cdots & \frac{u_{1m}}{P_{1m}} \\ \frac{u_{12}}{P_{12}} & \frac{2u_{22}}{P_{22}} & & \\ \vdots & & \ddots & \\ \frac{u_{1m}}{P_{1m}} & & & \frac{2u_{mm}}{P_{mm}} \end{bmatrix} w.$$

The right-hand side can be concisely written as $v^T(\frac{U}{P})w$, where $\frac{U}{P}$ is the Hadamard (element-wise) quotient of U by P. So criticality of P is equivalent to the statement that $v^T(\frac{U}{P})w$ vanishes for all v, w as in Lemma 3.3.3. This, in turn, is equivalent to the condition that $v^T R^{-1} Q R^{-1} w = 0$ for all v, w as in Lemma 3.3.3. We now state and prove our duality result in the symmetric case.

Theorem 3.3.4 (ML-duality for symmetric matrices)**.** *Let $U \in \mathbb{N}^{m\times m}$ be a sufficiently general symmetric data matrix, and let P be a critical point of ℓ_U on $\mathcal{SM}_r$. Define the matrix Q by $P_{ij}Q_{ij} = U_{i+}U_{ij}U_{j+}$. Then $4Q/(U_{++})^3$ is a critical point of ℓ_U on $\mathcal{SM}_{m-r+1}$.*

As in the rectangular case, the map $P \mapsto Q' := 4Q/(U_{++})^3$ is a bijection by virtue of the symmetry in P and Q, and the same conclusions for the cricital points with positive real entries can be drawn as in the rectangular case.

Proof. The normalizing factor was dealt with in Lemma 3.3.2 and will be largely ignored in what follows. As in the proof of Lemma 3.3.3, decompose $\mathbb{C}^m$ as $A\oplus B\oplus C$ with $A\oplus B = \operatorname{im} P$ and $A\oplus C = \mathbf{1}^\perp$. So A has dimension $r-1$, C has dimension $m-r$, and B has dimension 1. We take B to be spanned by $P\mathbf{1}$, which is a non-zero scalar multiple of $R\mathbf{1}$ by Lemma 3.3.3.

First we bound the rank of Q. To do so we prove that the image of Q is contained in a space of dimension $m-r+1$. Indeed, by criticality of P we have $v^T R^{-1} Q K^{-1} w = 0$ for $w \in \mathbb{C}^m$, $v \in \operatorname{im} P$ such that $v \perp \mathbf{1}$ or $w \perp \mathbf{1}$. Taking w arbitrary and v in A, we find that $\operatorname{im} Q \subseteq (R^{-1}A)^\perp$, which has dimension $m-r+1$.

Next we show that

$$x^T R^{-1} P K^{-1} y = 0$$

for any $x \in (R^{-1}A)^\perp$ and $y \in \mathbb{C}^m$ with $x \perp \mathbf{1}$ or $y \perp \mathbf{1}$. First, suppose $x \perp \mathbf{1}$. Since $PK^{-1}y$ may be written as $a + cR\mathbf{1}$ with $a \in A$ and scalar c, we find

$$x^T R^{-1} P K^{-1} y = x^T R^{-1} a + c x^T R^{-1} R\mathbf{1} = x^T R^{-1} a + 0 = 0.$$

Otherwise, we have $y \perp \mathbf{1}$ and we may assume $x = cR\mathbf{1}$ with c a scalar. In this case, we have

$$x^T R^{-1} P K^{-1} y = c\mathbf{1}^T P K^{-1} y = c\mathbf{1}^T K K^{-1} y = \mathbf{1}y = 0,$$

where we use Lemma 3.3.1.

Let k be the rank of Q. Since $\operatorname{im} Q \subset (R^{-1}A)^\perp$ we conclude that $x^T R^{-1} P K^{-1} y = 0$ holds, in particular, for all matrices $xy^T + yx^T$ spanning the tangent space to $\mathcal{SM}_k$ at Q', so that Q' is critical. By reversing the roles of P and Q and using the involution argument at the end of the proof of Theorem 3.2.4, we conclude that for generic U the value of k equals $m-r+1$ (rather than being strictly smaller). This proves the theorem. □

3.4 Duality in the skew-symmetric case

The skew-symmetric case, while perhaps not of direct relevance to statistics, is of considerable algebro-geometric interest [23], since the variety $\mathcal{AM}_r$, consisting of skew-symmetric matrices of *even* rank r whose upper-triangular entries are non-zero and add up to 1, is (an open subset

of a hyperplane section of the affine cone over) a secant variety of the Grassmannian of 2-spaces in $\mathbb{C}^m$. Recall that we want to prove a bijection between critical points of ℓ_U on (the *intersection* of a determinantal variety with an affine hyperplane) $\mathcal{AM}_r$ and critical points of ℓ_U on the *affine translate* $\mathcal{AM}'_s$ of a determinantal variety.

A point P of $\mathcal{AM}_r$ and data matrix U will be denoted by

$$P = \begin{bmatrix} 0 & p_{12} & \cdots & p_{1m} \\ -p_{12} & 0 & & \\ \vdots & & \ddots & \\ -p_{1m} & & & 0 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 0 & u_{12} & \cdots & u_{1m} \\ u_{12} & 0 & & \\ \vdots & & \ddots & \\ u_{m1} & & & 0 \end{bmatrix}.$$

Note that U is *symmetric* rather than alternating. We fix a sufficiently general data matrix U and a critical point P for ℓ_U on $\mathcal{AM}_r$. The tangent space $T_P\mathcal{AM}_r$ equals

$$T_P\mathcal{AM}_r = \{X \in \mathbb{C}^{m\times m} \text{ skew} \mid X \ker P \subseteq \operatorname{im} P \text{ and } \sum_{i<j} x_{ij} = 0\}. \tag{3.3}$$

The derivative of ℓ_U at P in the direction X equals $\sum_{i<j} \frac{x_{ij}u_{ij}}{p_{ij}}$, up to a factor irrelevant for its vanishing. The following lemma is the skew analogue of Lemmas 3.2.1 and 3.3.1.

Lemma 3.4.1. *The vector $a = (\sum_{j<i} p_{ji} + \sum_{j>i} p_{ij})_i$ is a scalar multiple of $U\mathbf{1}$.*

Proof. We need to show that 2×2-minors of the matrix $(a|U\mathbf{1})$ are zero, and do so for the first minor. Let v_1, v_2 be the first and second column of P, respectively, and set $w_1 := (a_2, 0, \ldots, 0)$ and $w_2 := (0, -a_1, 0, \ldots, 0)$. Then each of the matrices $v_i w_i^T - w_i v_i^T$ is tangent at P to the variety of skew-symmetric rank-r matrices, and their sum

$$X = \begin{bmatrix} 0 & (a_2-a_1)p_{12} & a_2p_{13} & \cdots & a_2p_{1m} \\ -(a_2-a_1)p_{12} & 0 & -a_1p_{23} & \cdots & -a_1p_{2m} \\ -a_2p_{13} & a_1p_{23} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_2p_{1m} & a_1p_{2m} & 0 & \cdots & 0 \end{bmatrix}$$

has upper-triangular entries adding up to 0, so that X is tangent at P to $\mathcal{AM}_r$. The derivative of ℓ_U at P in the direction X, which is zero by criticality of P, equals

$$(a_2-a_1)u_{12} + a_2u_{13} + \ldots + a_2u_{1m} - a_1u_{23} - \ldots - a_1p_{2m} = a_2u_{1+} - a_1u_{2+},$$

which is the minor whose vanishing was required. □

Next we determine rank-two elements spanning $T_P\mathcal{AM}_r$. For this we introduce the skew bilinear form $\langle .,. \rangle$ on $\mathbb{C}^m$ defined by $\langle v, w\rangle = v^T S w = \sum_{i<j}(v_iw_j - v_jw_i)$, where S is the skew-symmetric matrix

$$S = \begin{bmatrix} 0 & 1 & \cdots & 1 \\ -1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ -1 & \cdots & -1 & 0 \end{bmatrix}$$

from the introduction. By elementary linear algebra, this form is non-degenerate if m is even and has a one-dimensional kernel spanned by $(1,-1,1,-1,\ldots,1) \in \mathbb{C}^m$ if m is odd.

In what follows, it will be convenient to think of skew-symmetric matrices also as elements of $\bigwedge^2 \mathbb{C}^m$ or as alternating tensors.

Lemma 3.4.2. *The tangent space $T_P\mathcal{AM}_r$ is spanned by skew-symmetric matrices of the form $vw^T - wv^T$ with $v \in \operatorname{im} P$ and $\langle v, w\rangle = 0$.*

In the proof we will use that $\operatorname{im} P$ is non-degenerate with respect to $\langle ., .\rangle$. This condition will be satisfied for general U.

Proof. The proof is similar to the symmetric case and the rectangular case: a skew-symmetric matrix X lies in the tangent space if and only if $X \ker P \subseteq \operatorname{im} P$ and $\sum_{i<j} x_{ij} = 0$. The condition $v \in \operatorname{im} P$ ensures the first property and the condition that $\langle v, w\rangle = 0$ ensures the second property.

To complete the proof, decompose $\mathbb{C}^m$ as $A \oplus C$ with $A = \operatorname{im} P$ and $\langle A, C\rangle = 0$, so that $\bigwedge^2 \mathbb{C}^m$ decomposes as $\bigwedge^2 A \oplus (A \otimes C) \oplus \bigwedge^2 C$. Taking the vector w in $v^T w - wv^T$ from C we see that $A \otimes C$ is contained in the span of the matrices in the lemma. Next we argue that a codimension-one subspace of $\bigwedge^2 A$ is also contained in their span. Indeed, the (non-zero) tensors $v^T w - wv^T \in \bigwedge^2 A$ with $v, w \in A$ perpendicular with respect to $\langle ., .\rangle$ form a single orbit under the symplectic group $\mathrm{Sp}(A) = \mathrm{Sp}_r$ (recall that r is even, so that this is a reductive group), and hence their span is an $\mathrm{Sp}(A)$-submodule of $\bigwedge^2 A$. But $\bigwedge^2 A$ splits as a direct sum of only two irreducible modules under $\mathrm{Sp}(A)$: a one-dimensional trivial module corresponding to (the restriction of) $\langle ., .\rangle$ and a codimension-one module. Hence the tensors $v^T w - wv^T$ must span that codimension-one module.

Summarizing, we find that the matrices in the lemma span a space of dimension $r(n - r) + \binom{r}{2} - 1$, which equals $\dim \mathcal{AM}_r$. □

Recall that in the alternating case the likelihood function is given by $\ell_U(P) = \prod_{i<j} p_{ij}^{u_{ij}}$. The derivative of this expression in the direction of a skew-symmetric matrix X of the form $vw^T - wv^T$ equals (up to a factor irrelevant for its vanishing)

$$\sum_{i<j} x_{ij}\frac{u_{ij}}{p_{ij}} = \sum_{i<j} \frac{u_{ij}}{p_{ij}}(v_i w_j - v_j w_i) = v^T \begin{bmatrix} 0 & \frac{u_{12}}{p_{12}} & \cdots & \frac{u_{1m}}{p_{1m}} \\ -\frac{u_{12}}{p_{12}} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{u_{m-1,m}}{p_{m-1,m}} \\ -\frac{u_{1m}}{p_{1m}} & \cdots & -\frac{u_{m-1,m}}{p_{m-1,m}} & 0 \end{bmatrix} w.$$

Define the skew matrix Q by $P * Q = U$. Then criticality of P translates into $v^T Q w = 0$ for all $v \in \operatorname{im} P$ and $w \in \mathbb{C}^m$ with $\langle v, w\rangle = 0$.

Theorem 3.4.3 (Duality for skew matrices)**.** *Let $U = (u_{ij})_{ij}$ be a sufficiently general symmetric data matrix with zeroes on the diagonal, and let P be a critical point of ℓ_U on $\mathcal{AM}_r$, where $r \in \{2, \ldots, m\}$ is even. Let $s \in \{0, \ldots, m-2\}$ be the largest even integer less than or equal to $m - r$. Define the matrix Q by $P * Q = U$. Then the skew matrix $Q' := 2Q/U_{++}$ is a critical point of ℓ_U on the translated determinantal variety $\mathcal{AM}'_s$. Moreover, the map $P \to Q'$ is a bijection between the critical points of ℓ_U on $\mathcal{AM}_r$ and those $\mathcal{AM}'_s$.*

As in the rectangular and symmetric cases, the bijection $P \to Q'$ maps real, positive critical points to real, positive critical points in such a way that the sum of the log-likelihoods of P and Q' is constant.

Proof. By construction of Q we have $v^T Q w = 0$ for all $v \in \operatorname{im} P$ and $w \in \mathbb{C}^m$ with $v^T S w = 0$. This means that the quadratic form $(v, w) \mapsto v^T Q w$ on $\operatorname{im} P \times \mathbb{C}^m$ is a scalar multiple of the quadratic form $(v, w) \mapsto v^T S w$, denoted $\langle ., . \rangle$ earlier, on that same space. The scalar is computed by computing

$$(0, -p_{12}, \ldots, -p_{1m}) Q (1, 0, \ldots, 0)^T = U_{1+}$$

and

$$(0, -p_{12}, \ldots, -p_{1m}) S (1, 0, \ldots, 0)^T = P_{1+} = a_1,$$

where a is the vector of Lemma 3.4.1. Using that lemma and the fact that $\sum_i a_i = 2$ we find that $a_1 = 2U_{1+}/U_{++}$. We conclude that the skew bilinear form associated to $B := S - \frac{2}{U_{++}} Q$ is identically zero on $\operatorname{im} P \times \mathbb{C}^m$, hence $\ker B$ contains $\operatorname{im} P$ and $\operatorname{im} B = (\ker B)^\perp$ (where $\perp$ refers to the standard bilinear form on $\mathbb{C}^m$) is contained in $\ker P = (\operatorname{im} P)^\perp$. In particular, B has rank at most s; let $k \leq s$ denote the actual rank of B.

Next we argue that $Q' := \frac{2}{U_{++}} Q$ is critical for ℓ_U on $\mathcal{AM}'_k$. By arguments similar to (but easier than) those in Lemma 3.4.2 the tangent space $T_{Q'} \mathcal{AM}'_k$ is spanned by rank-two matrices $vw^T - wv^T$ with $v \in \operatorname{im} B$ and $w \in \mathbb{C}^m$ arbitrary. Thus proving that Q' is critical boils down to proving that $v^T P w = 0$ for all $v \in \operatorname{im} B$ and $w \in \mathbb{C}^m$. But this is immediate from $\operatorname{im} B \subseteq \ker P$. Thus Q' is critical.

Furthermore, we need to show that (for generic U) the rank k of $B = S - Q'$ is equal to s rather than strictly smaller, and that the map $P \mapsto Q'$, which is clearly injective, is also surjective on the set of critical points for ℓ_U on $\mathcal{AM}'_s$. For these purposes we reverse the arguments above: assume that Q' is a critical point on $\mathcal{AM}'_k$, where k is an even integer in the range $\{0, \ldots, m-2\}$. Define $Q := \frac{U_{++}}{2} Q'$ and define P by $P * Q = U$. Also, define $B := S - Q'$. Then criticality of Q' implies that $v^T P w = 0$ for all $v \in \operatorname{im} B$ and $w \in \mathbb{C}^m$, and this implies that $\ker P \supseteq \operatorname{im} B$. Thus $l := \operatorname{rk} P$ is at most $m - k$.

Moreover, B itself lies in the tangent space $T_{Q'} \mathcal{AM}'_k$, and criticality of Q' implies that $\sum_{i<j} B_{ij} \frac{U_{ij}}{Q_{ij}} = 0$. Substituting the expression for B into this we find that

$$0 = \sum_{i<j} (1 - \frac{2}{U_{++}} Q_{ij}) \frac{U_{ij}}{Q_{ij}} = \sum_{i<j} (P_{ij} - \frac{2}{U_{++}}) = (\sum_{i<j} P_{ij}) - 1,$$

i.e., the upper-triangular entries of P add up to one. We conclude that P lies in $\mathcal{AM}_l$. Next, we argue that P is critical. Indeed, for $v \in \operatorname{im} P$ and $w \in \mathbb{C}^m$ such that $\langle v, w \rangle = (v^T S w =) 0$ we find

$$v^T Q w = v^T (\frac{U_{++}}{2} (S - B)) w = \frac{U_{++}}{2} (v^T S w - v^T B w) = 0 + 0 = 0,$$

where we have used that $\operatorname{im} P \subseteq \ker B$.

Summarizing, we have found rational maps

$$\psi_r : \operatorname{Crit}(\mathcal{AM}_r) \dashrightarrow \operatorname{Crit}(\mathcal{AM}'_{f(r)}), \qquad (P, U) \mapsto (\frac{2}{U_{++}} \cdot \frac{U}{P}, U) = (Q', U) \text{ and}$$

$$\psi'_k : \mathrm{Crit}(\mathcal{AM}'_k) \dashrightarrow \mathrm{Crit}(\mathcal{AM}_{g(k)}), \qquad (Q', U) \mapsto (\frac{2}{U_{++}} \cdot \frac{U}{Q'}, U)$$

for some map f mapping even integers $r \in \{2, \ldots, m\}$ to even integers $k \in \{0, \ldots, m-2\}$, and some map g in the opposite direction. By the argument in the proof of Theorem 3.2.4, both ψ_r and ψ'_k are birational and $g(f(r)) = r$. Hence f is a bijection, and by the above it satisfies $f(r) \leq m - r$. The only such bijection is the one that maps r to the largest even integer less than or equal to $m - r$. This concludes the proof of the theorem. □

Example 3.4.4. Now we give an explicit example illustrating dual solutions in the alternating case. When $m = 4$ the ML degree of $\mathcal{AM}_2$ is 4 [23]. When

$$U = \frac{1}{41}\begin{bmatrix} 0 & 2 & 3 & 5 \\ 2 & 0 & 7 & 11 \\ 3 & 7 & 0 & 13 \\ 5 & 11 & 13 & 0 \end{bmatrix} \text{ and } P = \begin{bmatrix} 0 & 0.0386 & 0.0978 & 0.1075 \\ -0.0386 & 0 & 0.1563 & 0.2929 \\ -0.0978 & -0.1563 & 0 & 0.3069 \\ -0.1075 & -0.2929 & -0.3069 & 0 \end{bmatrix},$$

we have P is a critical point of l_U on $\mathcal{AM}_2$ and $U_{++} = 2$. Having Q defined as $P * Q = U$, we find that $Q(= Q')$ has full rank. But in the alternating case, we consider the affine translate of a determinantal variety. We find that $B = S - Q$ equals

$$B = \begin{bmatrix} 0 & -0.2638 & 0.2518 & -0.1344 \\ 0.2638 & 0 & -0.0924 & 0.0841 \\ -0.2518 & 0.0924 & 0 & -0.0332 \\ 0.1344 & -0.0841 & 0.0332 & 0 \end{bmatrix},$$

and indeed B has rank $4 - 2 = 2$. We can actually compute the ML degree of $\mathcal{AM}'_2$ symbolically to be 4 (even with the u_{ij} treated as symbols). For the data matrix U above, the minimal polynomial for q_{34} equals $434217q_{34}^4 - 1335767q_{34}^3 + 1536717q_{34}^2 - 764049q_{34} + 127426$.

3.5 Conclusion

We have proved that a number of natural determinantal varieties of matrices are *ML-dual* to other such varieties living in the same ambient spaces. However, we have done so without formalizing what exactly we mean by ML-duality. In Chapter 4 we will give a precise definition of ML-duality. In addition a generalization of Theorem 3.4.3 will be provided by Corollary 4.2.7.

Chapter 4

Maximum Likelihood for Dual Varieties

The content of this chapter has been submitted to the *Symbolic-Numeric Computation* conference to be held July 28-31, 2014 at the East China Normal University in Shanghai, China under the same title with some minor modifications for consistency with other chapters.

4.1 Introduction

Maximum likelihood estimation (MLE) is a fundamental problem in statistics that has been extensively studied from an algebraic viewpoint [9, 11, 12, 20, 23, 26]. We continue to follow an algebraic approach to MLE in this chapter, considering statistical models for discrete data in the probability simplex as irreducible varieties X in projective space $\mathbb{P}^n$.

An *algebraic statistical model* X in $\mathbb{P}^n$ will be defined by the vanishing of homogeneous polynomials in the unknowns $p_0, p_1, \ldots, p_n$. We assume that X is an irreducible generically reduced variety. When the coordinates $p_0, p_1, \ldots, p_n$ of a point p in X are positive and sum to one, we interpret p as a probability distribution, where the probability of observing event i is p_i. We let $u = (u_0, u_1, \ldots, u_n) \in (\mathbb{C}^*)^{n+1}$ be a vector of length $n+1$. This represents our *data*. When each entry u_i of the data vector is a positive integer, we interpret u_i as the number of observations of event i. We use the notation

$$u_+ := u_0 + \cdots + u_n \text{ and } p_+ := p_0 + \cdots + p_n,$$

always assuming $u_+ \neq 0$.

The likelihood function for u is defined as

$$\ell_u(p) := p_0^{u_0} p_1^{u_1} \cdots p_n^{u_n} / p_+^{u_+}.$$

When u and p are interpreted as data and a probability distribution respectively, the likelihood of observing u with respect to the distribution p is $\ell_u(p)$ divided by a multinomial coefficient depending only on u.

For fixed data u, to determine local maxima of $\ell_u(p)$ on a statistical model and give a solution to the MLE problem, we determine all complex critical points of $\ell_u(p)$ restricted to X. Of these critical points, we find the one with positive coordinates and greatest likelihood to determine the maximum likelihood estimate $\hat{p}$. The *(algebraic) maximum likelihood estimation problem* is solved by determining all critical points of $\ell_u(p)$ on X and maximizing $\ell_u(p)$ on this set.

To find the complex critical points, we determine when the gradient of $\ell_u(p)$ is orthogonal to the tangent space of X at p. So the set of critical points is

$$\{p \in X_{reg} \text{ such that } \nabla \ell_u(p) \perp T_p X\}.$$

Because the gradient of the likelihood function (up to a scalar) equals

$$\nabla \ell_u(p) = \left[\frac{u_0}{p_0} - \frac{u_+}{p_+}, \frac{u_1}{p_1} - \frac{u_+}{p_+}, \ldots, \frac{u_n}{p_n} - \frac{u_+}{p_+}\right],$$

critical points of $\ell_u(p)$ are $p \in X$ such that

$$\left[\frac{u_0}{p_0} - \frac{u_+}{p_+}, \frac{u_1}{p_1} - \frac{u_+}{p_+}, \ldots, \frac{u_n}{p_n} - \frac{u_+}{p_+}\right] \perp T_p(X),$$

implicitly forcing the condition $p_0 p_1 \cdots p_n (p_0 + \cdots + p_n) \neq 0$.

Definition 4.1.1. Given an algebraic statistical model X in $\mathbb{P}^n$, the *maximum likelihood degree* (ML degree) of X is the number of critical points of $\ell_u(p)$ restricted to X for generic choices of data u,

$$\text{MLdegree}(X) = \#\left\{p \in X : \nabla \ell_u(p) \perp T_p(X)\right\}.$$

The main result of this chapter is to give a formulation that relates maximum likelihood estimation to a conormal variety derived from X [Theorem 4.2.5]. With this perspective, we use the dual likelihood equations [Theorem 4.3.2] to solve the MLE problem for X when only given the defining equations of its dual variety X^*.

The computations in this chapter were done using `Bertini` [5] and `Macaulay2` [16].

4.2 MLE and conormal varieties

In this section, we consider an algebraic statistical model X in $\mathbb{P}^n$ and will define X' to be an embedding of X in $\mathbb{P}^{n+1}$. We will present our first result in Theorem 4.2.5. It gives a formulation of the MLE problem in terms of conormal varieties and dual varieties. In Corollary 4.2.7 we present a bijection between critical points of the likelihood function on two different varieties. In Corollary 4.2.9 we furnish equations to solve the MLE problem when the defining equations of a conormal variety are known. We will also recall how to compute conormal varieties and dual varieties of X and X'.

Let $X \subset \mathbb{P}^n$ be a codimension c algebraic statistical model defined by homogenous polynomials $f_1, f_2, \ldots, f_k$. We let $\text{Jac}(X)$ denote the $k \times (n+1)$ matrix of partial derivatives of $f_1, \ldots, f_k$ with respect to $p_0, \ldots, p_n$. We say this is the *Jacobian of* X.

To keep track of the sum of the coordinates $p_0, p_1, \ldots, p_n$ we introduce the coordinate p_s and a hyperplane of $\mathbb{P}^{n+1}$ defined by the vanishing of the polynomial

$$H(p) := -p_0 - p_1 + \cdots - p_n + p_s. \tag{4.1}$$

If X is defined by $f_1, \ldots, f_k$, then X' in the coordinates $p_0, p_1, \ldots, p_n, p_s$ is defined by the vanishing of $f_1, \ldots, f_k$ and H. With this definition, we have the following proposition.

Proposition 4.2.1. *Suppose X is defined by the homogeneous polynomials $f_1, f_2, \ldots, f_k$. Then, the Jacobian of X' is the $(k+1) \times (n+2)$-matrix*

$$\mathrm{Jac}(X') = \begin{bmatrix} -1 & -1 & \cdots & -1 & 1 \\ & & & & 0 \\ & \mathrm{Jac}\,(X) & & & \vdots \\ & & & & 0 \end{bmatrix}.$$

The important fact about the construction of X' is that there is a bijection between the critical points of the function $\ell_u(p)$ on X and the critical points of the Laurent monomial

$$\ell'_u(p) := p_0^{u_0} p_1^{u_1} \cdots p_n^{u_n} p_s^{-u_+} \text{ on } X'$$

given by Lemma 4.2.2.

By a slight abuse of notation, the "p" in $\ell_u(p)$ and the "p" in $\ell'_u(p)$ represent two different things. The first p represents a point $[p_0 : p_1 : \cdots : p_n] \in X$, while the second represents a point $[p_0 : p_1 : \cdots : p_n : p_s] \in X'$.

Lemma 4.2.2. *There is a bijection between the critical points of the function $\ell_u(p)$ on X and the critical points of $\ell'_u(p)$ on X'. Under this bijection, $[p_0 : p_1 : \cdots : p_n] \in \mathbb{P}^n$ is a critical point of $\ell_u(p)$ on X if and only if $[p_0 : p_1 : \cdots : p_n : p_s] \in \mathbb{P}^{n+1}$ is a critical point of $\ell'_u(p)$ on X'.*

Proof. To prove this we need to show that

$$[p_0 : \cdots : p_n : p_s] \in X'_{reg} \text{ satisfies } \nabla \ell'_u(p) \perp T_p X'$$

if and only if

$$[p_0 : \cdots : p_n] \in X_{reg} \text{ satisfies } \nabla \ell_u(p) \perp T_p X.$$

By Proposition 4.2.1, it follows that $[p_0 : \cdots : p_n : p_s] \in X'_{reg}$ if and only if $[p_0 : \cdots : p_n] \in X_{reg}$. So it remains to show that $\nabla \ell'_u(p) \perp T_p X'$ if and only if $\nabla \ell_u(p) \perp T_p X$. To do so, we prove that $\nabla \ell'_u(p)$ is in the row space of $\mathrm{Jac}(X')$ implies $\nabla \ell_u(p)$ is in the row space of $\mathrm{Jac}(X)$ and vica versa. To see this, observe

$$\begin{bmatrix} \nabla \ell'_u(p) \\ \mathrm{Jac}\,(X') \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ 1 & \cdots & 1 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} \frac{u_0}{p_0} - \frac{u_+}{p_s} & \frac{u_1}{p_1} - \frac{u_+}{p_s} & \cdots & \frac{u_n}{p_n} - \frac{u_+}{p_s} & -\frac{u_+}{p_s} \\ 0 & 0 & \cdots & 0 & 1 \\ & & & & 0 \\ & \mathrm{Jac}\,(X) & & & \vdots \\ & & & & 0 \end{bmatrix}$$

Since $p_s = p_+$ we have completed the proof because the top row in the matrix above is $\left[\nabla \ell_u(p), -\frac{u_+}{p_+}\right]$. □

The *conormal variety of* X is defined to be the Zariski closure in $\mathbb{P}^n \times \mathbb{P}^n$ of the set

$$N_X := \overline{\{(p,q) : q \perp T_p X\}}.$$

To determine the defining equations of N_X, we let M denote a $(k+1) \times (n+1)$ that is an extended Jacobian whose top row is $[q_0, q_1, \ldots, q_n]$ and whose bottom rows are $\mathrm{Jac}(X)$. The defining equations of the conormal variety can be computed by taking the ideal generated by $f_1, \ldots, f_k$ and the $(c+1) \times (c+1)$-minors of M and saturating by the $c \times c$-minors of $\mathrm{Jac}(X)$.

The *dual variety* X^* is the projection of the conormal variety N_X to the dual projective space $\mathbb{P}^n$ associated to the q-coordinates. To compute the equations of the dual variety, one eliminates the unknowns $p_0, p_1, \ldots, p_n$ from the equations defining N_X. For additional information on computing conormal varieties and dual varieties see [38].

Since X' is contained in a hyperplane defined by H, the dual variety of X' is known to be a cone of X^* over the point $h := [-1 : -1 : \cdots : -1 : 1]$. The dual of X' in $\mathbb{P}^{n+1}$ is given by

$$X'^* = \overline{\{[q_0 - b_s : q_1 - b_s : \cdots : q_n - b_s : b_s] : [q_0 : \cdots : q_n] \in X^*\}}.$$

It is easy to go between the coordinates of X and coordinates of X' because there is a birational map between these two varieties. But there does not have to be a birational map between X^* and X'^* as in [Example 4.2.4]. Indeed, the dimension of X^* and X'^* are not necessarily equal. For this reason, the coordinates of the former are in $q_0, \ldots, q_n$, and the coordinates of the latter are in $b_0, \ldots, b_n, b_s$. Our notation is to let q denote a point $[q_0 : q_1 : \cdots : q_n] \in X^*$ and let b denote a point $[b_0 : b_1 : \cdots : b_n : b_s] \in X'^*$.

The next proposition shows that if given the defining equations of X^* in the unknowns $q_0, \ldots, q_n$, then we can determine the defining equations of X'^* in the unknowns $b_0, \ldots, b_n, b_s$ using the relations

$$q_0 = b_0 + b_s, q_1 = b_1 + b_s, \ldots, q_n = b_n + b_s. \tag{4.2}$$

Specifically, if $g(q_0, q_1, \ldots, q_n)$ vanishes on X^*, then $g(b_0 + b_s, b_1 + b_s, \ldots, b_n + b_s)$ vanishes on X'^*. Moreover, if given the Jacobian of X^*, we can easily determine the Jacobian of X'^* as well using the relations in (4.2).

Proposition 4.2.3. *If $g_1(q), \ldots, g_l(q)$ are defining equations for the variety $X^* \subset \mathbb{P}^n$ in coordinates $q_0, q_1, \ldots, q_n$, then the defining equations of X'^* in coordinates $b_0, b_1, \ldots, b_n, b_s$ are*

$$\begin{aligned} g_1(b_0 + b_s, b_1 + b_s, \ldots, b_n + b_s) &= 0 \\ &\vdots \\ g_l(b_0 + b_s, b_1 + b_s, \ldots, b_n + b_s) &= 0. \end{aligned}$$

Moreover, the Jacobian of X'^* is given by

$$\mathrm{Jac}\,(X'^*) = \mathrm{Jac}\,(X^*)|_{(b_0+b_s,\ldots,b_n+b_s)} \begin{bmatrix} 1 & & & 1 \\ & 1 & & \vdots \\ & & \ddots & 1 \\ & & 1 & 1 \end{bmatrix}.$$

Proof. The first part of proposition follows immediately from the relations in (4.2). By

$$\text{Jac}\,(X^*)|_{(b_0+b_s,\ldots,b_n+b_s)}$$

we mean evaluate the Jacobian of X^* at $(b_0+b_s,\ldots,b_n+b_s)$. Since the defining equations of X'^* are found by evaluating each $g_i(q)$ at $(b_0+b_s,\ldots,b_n+b_s)$, it follows by the chain rule that $\text{Jac}(X'^*)$ equals the desired matrix product. □

Example 4.2.4. Consider X in $\mathbb{P}^3$ that is defined by

$$f = 2p_0p_1p_2 + p_1^2p_2 + p_1p_2^2 - p_0^2p_{12} + p_1p_2p_{12}.$$

The Jacobian of X and the defining polynomial $g(q)$ of the dual variety X^* are

$$\text{Jac}(X) = [2p_1p_2 - 2p_0p_{12}, p_2(2p_0+2p_1+p_2+p_{12}), p_1(2p_0+p_1+2p_2+p_{12}), -p_0^2+p_1p_2]$$

and

$$g(q) = q_0^4 - 8q_0^2q_1q_2 + 16q_1^2q_2^2 - 8q_0^3q_{12} + 16q_0^2q_1q_{12} + 16q_0^2q_2q_{12} - 32q_0q_1q_2q_{12}.$$

The variety X' is defined by the two equations in $\mathbb{P}^4$,

$$f(p) = 0 \text{ and } p_s = p_0+p_1+p_2+p_{12},$$

but the dual variety X'^* is defined by one equation

$$\begin{gathered} g(b_0+b_s, b_1+b_s, b_2+b_s, b_{12}+b_s) = \\ (b_0+b_s)^4 - 8(b_0+b_s)^2(b_1+b_s)(b_2+b_s) + \\ 16(b_1+b_s)^2(b_2+b_s)^2 - 8(b_0+b_s)^3(b_{12}+b_s) + \\ 16(b_0+b_s)^2(b_1+b_s)(b_{12}+b_s) + \\ 16(b_0+b_s)^2(b_2+b_s)(b_{12}+b_s) \\ -32(b_0+b_s)(b_1+b_s)(b_2+b_s)(b_{12}+b_s). \end{gathered}$$

The Jacobian of X^* is

$$\text{Jac}(X^*) = \begin{bmatrix} 4q_0^3 - 16q_0q_1q_2 - 32q_{12}(\frac{3}{4}q_0^2 - q_0q_1 - q_0q_2 + q_1q_2) \\ -8q_0^2q_2 + 32q_1q_2^2 + 16q_0^2q_{12} - 32q_0q_2q_{12} \\ -8q_0^2q_1 + 32q_1^2q_2 + 16q_0^2q_{12} - 32q_0q_1q_{12} \\ -8q_0^3 + 16q_0^2q_1 + 16q_0^2q_2 - 32q_0q_1q_2 \end{bmatrix}^T.$$

The Jacobian of X'^* is found by evaluating $\text{Jac}(X^*)$ at $(b_0+b_s,\ldots,b_n+b_s)$ and multiplying the result on the right by the matrix

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

Now we are ready to state our first result.

Theorem 4.2.5. *Fix an algebraic statistical model X. A point*

$$\big([p_0 : p_1 : \cdots : p_n : p_s], [b_0 : b_1 : \cdots : b_n : b_s]\big) \in N_{X'}$$

satisfies the relation

$$[p_0 b_0 : p_1 b_1 : \cdots : p_n b_n : p_s b_s] = [u_0 : u_1 : \cdots : u_n : -u_+]$$

if and only if $[p_0 : p_1 : \cdots : p_n : p_s]$ is a critical point of $\ell'_u(p) = p_0^{u_0} p_1^{u_1} \cdots p_n^{u_n} p_s^{-u_+}$ on X'.

Proof. To determine critical points of $\ell'_u(p)$ on X' we find when

$$\nabla \ell'_u(p) = [\partial l'_u/\partial p_0 : \cdots : \partial l'_u/\partial p_s]$$

is orthogonal to the tangent space of X' at the point p. This is the same as determining when

$$\big([p_0 : p_1 : \cdots : p_s], \nabla \ell'_u(p)\big) \in N_{X'}.$$

As a point in projective space, we have that whenever $p_0 p_1 \cdots p_s \neq 0$ that

$$\nabla \ell'_u(p) = \left[\frac{u_0}{p_0} : \cdots : \frac{u_n}{p_n} : -\frac{u_+}{p_s}\right].$$

So we immediately have that a critical point of $\ell'_u(p)$ satisfies the desired relations when we take the coordinate-wise product of $[p_0 : p_1 : \cdots : p_s]$ and $\nabla \ell'_u(p)$. □

In summary, Theorem 4.2.5, together with Lemma 4.2.2, says if $[p, b] \in N_{X'}$ and the coordinate-wise product of p and b is

$$[p_0 b_0 : \cdots : p_n b_n : p_s b_s] = [u_0 : \cdots : u_n : -u_+], \tag{4.3}$$

then $[p_0 : \cdots : p_n]$ is a critical point of $\ell_u(p)$ on X.

Definition 4.2.6. The *extended likelihood locus of X* for the data u is defined as the set of points in $N_{X'}$ satisfying the relations in (4.3), notated $E_X(u)$. We define $\mathcal{P}_u$ and $\mathcal{B}_u$ to be

$$\mathcal{P}_u := \{p : (p, b) \in E_X(u)\} \text{ and } \mathcal{B}_u := \{b : (p, b) \in E_X(u)\}.$$

For additional clarification, note that points in $E_X(u)$ are contained in the conormal variety $N_{X'} \subset \mathbb{P}^{n+1} \times \mathbb{P}^{n+1}$. These points are expressed as

$$(p, b) = \big([p_0 : p_1 : \cdots : p_s], [b_0 : b_1 : \cdots : b_s]\big) \in E_X(u).$$

In regards to ML degree, we have for generic choices of u

$$\text{MLdegree}(X) = \#E_X(u) = \#\mathcal{P}_u = \#\mathcal{B}_u.$$

There are two corollaries to Theorem 4.2.5. The first corollary gives a bijection between critical points of $\ell'_u(p)$ on X' and critical points of $\ell'_u(b)$ on X'^*. The second corollary gives equations to determine critical points of $\ell'_u(p)$ on X'.

Corollay 4.2.7. *There is a bijection between critical points of $\ell'_u(p)$ on X' and critical points of $\ell'_u(b)$ on X'^* given by (4.3). Moreover, the product $\ell'_u(p)\ell'_u(b)$ remains constant over $E_X(u)$.*

Proof. The first part follows by noticing that the relation forces us to have

$$[p_0 : p_1 : \cdots : p_s] = [u_0/b_0 : u_1/b_1 : \cdots : -u_+/b_s]$$

which is also the gradient of $\ell'_u(b)$. The second part follows as on $E_X(u)$ we have

$$\ell'_u(p)\ell'_u(b) = u_0^{u_0}u_1^{u_1}\cdots u_n^{u_n}(-u_+)^{-u_+}.$$

□

When $u_0, \ldots, u_n$ are positive integers, the bijection in Corollary 4.2.7 pairs positive critical points of $\ell'_u(p)$ ordered by increasing likelihood with positive critical points of $\ell'_u(b)$ ordered by decreasing likelihood!

Example 4.2.8. We will compute the ML degree of X in Example 4.2.4 to be 3. We fix the data vector $(u_0, u_1, u_2, u_{12}) = \frac{1}{40}(2, 13, 5, 20)$, and determine the points of $E_X(u)$ as follows:

p_0	p_1	p_2	p_{12}	p_s
.167493	.242186	.0532836	.537037	1
−.485608	.632011	.35886	.494736	1
−2.58189	5.56009	6.19312	−8.17133	1
b_0	b_1	b_2	b_{12}	b_s
.29852	1.34194	2.34594	.931035	−1
−.102964	.514232	.348325	1.01064	−1
−.0193657	.0584523	.0201837	−.0611895	−1.

The eliminants for p_0, p_1, p_2, and p_{12} are

$$(100p_0^3 + 290p_0^2 + 74p_0 - 21),$$
$$(62700p_1^3 - 403430p_1^2 + 314358p_1 - 53361),$$
$$(1900p_2^3 - 12550p_2^2 + 4886p_2 - 225),$$
$$(62700p_{12}^3 + 447650p_{12}^2 - 511962p_{12} + 136125).$$

The eliminants for b_0, b_1, b_2, b_{12} of $E_X(u)$ are

$$(1680b_0^3 - 296b_0^2 - 58b_0 - 1),$$
$$(34151040b_1^3 - 65386464b_1^2 + 27271868b_1 - 1377519),$$
$$(28800b_2^3 - 78176b_2^2 + 25100b_2 - 475),$$
$$(272250b_{12}^3 - 511962b_{12}^2 + 223825b_{12} + 15675).$$

Note that we are *not* saying the ML degree of X equals the ML degree of X^*. In general,

$$\text{MLdegree}(X) \neq \text{MLdegree}(X^*).$$

The reason why equality fails is because

$$b_0 + b_1 + \cdots + b_n - b_s$$

does not vanish on X'^*. So there is no analog of Lemma 4.2.2 involving X'^* and X^*. In terms of the previous chapters, one should think of Corollary 4.2.7 as a generalization of Theorem 3.4.3 and not as a generalization of Theorem 4.2.5.

Corollay 4.2.9. *Fix a point $[p, b]$ of $N_{X'}$ such that $p_s b_s \neq 0$. The following are equivalent:*

1. *The point $[p, b]$ is in $E_X(u)$.*
2. *The point $[p, b]$ satisfies*
$$u_i p_s b_s = -u_+ p_i b_i \text{ for } i = 0, 1, 2, \ldots, n$$
3. *There exists $[q_0 : \cdots : q_n] \in X^*$ such that*
$$u_i p_s b_s = -u_+ p_i (q_i - b_s) \text{ for } i = 0, 1, \ldots, n$$

Proof. It is immediate that part 1 and part 2 are equivalent. To see part 2 and part 3 are equivalent, recall $q_i = b_i + b_s$ for $i = 0, 1, \ldots, n$, from the definition of X'^*. □

A consequence of these equations is that it removes the need for saturation by $p_0 p_1 \cdots p_n$ with Gröbner basis computations that involve the likelihood equations whenever the u_i are nonzero. In addition, if we restrict to the affine charts defined by $p_s = 1$ and $b_s = -u_+$, then the condition $p_s b_s \neq 0$ is immediately satisfied.

4.3 Dual likelihood equations

In this section we will define a system of equations whose solutions are precisely

$$\mathcal{B}_u = \{b : (p, b) \in E_X(u)\}.$$

Once we know the set $\mathcal{B}_u$, we determine the critical points of $\ell_u(p) = p_0^{u_0} \cdots p_n^{u_n} / p_+^{u_+}$ on X using Lemma 4.2.2 and Corollary 4.2.9. For this reason we have the following definition.

Definition 4.3.1. The *dual maximum likelihood estimation problem* for the algebraic statistical model X and data u is to determine $\mathcal{B}_u$, the set of critical points of $\ell'_u(b)$ on X'^*.

By Corollary 4.2.7, we find the critical points of $\ell'_u(b) = b_0^{u_0} b_1^{u_1} \cdots b_n^{u_n} b_s^{-u_+}$ on X'^* to determine the set $\mathcal{B}_u$. That is, we determine the points $b \in X'^*$ such that the gradient

$$\nabla \ell'_u(b) = \left[\frac{u_0}{b_0} : \frac{u_1}{b_1} : \cdots : \frac{u_n}{b_n} : \frac{-u_+}{b_s} \right]$$

is orthogonal to the tangent space of X'^* at b.

If X^* in $\mathbb{P}^n$ has codimension c, which means X'^* in $\mathbb{P}^{n+1}$ has codimension c, then the *dual likelihood equations* are obtained by taking the sum of ideals generated by

- the polynomials defining X'^*, and
- the $(c+1) \times (c+1)$ minors of an extended Jacobian multiplied by a diagonal matrix with entries $b_0, b_1, \ldots, b_n, b_s$,

$$\begin{bmatrix} \nabla l_u(b) \\ \operatorname{Jac}(X'^*) \end{bmatrix} \begin{bmatrix} b_0 & & \\ & \ddots & \\ & & b_s \end{bmatrix}, \tag{4.4}$$

and saturating by the product of two ideals,

- the principal ideal generated by $b_0 b_1 \cdots b_n b_s$, and
- the ideal generated by the $c \times c$-minors of $\operatorname{Jac}(X'^*)$.

This gives us a formulation of the dual likelihood equations. Now we make some simplifications to these equations to get Theorem 4.3.2.

By Euler's relations for partial derivatives, the columns of the matrix product in (4.4) are linearly dependent. Indeed the columns sum to zero, so we may drop the last column of the product without effecting the rank.

By Proposition 4.2.3, if $g_1(q), \ldots, g_l(q)$ define the variety X^*, then the defining equations of X'^* are

$$\begin{aligned} g_1(b_0 + b_s, b_1 + b_s, \ldots, b_n + b_s) &= 0 \\ &\vdots \\ g_l(b_0 + b_s, b_1 + b_s, \ldots, b_n + b_s) &= 0. \end{aligned}$$

and the Jacobian of X'^* is

$$\operatorname{Jac}(X'^*) = \operatorname{Jac}(X^*)|_{(b_0+b_s,\ldots,b_n+b_s)} \begin{bmatrix} 1 & & & 1 \\ & 1 & & \vdots \\ & & \ddots & 1 \\ & & 1 & 1 \end{bmatrix}.$$

Since the last column of $\operatorname{Jac}(X'^*)$ is the sum of the first columns, it follows the dual likelihood equations can be calculated by the next theorem.

Theorem 4.3.2. *Let $g_1(q), \ldots, g_l(q)$ define $X^* \subset \mathbb{P}^n$ with codimension c. Then the* dual likelihood equations of X *are calculated by taking the sum of the ideals generated by*

- $g_1(b_0 + b_s, \ldots, b_n + b_s), \ldots, g_l(b_0 + b_s, \ldots, b_n + b_s)$ *and*
- *the $(c+1) \times (c+1)$ minors of*

$$\begin{bmatrix} \frac{u_0}{b_0} & \frac{u_1}{b_1} & \cdots & \frac{u_{n-1}}{b_{n-1}} & \frac{u_n}{b_n} \\ & & \operatorname{Jac}(X^*)|_{(b_0+b_s,\ldots,b_n+b_s)} & & \end{bmatrix} \begin{bmatrix} b_0 & & \\ & \ddots & \\ & & b_n \end{bmatrix},$$

and saturating by the product of two ideals,

- *the principal ideal generated by* $b_0 b_1 \cdots b_n b_s$, *and*
- *the ideal of* $c \times c$*-minors of* $\operatorname{Jac}(X^*)|_{(b_0+b_s,\ldots,b_n+b_s)}$.

The point of Theorem 4.3.2, is that the dual likelihood equations define a homogenous ideal in the polynomial ring $\mathbb{C}[b_0, b_1, \ldots, b_n, b_s]$ whose variety is $\mathcal{B}_u$, the set of critical points of $\ell'_u(b)$ on X'^*.

Example 4.3.3. Let X be defined by $f(p) = 4p_0p_2 - p_1^2$ in $\mathbb{P}^2$. Then X^* is defined by $g(q) = q_0q_2 - q_1^2$ in $\mathbb{P}^2$. So

$$f(p) = \det \begin{bmatrix} 2p_0 & p_1 \\ p_1 & 2p_2 \end{bmatrix} \text{ and } g(q) = \det \begin{bmatrix} q_0 & q_1 \\ q_1 & q_2 \end{bmatrix}.$$

The dual likelihood equations are computed by taking the ideal generated by

- $g(b_0 + b_s, b_1 + b_s, b_2 + b_s) = (b_0 + b_s)(b_2 + b_s) - (b_1 + b_s)^2$, and
- 2×2 minors of

$$\begin{bmatrix} \frac{u_0}{b_0} & \frac{u_1}{b_1} & \frac{u_2}{b_2} \\ (b_2 + b_s) & -2(b_1 + b_s) & (b_0 + b_s) \end{bmatrix} \begin{bmatrix} b_0 & & \\ & b_1 & \\ & & b_2 \end{bmatrix}$$

and saturating by the product of two ideals

- the principal ideal $(b_0b_1b_2b_s)$ and
- the 1×1 minors of

$$\begin{bmatrix} (b_2 + b_s) & -2(b_1 + b_s) & (b_0 + b_s) \end{bmatrix}.$$

We find that there is a unique critical point of $\ell'_u(b)$ on X'^* whose coordinates are derived from the matrix equality

$$\frac{1}{b_s} \begin{bmatrix} b_0 & b_1 \\ b_1 & b_2 \end{bmatrix} = \begin{bmatrix} \frac{4u_0u_+}{(2u_0+u_1)^2} & \frac{4u_1u_+}{2(u_1+2u_2)(2u_0+u_1)} \\ \frac{4u_1u_+}{2(u_1+2u_2)(2u_0+u_1)} & \frac{4u_2u_+}{(2u_2+u_1)^2} \end{bmatrix}.$$

So by Corollary 4.2.9, the coordinates of the critical point of $\ell_u(p)$ on X are derived from the matrix equality

$$\frac{1}{p_s} \begin{bmatrix} 2p_0 & p_1 \\ p_1 & 2p_2 \end{bmatrix} = \frac{1}{2u_+^2} \begin{bmatrix} (2u_0 + u_1) \\ (u_1 + 2u_2) \end{bmatrix} \begin{bmatrix} (2u_0 + u_1) \\ (u_1 + 2u_2) \end{bmatrix}^T.$$

To calculate ML degrees when X^* is not a complete intersection [Computation 4.3.5], we will work with an adjusted formulation of the dual likelihood equation. This formulation introduces codimension X^* auxiliary unknowns (Lagrange multipliers). Also, instead of working with every generator of the ideal of X^*, we work with codimension X^* generators. These generators should be chosen so that they define a reducible variety whose only irreducible component not contained in the coordinate hyperplanes is X^*.

Example 4.3.4. Consider $2\times 2\times 2$-tensors of the form $[p_{ijk}]$ with $i,j,k,\in\{0,1\}$. If X is the hyperdeterminant of these tensors, then X^* is defined by the 2×2 minors of all flattenings of the tensor $[q_{ijk}]$. The codimension of X^* is 4. The 4 minors below define X^* after saturating by q_{111}.

$$g_1(q) = q_{011}q_{101} - q_{001}q_{111} \quad g_2(q) = q_{011}q_{110} - q_{010}q_{111}$$
$$g_3(q) = q_{001}q_{110} - q_{000}q_{111} \quad g_4(q) = q_{011}q_{101} - q_{001}q_{111}$$

So by introducing auxiliary unknowns $\lambda_0,\lambda_1,\lambda_3,\lambda_4$ we create a square system of 12 equations in the homogeneous variable groups $(b_{000},\ldots,b_{111},b_s)$ and $(\lambda_0,\ldots,\lambda_4)$:

$$\begin{aligned} g_1 &= (b_{011}+b_s)(b_{101}+b_s) - (b_{001}+b_s)(b_{111}+b_s)\\ g_2 &= (b_{011}+b_s)(b_{110}+b_s) - (b_{010}+b_s)(b_{111}+b_s)\\ g_3 &= (b_{001}+b_s)(b_{110}+b_s) - (b_{000}+b_s)(b_{111}+b_s)\\ g_4 &= (b_{011}+b_s)(b_{101}+b_s) - (b_{001}+b_s)(b_{111}+b_s) \end{aligned}$$

$$[\lambda_0,\lambda_1,\lambda_2,\ldots,\lambda_4]\begin{bmatrix} \nabla \ell_u'(b) \\ \mathrm{Jac}(g) \end{bmatrix}\begin{bmatrix} b_{000} & & \\ & \ddots & \\ & & b_{111} \end{bmatrix} = 0.$$

The solutions with $\lambda_0 b_s \neq 0$ give the critical points. We find that there are 13 critical points of $\ell_u'(b)$ on X^*. For more details on the statistical model X we refer to [12], Example 2.2.10.

The next example is a new computational result to determine the ML degree of a hyperdeterminant.

Computation 4.3.5. Let X denote the hyperdeterminant of $2\times 2\times 3$ tensors of the form $[p_{ijk}]$ for $i\in\{0,1\}$, $j\in\{0,1\}$, $k\in\{0,1,2\}$. Then the ML degree of X is 71.

Proof. The variety X is dual to the variety X^* defined by the 2×2-minors of the flattenings of the $2\times 2\times 3$ tensor $[q_{ijk}]$ with $i\in\{0,1\}$, $j\in\{0,1\}$, $k\in\{0,1,2\}$. The variety X^* has codimension 7, degree 12, and 24 generators. We consider 7 of the 24 generators,

$$\begin{aligned} g_1(q) &= q_{102}q_{111} - q_{101}q_{112}\\ g_2(q) &= q_{102}q_{110} - q_{100}q_{112}\\ g_3(q) &= q_{002}q_{111} - q_{001}q_{112}\\ g_4(q) &= q_{012}q_{102} - q_{002}q_{112}\\ g_5(q) &= q_{012}q_{111} - q_{011}q_{112}\\ g_6(q) &= q_{012}q_{110} - q_{010}q_{112}\\ g_7(q) &= q_{002}q_{110} - q_{000}q_{112} \end{aligned}$$

such that when saturated by q_{112} we recover the dual variety X^*. We solve the following square system of equations: the seven equations

$$g_1(b_0+b_s,\ldots,b_{112}+b_s) = \cdots = g_7(b_0+b_s,\ldots,b_{112}+b_s) = 0$$

and the 12 equations

$$[1,\lambda_1,\lambda_2,\ldots,\lambda_7]M\begin{bmatrix} b_{101} & & \\ & \ddots & \\ & & b_{112} \end{bmatrix} = 0,$$

with M in (4.5) for a choice of u consisting of random complex numbers to determine the ML degree of X is 71.

(4.5)

$$\left[\begin{array}{cccccccccccc}
\frac{u_{101}}{b_{101}} & \frac{u_{011}}{b_{011}} & \frac{u_{100}}{b_{100}} & \frac{u_{010}}{b_{010}} & \frac{u_{001}}{b_{001}} & \frac{u_{000}}{b_{000}} & \frac{u_{002}}{b_{002}} & \frac{u_{012}}{b_{012}} & \frac{u_{102}}{b_{102}} & \frac{u_{110}}{b_{110}} & \frac{u_{111}}{b_{111}} & \frac{u_{112}}{b_{112}} \\
-q_{112} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_{111} & 0 & q_{102} & -q_{101} \\
 & -q_{112} & 0 & 0 & 0 & 0 & 0 & q_{111} & 0 & 0 & q_{012} & -q_{011} \\
 & & -q_{112} & 0 & 0 & 0 & 0 & 0 & q_{110} & q_{102} & 0 & -q_{100} \\
 & & & -q_{112} & 0 & 0 & 0 & q_{110} & 0 & q_{012} & 0 & -q_{010} \\
 & & & & -q_{112} & 0 & q_{111} & 0 & 0 & 0 & q_{002} & -q_{001} \\
 & & & & & -q_{112} & q_{110} & 0 & 0 & q_{002} & 0 & -q_{000} \\
 & & & & & & -q_{112} & q_{102} & q_{012} & 0 & 0 & -q_{002}
\end{array}\right]\Bigg|_{(b_{101}+b_s,\ldots,b_{112}+b_s)}$$

□

4.4 The dual MLE problem and ML duality

In this section we introduce two examples and show how the results presented in this chapter fit in context with Chapter 3.

Definition 4.4.1. A pair of algebraic statistical models X and Y in $\mathbb{P}^n$ are said to be *ML-dual* if for generic u there is a involutive bijection between points of $E_X(u)$ and points of $E_Y(u)$. Moreover, this bijection pairs points of $E_X(u)$ with points of $E_Y(u)$ such that the coordinate-wise product of each pair can be expressed terms of the data u alone.

Example 4.4.2. Suppose $r \leq m \leq n$, and let $V_{m,n,r}$ denote the Zariski closure in $\mathbb{P}^{mn-1}$ of rank r matrices of the form

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{m1} & & & p_{mn} \end{bmatrix}.$$

Then $V^*_{m,n,r}$ is known to be the Zariski closure in $\mathbb{P}^{mn-1}$ of rank $m-r$ matrices of the form

$$\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & & \\ \vdots & & \ddots & \\ q_{m1} & & & q_{mn} \end{bmatrix}.$$

Fix a choice of m, n, r. If we take $X = V_{m,n,r}$, then points in X' will be represented as

$$[p_{ij} : p_s] \in X' \subset \mathbb{P}^{mn}$$

and points in X'^* will be represented as

$$[b_{ij} : b_s] \in X'^* \subset \mathbb{P}^{mn}.$$

With Corollary 4.2.7, it follows that there is a bijection between $\mathcal{P}_u$ and $\mathcal{B}_u$ for $X = V_{m,n,r}$.

On the other hand, we know that $V_{m,n,r}$ and $V_{m,n,m-r}$ are ML-dual by Theorem 3.1.1. This means, if we take Y to be $V_{m,n,m-r}$ there is a involutive bijection between critical points

$E_X(u)$ and $E_Y(u)$ for generic choices of u. In particular, the bijection is such that the coordinate-wise product of the paired points is

$$\left(\left[\frac{u_{i+}u_{+j}u_{ij}}{u_{++}^3} : 1\right], \left[\frac{u_{ij}u_{++}}{u_{i+}u_{+j}} : 1\right]\right) \in \mathbb{P}^{mn} \times \mathbb{P}^{mn}.$$

Here $u_{++} := \sum_{i,j} u_{ij}$, $u_{i+} := \sum_j u_{ij}$, and $u_{+j} := \sum_i u_{ij}$, and likewise for p_{++}, p_{i+}, p_{+j}.

In the next example will provide a another pair of varieties that are ML-dual. Afterwards we generalize the example to provide a family of statistical models for which we can find ML-duals.

Example 4.4.3. Let X in $\mathbb{P}^3$ be defined by

$$\begin{aligned} f_1(p) &= (3p_0^2 + 22p_0p_1 + 3p_1^2 - 6p_0p_2 - 6p_1p_2 + 23p_2^2 \\ &\quad -22p_0p_3 - 22p_1p_3 - 34p_2p_3 + 39p_3^2) \end{aligned}$$

and fix

$$(v_0, v_1, v_2, v_3, v_s) = (1, 1, 1, 1, 4) \text{ and } (u_0, u_1, u_2, u_3) = (2, 5, 9, 7).$$

Then, the variety X' is a cone over the point

$$v = [v_0 : v_1 : v_2 : v_3 : v_s],$$

and also contained in the hyperplane defined by

$$p_0 + p_1 + p_2 + p_3 - p_s.$$

The defining equations of X'^* are

$$\begin{aligned} g_1(b) &= (b_0 + b_1 + b_2 + b_3 + 4b_s) \\ g_2(b) &= -152b_1^2 - 152b_1b_2 + b_2^2 - 152b_1b_3 - 42b_2b_3 + \\ &\quad -15b_3^2 - 608b_1b_s - 192b_2b_s - 224b_3b_s - 512b_s^2 \end{aligned} \tag{4.6}$$

Note that X'^* is contained in the hyperplane defined by

$$g_1(b) = v_0b_0 + v_1b_1 + v_2b_2 + v_3b_3 + v_sb_s.$$

This is because X' is a cone over the point v. In addition, since X' is contained in a hyperplane then X'^* is a cone as well. In this case X'^* is a cone over the point

$$[1, 1, 1, 1 - 1].$$

Now, let Y be defined the polynomial

$$32p_0^2 - 88p_0p_1 + 32p_1^2 + 16p_0p_2 + 16p_1p_2 - 17p_2^2 + 8p_0p_3 + 8p_1p_3 + 2p_2p_3 - 9p_3^2. \tag{4.7}$$

Doing so, it follows that Y' is a linear change of coordinates of X'^*. If we replace b_0, b_1, b_2, b_3, b_s with $p_0, p_1, p_2, p_3, \frac{-1}{4}p_s$ respectively in (4.6), then we have the defining equations of Y':

$$\begin{gathered} p_0 + p_1 + p_2 + p_3 - p_s \text{ and} \\ 32p_0^2 - 88p_0p_1 + 32p_1^2 + 16p_0p_2 + 16p_1p_2 - 17p_2^2 + 8p_0p_3 + 8p_1p_3 + 2p_2p_3 - 9p_3^2. \end{gathered}$$

Eliminating the unknown p_s from Y' yields the relation (4.7) defining Y.

The varieties X and Y in $\mathbb{P}^n$ are ML-dual, meaning there is a bijection between the points of $E_X(u)$ and $E_Y(u)$. The bijection pairs a point of $E_X(u)$ with a point of $E_Y(u)$ such that the coordinate-wise product of these two points is

$$\left(\left[\frac{u_0}{v_0} : \cdots : \frac{u_3}{v_3} : \frac{-u_+}{v_s} \right], \left[\frac{u_0}{v_0^{-1}} : \cdots : \frac{u_3}{v_3^{-1}} : \frac{-u_+}{v_s^{-1}} \right] \right).$$

For the fixed choice of $u = (2, 5, 9, 7)$ we find that there are 12 points in $E_X(u)$ and 12 points in $E_Y(u)$.

Theorem 4.4.4. *Fix an algebraic statistical model X. Suppose X' is a cone over the point $v = [v_0 : v_1 : \cdots : v_n : v_s] \in (\mathbb{C}^*)^{n+2}$. Then, there exists an algebraic statistical model Y that is ML dual to X. Explicitly, if the defining equations of X'^* are*

$$\begin{array}{c} g_1(b_0, b_1, \ldots, b_n, b_s), \\ g_2(b_0, b_1, \ldots, b_n, b_s), \\ \vdots \\ g_k(b_0, b_1, \ldots, b_n, b_s), \end{array}$$

then the defining equations of Y' are

$$\begin{array}{c} g_1\left(\frac{p_0}{v_0}, \frac{p_1}{v_1}, \ldots, \frac{p_n}{v_n}, \frac{p_s}{v_s} \right), \\ g_2\left(\frac{p_0}{v_0}, \frac{p_1}{v_1}, \ldots, \frac{p_n}{v_n}, \frac{p_s}{v_s} \right), \\ \vdots \\ g_k\left(\frac{p_0}{v_0}, \frac{p_1}{v_1}, \ldots, \frac{p_n}{v_n}, \frac{p_s}{v_s} \right). \end{array}$$

To determine Y, eliminate p_s from the equations defining Y'. Moreover, there is a bijection between $E_X(u)$ and $E_Y(u)$. This bijection pairs a critical point of $E_X(u)$ with a critical point of $E_Y(u)$ such that coordinate-wise product of the pair is given by the relation

$$\left(\left[\frac{u_0}{v_0} : \frac{u_1}{v_1} : \cdots : \frac{u_n}{v_n} : \frac{-u_+}{v_s} \right], \left[\frac{u_0}{v_0^{-1}} : \frac{u_1}{v_1^{-1}} : \cdots : \frac{u_n}{v_n^{-1}} : \frac{-u_+}{v_s^{-1}} \right] \right).$$

It is too strong of a hypothesis to expect a statistical model to be a cone over a point. But the following example shows why Theorem 4.4.4 is still of great interest. It shows that the critical points of the likelihood function over common statistical models can be found by determining critical points of the likelihood function over cone.

Example 4.4.5. Fix u to be the data

$$(u_{11}, u_{12}, u_{13}, u_{22}, u_{23}, u_{33}) = (10, 9, 1, 21, 3, 7). \tag{4.8}$$

Let X' be defined by $p_s = p_{11} + p_{12} + p_{13} + p_{22} + p_{23} + p_{33}$ and

$$\begin{array}{rcl} f_1(p) & = & (-p_{12} + 3p_{13} - 2p_{22} + 2p_{23} + 6p_{33}) \\ f_2(p) & = & (6p_{11} + 7p_{13} - 6p_{22} + p_{23} + 8p_{33}) \\ f_3(p) & = & (6p_{13}p_{22} - 18p_{13}p_{23} + 6p_{22}p_{23} - p_{23}^2 + \\ & & \quad 54p_{13}p_{33} - 56p_{22}p_{33} + 18p_{23}p_{33} + 108p_{33}^2). \end{array}$$

Here, X' is a cone over the point $v = [v_{11} : v_{12} : v_{13} : v_{22} : v_{23} : v_{33} : v_s]$ such that

$$\begin{aligned} v_{11} &= (2u_{11} + u_{12} + u_{13})^2 \\ v_{12} &= 2(u_{12} + 2u_{22} + u_{23})(2u_{11} + u_{12} + u_{13}) \\ v_{13} &= 2(u_{13} + u_{23} + 2u_{33})(2u_{11} + u_{12} + u_{13}) \\ v_{22} &= (u_{12} + 2u_{22} + u_{23})^2 \\ v_{23} &= 2(u_{13} + u_{23} + 2u_{33})(u_{12} + 2u_{22} + u_{23}) \\ v_{33} &= (u_{13} + u_{23} + 2u_{33})^2 \\ v_s &= -4(u_{11} + u_{12} + u_{13} + u_{22} + u_{23} + u_{33})^2. \end{aligned}$$

The dual variety X'^* is defined by the polynomials

$$\begin{aligned} g_1(b) = \;& v_{11}b_{11} + v_{12}b_{12} + v_{13}b_{13} + v_{22}b_{22} + v_{23}b_{23} + v_s b_s \\ g_2(b) = \;& 2(b_{12}b_{13} - b_{13}b_{22} - b_{11} + b_{12}b_{23} + b_{13}b_{23} - b_{12}b_{33}) \\ & -b_{13}^2 + b_{11}b_{22} - b_{12}^2 + b_{23} - b_{23}^2 + b_{11}b_{33} + b_{22}b_{33}. \end{aligned}$$

If we take Y' to be as in Theorem 4.4.4, then Y is ML dual to X. We determine there are 15 points in $E_X(u)$ and 15 points in $E_Y(u)$.

Six of these 15 points are even more interesting when we consider the algebraic statistical model Z defined by the determinant of

$$\begin{bmatrix} 2p_{11} & p_{12} & p_{13} \\ 2p_{12} & 2p_{22} & p_{23} \\ 2p_{13} & p_{23} & 2p_{33} \end{bmatrix}.$$

So Z consists of symmetric matrices of rank at most 2. The extended likelihood locus $E_Z(u)$ for the prescribed data (4.8) consists of six points

$$E_Z(u) = \{z_1, z_2, \ldots, z_6\}.$$

The surprising result is that the set $E_Z(u)$ is a subset of $E_X(u)$.

From this final example we saw that we can determine critical points of standard algebraic statistical models by considering critical points of varieties which are cones.

4.5 Conclusion

In this chapter, we have given an elegant formulation of the MLE problem involving conormal varieties. This formulation allows one to forgo the expensive computation of saturation by a product of unknowns. We also define the dual likelihood equations that allows one to compute critical points on X even if the defining equations of X are not known using a dual variety. We showed that if we solve the dual likelihood equations equations, we recover the critical points on X by Theorem 4.2.5. More broadly, we showed that if there is a bijection between critical points of a function restricted to a variety and critical points of a Laurent monomial restricted to a different variety, then we can formulate a new set of equations to determine these points.

Chapter 5

Maximum likelihood geometry in the presence of data zeros

The content of this chapter has been submitted to the *International Symposium on Symbolic and Algebraic Computation* to be held July 23-25, 2014 at Kobe University, Japan under the same title. This is joint work with Elizabeth Gross with some minor modifications for consistency with other chapters.

5.1 Introduction

The method of maximum likelihood estimation for a statistical model $\mathcal{M}$ and an observed data vector $u \in \mathbb{R}^{n+1}$ involves maximizing the likelihood function l_u over all distributions in $\mathcal{M}$. This involves understanding the zero-set of a system of equations, and, thus, when the models of interest are algebraic, the process lends itself to investigation using algebraic geometry. In fact, likelihood geometry has been studied in a series of papers in the field of algebraic statistics beginning with [9] and [23]. Subsequent papers include [7, 25, 17, 20, 45, 26] covering both discrete and continuous models. In this chapter, we look at discrete models and the case where the observed data vector contains zero entries.

In [23], Hoşten, Khetan, and Sturmfels introduce the likelihood locus and its associated incidence variety for discrete statistical models. In [27], Huh and Sturmfels study this incidence variety further under the name of the *likelihood correspondence*. Given a discrete algebraic statistical model with sample space of size $n+1$ and Zariski closure X, the likelihood correspondence $\mathcal{L}_X$ is a closed algebraic subset of $\mathbb{P}^n \times \mathbb{P}^n$. We view $\mathbb{P}^n \times \mathbb{P}^n$ as the product of the probability space $\mathbb{P}^n_p$ with homogeneous coordinates $p_0, p_1, \ldots, p_n$ and the data space $\mathbb{P}^n_u$ with homogeneous coordinates $u_0, u_1, \ldots, u_n$. In this chapter, we are concerned with special fibers of the projections $pr_1 : \mathcal{L}_X \to \mathbb{P}^n_p$ and $pr_2 : \mathcal{L}_X \to \mathbb{P}^n_u$. Specifically, we set out to understand $pr_2^{-1}(u)$ when u contains zero entries and show how our understanding of $pr_2^{-1}(u)$ yields information about generic fibers of pr_2. In particular, we want to understand the degree of a generic fiber of pr_2. That quantity is the *ML degree* (maximum likelihood degree) of X as discussed in previous chapters.

A statistical model $\mathcal{M}$ is a subset of the probability simplex

$$\Delta_n = \left\{ (p_0, p_1, \ldots, p_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} p_i = 1 \text{ and } p_i \geq 0 \text{ for } i = 0, 1, \ldots n \right\}.$$

Given positive integer data $u \in \mathbb{Z}_{\geq 0}^{n+1}$, the *maximum likelihood estimation* problem is to determine $\hat{p} \in \mathcal{M}$ that maximizes the likelihood function

$$l_u = p_0^{u_0} p_1^{u_1} \cdots p_n^{u_n}$$

restricted to $\mathcal{M}$. The point $\hat{p} \in \mathcal{M}$ is called the *maximum likelihood estimate*, or mle. The family of models we are interested in are *algebraic statistical* models, which are defined by the vanishing of polynomial equations restricted to the probability simplex.

To use algebraic methods, we consider points of $\mathcal{M} \subset \mathbb{R}^{n+1}$ as representatives of points in $\mathbb{P}^n$ and study the Zariski closure $\overline{\mathcal{M}} = X \subset \mathbb{P}^n$. This makes the problem easier by relaxing the nonnegative and real constraints, which allows us to obtain an understanding about the number of possible modes of the likelihood surface. There are subtleties when performing this relaxation as mentioned for example in Section 2.5 of Chapter 2 regarding rank and non-negative rank.

Let $p_+ := p_0 + p_1 + \cdots + p_n$ and $\mathcal{H}_n$ be the set of points where $p_+ p_0 p_1 \cdots p_n$ equals zero. With algebraic methods, our goal is to determine all complex critical points of $\ell_u(p) := l_u / p_+^{u_+}$ when restricted to $X_{reg} \backslash \mathcal{H}_n \subset \mathbb{P}^n$, where X_{reg} is the set of regular points of X. We work with $\ell_u(p)$ since it is a rational function of degree zero and thus a function on $\mathbb{P}^n$ (see [12, §2.2]).

A point $p \in X_{reg}$ is said to be a *critical point* if the gradient of $\ell_u(p)$ is orthogonal to the tangent space of X at p, that is

$$\nabla \ell_u(p) \perp T_p X.$$

If the maximum likelihood estimate $\hat{p}$ for the data vector u is in the interior of $\mathcal{M}$, then $\hat{p}$ will be a critical point of $\ell_u(p)$ over X. By determining the critical points of $\ell_u(p)$ on X, we find all local maxima of l_u on $\mathcal{M}$.

If the data vector u contains zero entries, in the statistics literature each zero entry is called either a sampling zero or a structural zero. Considering u as a flattened contingency table, a sampling zero at u_i occurs when no observations fall into cell i even though p_i is nonzero. A structural zero occurs at u_i when the probability of an observation falling into cell i is zero. Structural and sampling zeros occur commonly in practice, for example, in large sparse data sets (for more on sampling and structural zeros see [6, §5.1.1]).

The terms "sampling zero" and "structural zero" are denotationally about contingency tables, but they also carry implications about X as well. For example, the term "structural zero" connotes that maximum likelihood estimation should proceed over a projection of X (see [36]). Due to this secondary definition imparted to the term "structural zero," and in view of the fact that this chapter is concerned with the intersection of X with the hyperplane $p_i = 0$ as opposed to the projection of X, we introduce the definition of a model zero.

Definition 5.1.1 (**Model zeros**). Given a model $\mathcal{M}$ with $\overline{\mathcal{M}} = X \subset \mathbb{P}^n$ and data vector u with $u_i = 0$, a *model zero* at cell i is a zero such that the maximum likelihood estimate $\hat{p}$ for u is a critical point of $\ell_u(p)$ over $X \cap \{p_i = 0\}$.

Remark 5.1.2. From the remainder of the chapter, we will use "structural zero" to mean a zero at cell i such that *1)* $p_i = 0$, and *2)* maximum likelihood estimation proceeds over the projection of X onto all coordinates except the ith coordinate.

In this chapter we explore in depth the algebraic considerations of maximum likelihood estimation when the data contains sampling zeros or model zeros. The main theorem of Section 5.2, Theorem 5.3.6, shows how solutions to the maximum likelihood estimation problem for data with zeros on X are contained in the likelihood correspondence of X. This result gives statistical meaning to the likelihood correspondence when u_i is equal to zero and has interesting theoretical and computational implications. On the theoretical side, we can use Theorem 5.3.6 to compute a lower bound on the ML degree of a variety X. On the computational side, Theorem 5.3.6 can be applied using coefficient-parameter homotopies to quickly find critical points of $\ell_u(p)$ over X (Algorithm 5.4.2) and can make the problem of computing the ML degree for multi-way tables tractable (Section 5.4).

This chapter is organized as follows. In Section 5.2, we give preliminary definitions and introduce a square parameterized system called the Lagrange likelihood equations. Proposition 5.2.2 describes the properties of the Lagrange likelihood equations that will be referenced in later sections. We also describe how the variety of the Lagrange likelihood equations is related to the likelihood correspondence of Huh and Sturmfels [27].

In Section 5.2, we discuss how sampling and model zeros change the maximum likelihood problem. Theorem 5.3.6 describes the special fiber $pr_2^{-1}(u)$ when u contains zero entries. We use this theorem to give a lower bound on the ML degree of X. The section continues with exploring how solutions to the Lagrange likelihood equations partition into solutions for different maximum likelihood estimation problems for sampling and model zeros; these partitions are captured in the ML tables introduced in this section. We end this section by fully characterizing the ML degree for different sampling and model zero configurations of a generic hypersurface of degree d in $\mathbb{P}^n$.

We conclude with Section 5.3, which illustrates several computational advantages that can be achieved in ML degree computation by first considering data vectors with zeros. Algorithm 5.4.2 gives a method to find critical points of $\ell_u(p)$ over X by computing the critical points of $\ell_u(p)$ when u contains model zeros; these solutions are significantly easier to compute. We continue the section by looking at Grassmannian and tensor examples. We conclude by extending maximum likelihood duality to u with zero entries and showing how ML duality offers further computational benefits.

5.2 Likelihood equations and ML degree

The *maximum likelihood degree* (ML degree) of a variety $X \subset \mathbb{P}^n$ is defined as the number of critical points of the likelihood function $\ell_u(p)$ on $X_{reg} \setminus \mathcal{H}_n$ for generic data u. The ML degree of X quantifies the algebraic complexity of the maximum likelihood estimation problem over the model $\mathcal{M}$, indicating how feasible symbolic algebraic methods are for finding the maximum likelihood estimate. The ML degree has an explicit interpretation in numerical algebraic geometry as well. Assuming that the *ab initio* stage of a coefficient-parameter homotopy has been run [41, §7] the ML degree is the number of paths that need to be followed for every subsequent run.

For each u, all critical points of $\ell_u(p)$ over X form a variety. Thus, by varying u over $\mathbb{P}^n_u$ we obtain a family of projective varieties with base $\mathbb{P}_u$. In algebraic geometry, the natural way to view this family of parameterized varieties is as a subvariety $\mathcal{L}_X$ of the product variety $\mathbb{P}^n_p \times \mathbb{P}^n_u$ where the elements of the family are the fibers of the canonical projection $pr_2 : \mathbb{P}^n_p \times \mathbb{P}^n_u \to \mathbb{P}^n_u$ over the points u in $\mathbb{P}^n_u$. The subvariety $\mathcal{L}_X$ is called the likelihood correspondence [27], which is the closure in $\mathbb{P}^n_p \times \mathbb{P}^n_u$ of

$$\{(p,u) : p \in X_{reg} \setminus \mathcal{H}_n \text{ and } \operatorname{dlog}(\ell_u(p)) \text{ vanishes at } p\}.$$

Just as we can talk about a parameterized family of varieties, we can also talk about a parameterized system of polynomial equations. For us, a *parameterized polynomial system* is a family $\mathcal{F}$ of polynomial equations in the variables $p_0, \ldots, p_n$ and the parameters $u_0, \ldots, u_n$. A member of the family is chosen by assigning a complex number to each parameter u_i. If u is a generic vector in $\mathbb{P}^n$, we call the resulting system *generic*. A system of equations is said to be *square* if the number of unknowns (variables) equals the number of equations of the system. Algebraic homotopies are an effective way to solve many members of a family $\mathcal{F}$. By solving a generic member of the family, we determine the solutions to another system of the family using a *coefficient-parameter homotopy* (see [34]), thus, this viewpoint can be computationally advantageous.

Now we define a parameterized square system of polynomial equations called the Lagrange likelihood equations. The Lagrange likelihood equations for a variety $X \subset \mathbb{P}^n$ of codimension c consists of $n+1+c$ equations. The $n+1+c$ unknowns are $p_0, p_1, \ldots p_n, \lambda_1, \ldots, \lambda_c$ and the parameters are $u_0, \ldots, u_n$. The advantage of the Lagrange likelihood equations, in addition to being a parameterized square system, is that properties of a point (p,u) in the likelihood correspondence become apparent. These properties are summarized in Proposition 5.2.2.

Definition 5.2.1 (**Lagrange likelihood equations**)**.** Let X be a codimension c irreducible variety. If X is an irreducible component of the variety of $h_1, h_2, \ldots, h_c$, then the Lagrange likelihood equations of X denoted by $\operatorname{LL}(X,u)$ are

$$h_1 = h_2 = \cdots = h_c = 0 \tag{5.1}$$

$$(u_+ p_i - u_i) = p_i \left(\lambda_1 \partial_i h_1 + \lambda_2 \partial_i h_2 + \cdots + \lambda_c \partial_i h_c\right) \text{ for } i = 0, \ldots, n \tag{5.2}$$

If X is a complete intersection, then $h_1, \ldots, h_c$ are the minimal generators of $I(X)$. Otherwise, in order to satisfy the conditions imposed on X, one can choose $h_1, \ldots, h_c$ to be c random linear combinations of the minimal generators of $I(X)$.

Proposition 5.2.2. *The Lagrange likelihood equations have the following properties.*

1. *If (p,λ) is a solution of $\operatorname{LL}(X,u)$ and $u_+ \neq 0$, then $\sum p_i = 1$.*
2. *If $p_i = 0$, then $u_i = 0$.*
3. *If the point p is a critical point of $\ell_u(p)$ restricted to $X_{reg} \setminus \mathcal{H}_n$, then there exists an unique λ such that (p,λ) is a solution to $\operatorname{LL}(X,u)$.*
4. *If $p \in X_{reg}$ and (p,λ) is a regular isolated solution to $\operatorname{LL}(X,u)$, then p is a critical point of $\ell_u(p)$ on $X_{reg} \setminus \mathcal{H}_n$.*

5. *For generic choices of u, the number of solutions of* $\mathrm{LL}(X, u)$ *with* $p \in X_{reg} \backslash \mathcal{H}_n$ *equals the MLdegree of* X.

Proof. To arrive at property (1), we sum the equations of (5.2) to get

$$\sum_{i=0}^{n}(u_+p_i - u_i - p_i(\lambda_1\partial_i h_1 + \cdots + \lambda_n \partial_i h_c)) = \sum_{i=0}^{n} p_i u_+ - u_+ = u_+(\sum_{i=0}^{n} p_i - 1).$$

The first equality above follows by Euler's relation of homogeneous polynomials.

The implication stated in property (2) is clearly seen by setting p_i equal to zero in the ith equation of Equations (5.2).

For properties (3) and (4), we note that, as discussed in Chapter 4, $p \in X_{reg} \backslash \mathcal{H}_n$ is a critical point of $\ell_u(p)$ on X if and only if the linear subspace $T_p^{\perp}$ contains the point

$$\left(\frac{u_0}{p_0} - \frac{u_+}{p_+} : \ldots : \frac{u_n}{p_n} - \frac{u_+}{p_+}\right).$$

When X is of codimension c, this is equivalent to saying that $p \in X_{reg} \backslash \mathcal{H}_n$ is a critical point for $\ell_u(p)$ on X if and only if there exist $\lambda_1, \ldots, \lambda_c \in \mathbb{C}$ such that for all $0 \leq i \leq n$,

$$\frac{u_i}{p_i} - \frac{u_+}{p_+} = \lambda_1 \cdot \partial_i h_1 + \ldots + \lambda_c \cdot \partial_i h_c.$$

The Langrange likelihood equations are a restatement of this condition with the denominators cleared. Property (5) follows from (3) and (4). □

If we homogenenize the Lagrange likelihood equations using p_+ and u_+ so that each equation is homogeneous in both the coordinates $p_0, \ldots, p_n$ and the coordinates $u_0, \ldots, u_n$ $\lambda_1, \ldots, \lambda_c$, the Lagrange likelihood equations define a variety $\hat{\mathcal{L}}_X$ in the product space $\mathbb{P}^n_p \times \mathbb{P}^{n+c}_{u,\lambda}$. The variety $\hat{\mathcal{L}}_X$ is related to the likelihood correspondence as follows. Let

$$\pi : \mathbb{P}^n \times \mathbb{P}^{n+c} \to \mathbb{P}^n \times \mathbb{P}^n$$
$$((p_0 : \ldots : p_n), (u_0 : \ldots : u_n : \lambda_1 : \ldots \lambda_n)) \mapsto ((p_0 : \ldots : p_n), (u_0 : \ldots : u_n)).$$

Then by Proposition 5.2.2, the morphism π maps a dense open set of $\hat{\mathcal{L}}_X$ to a dense open set of $\mathcal{L}(X)$, thus,

$$\mathcal{L}(X) = \pi(\hat{\mathcal{L}}_X).$$

The implication of this equality is that by studying the Lagrange likelihood equations, we are in fact studying fibers of the projection $pr_2 : \mathbb{P}^n_p \times \mathbb{P}^n_u \to \mathbb{P}^n_u$.

We conclude this section with an example that shows how the Lagrange likelihood equations are used to find critical points of $\ell_u(p)$.

Example 5.2.3. Let $X = \mathrm{Gr}_{2,6} \subset \mathbb{P}^{14}$ be the variety defined by

$$p_{ij}p_{kl} - p_{ik}p_{jl} + p_{ij}p_{jk}, \quad 1 \leq i < j < k < l \leq 6.$$

The Grassmannian $\mathrm{Gr}_{2,6}$ parameterizes lines in the projective space $\mathbb{P}^5$. It has codimension 6 and is not a complete intersection. However, the 6 equations

$$\begin{aligned} h_1 &= p_{36}p_{45} - p_{35}p_{46} + p_{34}p_{56}, \quad h_4 = p_{26}p_{45} - p_{25}p_{46} + p_{24}p_{56}, \\ h_2 &= p_{25}p_{34} - p_{24}p_{35} + p_{23}p_{45}, \quad h_5 = p_{16}p_{45} - p_{15}p_{46} + p_{14}p_{56}, \\ h_3 &= p_{15}p_{34} - p_{14}p_{35} + p_{13}p_{45}, \quad h_6 = p_{14}p_{23} - p_{13}p_{24} + p_{12}p_{34} \end{aligned}$$

define a reducible variety that has $\mathrm{Gr}_{2,6}$ as an irreducible component (the other components live in the coordinate hyperplanes). The system of equations $\mathrm{LL}\,(X, u)$ consists of 21 equations: the equations $h_1 = \cdots = h_6 = 0$ and the 15 below

$$\begin{aligned} u_{12} - u_+p_{12} &= p_{12}(\lambda_1 \cdot \tfrac{\partial h_1}{\partial p_{12}} + \ldots + \lambda_6 \cdot \tfrac{\partial h_6}{\partial p_{12}}) \\ u_{13} - u_+p_{13} &= p_{13}(\lambda_1 \cdot \tfrac{\partial h_1}{\partial p_{13}} + \ldots + \lambda_6 \cdot \tfrac{\partial h_6}{\partial p_{13}}) \\ &\vdots \\ u_{56} - u_+p_{56} &= p_{56}(\lambda_1 \cdot \tfrac{\partial h_1}{\partial p_{56}} + \ldots + \lambda_6 \cdot \tfrac{\partial h_6}{\partial p_{56}}). \end{aligned}$$

Solving $\mathrm{LL}\,(X, u)$, we find there are 156 regular isolated solutions (p, λ) with $p \in X$, thus, by Proposition 5.2.2 the ML degree of X is 156.

5.3 Sampling and model zeros

In this section, we determine what happens when the data vector u contains zero entries. By understanding the maximum likelihood estimation problems for sampling and model zeros we gain insight into the ML degree of a variety X.

For a subset $S \subseteq \{0, 1, \ldots, n\}$, we define

$$U_S := \{u \in \mathbb{P}^n \mid u_i = 0 \,\text{if}\, i \in S \text{ and nonzero otherwise}\}.$$

The set U_S specifies which entries of the data vector are zero. A partial order on the set of all $\{U_S \ : S \subseteq \{0, 1, \ldots, n\}\}$ is induced by inclusion and we notice $U_S \subseteq U_{S'}$ if and only if $S' \subseteq S$. For ease of notation, we define $U := U_\emptyset$. When $u \in U_S$, every u_i with $i \in S$ is considered a sampling zero or a model zero.

A sampling zero at cell i changes the likelihood function since the monomial $p_i^{u_i}$ no longer appears in l_u. In the case of a model zero at cell i, the model zero is not considered as part of the data, and thus, the likelihood function is changed as well: $p_i^{u_i}$ no longer appears in the function and p_i is set to zero in p_+. Below, we make precise how the maximum likelihood estimation problem changes in the presence of model zeros and sampling zeros and describe the maximum likelihood estimation problem on X for data $u \in U_S$ with model zeros R.

Let $S \subseteq \{0, 1, \ldots, n\}$ and $R \subseteq S$ and consider the following modified likelihood function

$$\ell_{u,S} := \prod_{i \notin S} p_i^{u_i}/p_+^{u_+}.$$

The set $X_R := X \cap \{p \in \mathbb{P}^n \mid p_i = 0 \text{ for all } i \in R\}$ will be called the *model zero variety* for X and R. We consider X_R as a projective variety in $\mathbb{P}^{n-|R|}$ and define $\mathcal{H}_R$ as the set of points in $\mathbb{P}^{n-|R|}$ where $\left(\prod_{i \notin R} p_i\right) \cdot p_+$ vanishes. The model zero variety X_R is called *proper* if the codimension of $X_R \subset \mathbb{P}^{n-|R|}$ equals the codimension of $X \subset \mathbb{P}^n$.

Definition 5.3.1. The *maximum likelihood estimation problem on X for data $u \in U_S$ with model zeros R*, denoted $ML_{R,S}$, is to determine the critical points of $\ell_{u,S}$ on $X_R \setminus \mathcal{H}_R$. The MLdegree (X_R, S) is defined to be the number of critical points of $\ell_{u,S}$ on $X_R \setminus \mathcal{H}_R$ for generic $u \in U_S$ when X_R is proper and zero otherwise.

In terms of the likelihood correspondence, the MLdegree(X_R, S) is the cardinality of the subset of points (p, u) of $pr_2^{-1}(u)$ such that $p_i = 0$ for all $i \in R$ for generic $u \in U_S$. Whenever $R = S$, then MLdegree(X_R, S) simply equals MLdegree$(X_R \subset \mathbb{P}^{n-|R|})$. In terms of optimization, the MLdegree (X_R, S) gives an upper bound on the local maxima of $l_{u,S} := \prod_{i \notin S} p_i^{u_i}$ on $\mathcal{M} \cap \{p_i = 0 \text{ for all } i \in R\}$.

Next, we take the time to explain the subtleties of sampling zeros, model zeros, and structural zeros. When given a model $\mathcal{M}$ with closure X and structural zeros R, common practice is to optimize $l_{u,R}$ restricted to $\pi_R(X)$, the closure of the *projection* of X onto all coordinates not indexed by R [6][36]. In contrast, given a model $\mathcal{M}$ with closure X and *model* zeros R, the goal is to optimize $l_{u,R}$ restricted to X_R. In general, $\pi_R(X) \neq X_R$, and so, the number of critical points will differ. The reason this occurs is because projections of intersections is not the same as intersecting projections. We illustrate the differences between model zeros, sampling zeros, and structural zeros in the next three examples.

Notation 5.3.2. We use S to denote the indices of the data zeros in u and $R \subset S$ to denote the indices of the model zeros. While we defined $S \subset \{0, 1, \ldots, n\}$, in some examples, it is more natural to index the entries of u by ordered pairs. In this case, S will be a set of ordered pairs indicating the positions of the data zeros and R will be a set of ordered pairs indicating the positions of the model zeros.

Example 5.3.3 (Model, sampling, and structural zeros). Let X denote the set of 3×3 matrices of rank 2 in $\mathbb{P}^8$. The variety X is a hypersurface defined by the polynomial $f = p_{11}p_{22}p_{33} - p_{11}p_{23}p_{32} - p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31}$. The ML degree of X is 10.

When we have data u as a 3×3 table and the upper left entry u_{11} is a model zero, then optimization proceeds over $X_R = X_{\{(1,1)\}}$. The model zero variety X_R is defined by the polynomial $-p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31}$, obtained by setting $p_{11} = 0$ in f. In this case, there are 5 complex critical points, that is, MLdegree$(X_R) = 5$.

When u_{11} is a sampling zero, optimization proceeds over X and critical points on the coordinate hyperplanes are ignored. In this case, there are 5 complex critical points whose coordinates are all non-zero, i.e., MLdegree$(X, \{(1,1)\}) = 5$.

When u_{11} is a structural zero, optimization proceeds over $\pi_R(X) = \mathbb{P}^7$. The projection is onto since X is a hypersurface. In this case, there is one complex critical point.

Example 5.3.4. Let X denote the set of 3×4 matrices of rank 2 in $\mathbb{P}^{11}$. The defining ideal of X is generated by the four 3×3 minors of p,

$$\begin{aligned} I(X) = \langle & p_{11}p_{22}p_{33} - p_{11}p_{23}p_{32} - p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31}, \\ & p_{11}p_{22}p_{34} - p_{11}p_{24}p_{32} - p_{12}p_{21}p_{34} + p_{12}p_{24}p_{31} + p_{14}p_{21}p_{32} - p_{14}p_{22}p_{31}, \\ & p_{11}p_{23}p_{34} - p_{11}p_{24}p_{33} - p_{13}p_{21}p_{34} + p_{13}p_{24}p_{31} + p_{14}p_{21}p_{33} - p_{14}p_{23}p_{31}, \end{aligned}$$

$$p_{12}p_{23}p_{34} - p_{12}p_{24}p_{33} - p_{13}p_{22}p_{34} + p_{13}p_{24}p_{32} + p_{14}p_{22}p_{33} - p_{14}p_{23}p_{32}\rangle.$$

The ML degree of X is 26.

Now let u_{11} be a model zero in the contingency table u. In this case, $R = \{(1,1)\}$ and the defining ideal of X_R is

$$\begin{aligned} I(X_R) = \langle & p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31}, \\ & p_{12}p_{21}p_{34} + p_{12}p_{24}p_{31} + p_{14}p_{21}p_{32} - p_{14}p_{22}p_{31}, \\ & p_{13}p_{21}p_{34} + p_{13}p_{24}p_{31} + p_{14}p_{21}p_{33} - p_{14}p_{23}p_{31}, \\ & p_{12}p_{23}p_{34} - p_{12}p_{24}p_{33} - p_{13}p_{22}p_{34} + p_{13}p_{24}p_{32} + p_{14}p_{22}p_{33} - p_{14}p_{23}p_{32}\rangle. \end{aligned}$$

The MLdegree$(X_R) = 13$.

When u_{11} is a structural zero, we follow [36] and eliminate p_{11} from the ideal $I(X)$ to obtain the defining ideal of $\pi_R(X)$,

$$I(\pi_R(X)) = \langle p_{12}p_{23}p_{34} - p_{12}p_{24}p_{33} - p_{13}p_{22}p_{34} + p_{13}p_{24}p_{32} + p_{14}p_{22}p_{33} - p_{14}p_{23}p_{32}\rangle.$$

Optimizing over $\pi_R(X)$, yields 10 complex critical points. Coincidently, 10 is also the ML degree for 3×3 rank 2 matrices.

Example 5.3.5. Let X be the set of 3×3 matrices of rank 1 in $\mathbb{P}^8$. It is well known that the ML degree of X equals 1 and that the corresponding critical point of $\ell_u(p)$ is $\frac{1}{u_{++}^3}[u_{i+}u_{+j}]$ for generic choices of data. Now consider the case when

$$u = \begin{bmatrix} 0 & u_{12} & u_{13} \\ u_{21} & 0 & u_{23} \\ u_{31} & u_{32} & 0 \end{bmatrix}.$$

The zeros of u are indexed by $S = \{(1,1),(2,2),(3,3)\}$.

If all zeros of u are sampling zeros, then we ask how many critical points of $\ell_{u,\emptyset} = p_{11}^{u_{11}}p_{12}^{u_{12}}\cdots p_{33}^{u_{33}}/p_{++}^{u_{++}}$ restricted to $X \subset \mathbb{P}^8$ there are. We find the unique critical point is again $\frac{1}{u_{++}^3}[u_{i+}u_{+j}]$.

If the zeros of u are model zeros, then we let $R = S$ and we ask how many critical points of $\ell_{u,R} = p_{12}^{u_{12}}p_{13}^{u_{13}}p_{21}^{u_{21}}p_{23}^{u_{23}}p_{31}^{u_{31}}p_{32}^{u_{32}}/p_{++}^{u_{++}}$ restricted to $X \cap \{p_{11} = p_{22} = p_{33} = 0\} \setminus \mathcal{H}_R \subset \mathbb{P}^5$ there are. We find there are no such critical points. This is because $X \cap \{p_{11} = p_{22} = p_{33} = 0\} \subseteq \mathcal{H}_R$.

If the zeros of u are structural zeros, then the model under consideration is a quasi-independence model; such models have been well-studied. The projection $\pi_R(X)$ is defined by one equation $p_{12}p_{23}p_{31} - p_{13}p_{21}p_{32}$, and we find the ML degree of $\pi_R(X)$ is 3.

We now come to the description of the special fiber $pr_2^{-1}(u)$ when u is a generic data vector in U_S. This connects the material in this chapter with previous work on the likelihood correspondence [27].

Theorem 5.3.6. *Let u be a generic data vector in U_S for some $S \subseteq \{0,\ldots,n\}$. Let $X \subseteq \mathbb{P}^n$ be a codimension c irreducible component of a projective variety defined by homogeneous*

polynomials $h_1, \dots, h_c$. Let X_R be a proper model zero variety for all $R \subseteq S$. Then, the special fiber $pr_2^{-1}(u)$ contains the critical points of the problem $ML_{R,S}$ for all $R \subseteq S$.

Moreover, if $(p, u) \in pr_2^{-1}(u)$ with $p \in (X_R \setminus \mathcal{H}_R)_{reg}$ then p is a critical point of the problem $ML_{R,S}$ for some $R \subseteq S$.

Proof. Most of the work of this proof comes from the formulation of the Lagrange likelihood equations. First, note that for a variety $Y \subseteq \mathbb{P}^n$ and $u \in U'_S$ for $S' \subseteq \{0, 1, \dots, n\}$, the point $p \in Y_{reg} \setminus \mathcal{H}$ is a critical point on Y for $l_{u,S'}$ if and only if the linear subspace $T_p^{\perp}$ contains the point $v \in \mathbb{P}^{n-|R|}$ where

$$v_i = \begin{cases} \frac{u_i}{p_i} - \frac{u_+}{p_+} & \text{if } i \notin S, \\ -\frac{u_+}{p_+} & \text{if } i \in S. \end{cases}.$$

This condition results in the same equations as in $\mathrm{LL}(Y, u)$ when $u_i = 0$ for all $i \in S$ and p_i is assumed not to be zero when $i \notin S$.

Second, note that when we substitute $p_i = 0$ into $\mathrm{LL}(X, u)$, we get the equations for $\mathrm{LL}(X_R, u)$. Thus, by substituting $p_i = 0$ for $i \in R$ and $u_i = 0$ for $i \in S$ into $\mathrm{LL}(X, u)$, we get a system of equations whose solutions are the critical points of $\ell_{u,S}$ on X_R.

This implies that if X_R is a proper model zero variety then p is a critical point on X_R for $\ell_{u,S}$ if and only if there exists λ such that (p, λ) is an isolated solution to $\mathrm{LL}(X, u)$, or equivalently, the point $(p, u) \in \mathcal{L}_X$.

From Proposition 5.2.2, we know $u_i \neq 0$ implies $p_i \neq 0$, thus, we can account for all solutions to $\mathrm{LL}(X, u)$ since we consider every subset $R \subseteq S$. □

In the proof of Theorem 5.3.6, we also proved the following statement (Proposition 5.3.7). We state Proposition 5.3.7 separately in order to highlight the equations for $ML_{R,S}$.

Proposition 5.3.7. *Fix $u \in U_S$ and $X \subset \mathbb{P}^n$ with codimension c that is an irreducible component of the projective variety defined by homogeneous polynomials $h_1, \dots, h_c$. Whenever X_R is proper, the critical points of $\ell_{u,S}$ restricted to X_R are regular isolated solutions of the equations:*

$$\begin{aligned} h_1 = h_2 = \cdots = h_c = 0 \\ p_i = 0 \text{ for } i \in R, \text{ and} \end{aligned} \tag{5.3}$$

$$\begin{aligned} u_+ &= (\lambda_1 \partial_i h_1 + \lambda_2 \partial_i h_2 + \cdots + \lambda_c \partial_i h_c) && \text{for } i \in S \setminus R \\ (u_+ p_i - p_+ u_i) &= p_i (\lambda_1 \partial_i h_1 + \lambda_2 \partial_i h_2 + \cdots + \lambda_c \partial_i h_c) && \text{for } i \notin S \end{aligned} \tag{5.4}$$

Moreover, the solutions to (5.3) and (5.4) for all $R \subseteq S$ account for all the solutions to $\mathrm{LL}(X, u)$.

An important consequence of Theorem 5.3.6 is that we can use a parameter homotopy to take the solutions of $\mathrm{LL}(X, u)$ for $u \in U$ to the solutions of $\mathrm{LL}(X, v)$ for $v \in U_S$. Such methods are discussed in [41] and can be implemented in `Bertini` [3] or `PHCpack` [46]. Doing so, we solve $2^{|S|}$ different optimization problems corresponding to the $2^{|S|}$ subsets of S. In the case $|\ S\ | = 1$, we get the following corollary.

Corollary 5.3.8 (ML degree bound). *Suppose $S = \{n\}$ and $X \subset \mathbb{P}^n$ is an irreducible projective variety. Then for generic $u \in U_S$, we have*

$$\mathrm{MLdegree}(X) \geq \mathrm{MLdegree}(X_S) + \mathrm{MLdegree}(X, S)$$

Moreover, when X is a generic complete intersection, the inequality becomes an equality.

Proof. This follows from Theorem 5.3.6 and the fact that the number of solutions to a parameterized family of polynomial systems for a generic choice of parameters can only decrease on nested parameter spaces [34]. Equality holds when u remains off an exceptional subset $\mathcal{E} \subset U$ which is defined by an algebraic relation among the p coordinates and u coordinates [41]. Since X is a generic intersection, we have U_S is not strictly contained in $\mathcal{E}$, and the equality holds. □

As we can see from Corollary 5.3.8, solutions to $\mathrm{LL}(X, u)$ with $u \in U_S$ get partitioned into sampling zero and model zero solutions, in fact, we see this same behavior even as we increase the size of S. We encode $\mathrm{MLdegree}\,(X_R, S)$ for all possible choices of (R, S) in a table called the *ML table* of X whose rows are indexed by $R \subset \{0, 1, \ldots, n\}$ and whose columns are indexed by $S \subset \{0, 1, \ldots, n\}$. Due to space considerations, in our examples, we often only print partial ML tables, i.e. that is subtables of the complete ML table.

Example 5.3.9. The ML table of a generic curve of degree d in $\mathbb{P}^2$ is below. The top left entry of the table is the ML degree of a generic curve of degree d in $\mathbb{P}^2$.

$R\backslash S$	$\{\}$	$\{0\}$	$\{1\}$	$\{2\}$
$\{\}$	$d + d^2$	d^2	d^2	d^2
$\{0\}$		d	0	0
$\{1\}$			d	0
$\{2\}$				d

Example 5.3.10. Let $X \subset \mathbb{P}^8$ be the projectivization of all 3×3 matrices of rank 2. A partial ML table of X is below.

$R\backslash S$	$\{\}$	$\{11\}$	$\{12\}$	$\{11, 12\}$
$\{\}$	10	5	5	1
$\{11\}$		5	–	4
$\{12\}$			5	4
$\{11, 12\}$				1

.

In Example 5.3.9 and Example 5.3.10 above, each of the columns of the $\mathrm{MLtable}(X)$ sum to $\mathrm{MLdegree}(X)$. This does not happen for all varieties, but, in general, the column sums are lower bounds of the ML degree of X.

Corollary 5.3.11. *The column sums of the ML table of X are less than or equal to* $\mathrm{MLdegree}(X)$*, meaning* $\mathrm{MLdegree}(X) \geq \sum_{R \subseteq S} \mathrm{MLdegree}(X_R, S)$*. Moreover, when X is a generic complete intersection, the inequality becomes an equality.*

The inequality in Corollary 5.3.11 above can be strict as the next example shows.

Example 5.3.12. Let $f = p_0^3 + p_1^3 + p_2^3 + p_3^3$ define a hypersurface $X \subset \mathbb{P}^3$. Some of the entries of the MLtable of X are below. We have $\mathrm{MLdegree}(X) = 30$ but for $S = \{0, 1\}$, we have $\sum_{R \subseteq S} \mathrm{MLdegree}(X_R, S) = 28$.

$R\backslash S$	$\{\}$	$\{0\}$	$\{0,1\}$
$\{\}$	30	21	12
$\{0\}$		9	7
$\{1\}$			7
$\{0,1\}$			2

Remark 5.3.13. Our definition for the entries of the ML table ignores multiplicities and singularities of the variety. We only take account regular isolated solutions. An interesting research direction would be to take into account multiplicities to obtain an equality in the statement of Corollary 5.3.8.

We conclude this section with a full description of the ML table for a generic hypersurface of degree d in $\mathbb{P}^n$.

Theorem 5.3.14. *Suppose X is a generic hypersurface of degree d in $\mathbb{P}^n$ and let $s =\mid S \mid$ and $r =\mid R \mid$. Then*

$$\text{MLdegree}\,(X_R, S) = \begin{cases} \frac{d}{d-1}\left(d^{n-s}-1\right) & s = r \\ d^{n-s+1}\left(d-1\right)^{s-r-1}, & s > r \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Since the entries of the ML table of generic degree d hypersurfaces $X \subset \mathbb{P}^n$ depend only on d, n, and the size of R and S, we ease notation and let $\text{MLdegree}\,(X_r \subset \mathbb{P}^n, s) := \text{MLdegree}\,(X_R \subset \mathbb{P}^n, S)$. By Proposition 5.3.7, it follows

$$\text{MLdegree}\left(X_r \subset \mathbb{P}^{n+1}, s\right) = \text{MLdegree}\,(X_{r-1} \subset \mathbb{P}^n, s-1) \text{ for } r, s \geq 1 \tag{5.5}$$

because a section of a generic hypersurface projected into a smaller projective space is again generic degree d hypersurface. We will use (5.5) to induct on n.

Recall by [23] the ML degree of a generic degree d hypersurface in $\mathbb{P}^n$ is $\frac{d}{d-1}\left(d^n-1\right)$. So when $s = r$, we have $\text{MLdegree}\,(X_R, S) = \frac{d}{d-1}\left(d^{n-s}-1\right)$ as desired. So for $n = 2$ we have

$$\text{MLdegree}\,(X_\emptyset, \emptyset) = \text{MLdegree}\,(X_\emptyset, \{0\}) + \text{MLdegree}\left(X_{\{0\}}, \{0\}\right).$$

Simple algebra reveals $\text{MLdegree}\,(X_\emptyset \subset \mathbb{P}^2, \{0\}) = d^2$. With this we have shown the theorem holds when $n = 2$. To complete the proof by induction, we need only show $\text{MLdegree}\,(X_r \subset \mathbb{P}^{n+1}, s)$ equals $d^{n-s+1}\left(d-1\right)^{s-r-1}$, when $r = 0$ and $r < s$. To show this we recall

$$\text{MLdegree}(X \subset \mathbb{P}^{n+1}) = \sum_{R \subset S} \text{MLdegree}\left(X_R \subset \mathbb{P}^{n+1}, S\right). \tag{5.6}$$

The right hand side of (5.6) becomes

$$\text{MLdegree}\left(X_\emptyset \subset \mathbb{P}^{n+1}, s\right) + \sum_{r=1}^{s} \binom{s}{r} \text{MLdegree}\,(X_{r-1} \subset \mathbb{P}^n, s-1).$$

Letting $D = \text{MLdegree}\,(X_\emptyset \subset \mathbb{P}^{n+1}, s)$ we have that (5.6) simplifies to

$$\frac{d}{d-1}(d^{n+1}-1) = D + \sum_{r=1}^{s-1} \binom{s}{r} d^{n-s+2}(d-1)^{s-r-1} + \frac{d}{d-1}(d^{n-s+1}-1).$$

With the binomial formula it follows $D = d^{n-s+2}\,(d-1)^{s-1}$ finishing the proof. □

Example 5.3.15. By Theorem 5.3.14 we have the following MLtable of a generic degree d hypersurface $X \subset \mathbb{P}^n$.

$R\backslash S$	$\{\}$	$\{0\}$	$\{0,1\}$	$\{0,1,2\}$	$\cdots$
$\{\}$	$\frac{d}{d-1}\,(d^n-1)$	$d^n\,(d-1)^0$	$d^{n-1}\,(d-1)$	$d^{n-2}\,(d-1)^2$	
$\{0\}$		$\frac{d}{d-1}\,(d^{n-1}-1)$	$d^{n-1}\,(d-1)^0$	$d^{n-2}\,(d-1)^1$	
$\{0,1\}$			$\frac{d}{d-1}\,(d^{n-2}-1)$	$d^{n-2}\,(d-1)^0$	
$\{0,1,2\}$				$\frac{d}{d-1}\,(d^{n-3}-1)$	
$\vdots$					$\ddots$

5.4 Applications and further directions

In this section we illustrate the computational gains acquired by working with model zero varieties. This section has four brief subsections focused on different applications: ML table homotopies, ML duality, tensors (multi-way tables), and Grassmannians.

ML table homotopy

Let $X \subset \mathbb{P}^n$ be a generic complete intersection of codimension c defined by homogeneous polynomials $h_1, \ldots, h_c$. Let u be generic data vector in U, and let u_s be a generic data vector in U_S with $S \subseteq \{0, 1, \ldots, n\}$. Our first application of Corollary 5.3.8 is the construction of a homotopy to determine critical points of $\ell_u(p)$ on X. We determine the critical points of $\ell_{u_s,S}$ on $X \cap \mathcal{H}_R$ for each subset R of S. So rather than doing a single expensive computation to determine the critical points of $\ell_u(p)$ on X, we perform several easier computations to determine critical points of $\ell_{u_s,S}$. Doing so allows us to use Proposition 5.3.7 to get the critical points of $\ell_u(p)$ using a coefficient-parameter homotopy. The homotopy requires two steps. Step 1 determines the start points by solving multiple systems of equations. Step 2 constructs the coefficient-parameter homotopy (see [41, §7]) that will do the path tracking.

Example 5.4.1. Let $X \subset \mathbb{P}^3$ be defined by $f = 2p_0^3 - 3p_1^3 + 5p_2^3 - 7p_3^3$. We note that MLdegree$(X) = 39$ and the ML table of X is:

$R\backslash S$	$\{\}$	$\{0\}$	$\{1\}$	$\{0,1\}$
$\{\}$	39	27	27	18
$\{0\}$		12	–	9
$\{1\}$			12	9
$\{0,1\}$				3

Let $S = \{0, 1\}$ and let u_s be a generic vector in U_S. For Step 1 of the algorithm, we solve four systems of equations. Each system of equations corresponds to a choice of R from $\mathcal{R} := \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$. For example, when $R = \{0, 1\}$, we solve the following system

$$\begin{aligned} f &= 0 \\ p_0 &= 0 \\ p_1 &= 0 \\ (u_+ p_2 - p_+ u_2) &= p_2 \lambda_1 \cdot \partial_2 f \\ (u_+ p_3 - p_+ u_3) &= p_3 \lambda_1 \cdot \partial_3 f \end{aligned}$$

and find 3 solutions. In general, we solve the equations in Proposition 5.3.7. So when $R = \emptyset, \{0\}, \{1\}, \{0, 1\}$ we determine there are $18, 9, 9, 3$ solutions for the respective systems for a total of 39 solutions. For Step 2, by Proposition 5.3.7, the computed 39 solutions are solutions to the Lagrange likelihood equations $\mathrm{LL}(X, u_s)$. By using the coefficient-parameter homotopy $\mathrm{LL}(X, u_s \to u)$, we can go from data with zeros u_s to generic data u.

Algorithm 5.4.2.

- Input: $u_s \in U_S$ and homogeneous polynomials $h_1, h_2, \ldots, h_c$ defining X with codimension c.
- (Step 1) Solve $\mathrm{LL}(X_R, u_s)$ for each $R \subset S$ to determine the start points of the homotopy.
- (Step 2) Construct and solve the coefficient-parameter homotopy $\mathrm{LL}(X, u_s \to u)$.
- Output solutions to $\mathrm{LL}(X, u)$ yielding the critical points of $\ell_u(p)$ on X.

The immediate advantage of this homotopy is that we can get several critical points of $\ell_u(p)$ quickly. Thus, we get some insight if the ML degree of X is small. Moreover, one can use monodromy methods [39] to attempt to recover additional solutions. One drawback is that by increasing the size of S we also increase the number of subproblems we need to solve, a second drawback is that we may not know *a priori* that $\sum_{R \subseteq S} \mathrm{MLdegree}(X_R, S)$ equals the ML degree. To address the first drawback, one can take advantage of the structure of the problem to lessen the number of subproblems. For example, in the case when X is a generic hypersurface, we know that the ML degree of X depends only on the size of R and S. Taking advantage of this structure and pairing change of variables with parameter homotopies, we preprocess much fewer subproblems—namely $|S|$ subproblems versus $2^{|S|}$. While we do not have equality in Corollary 5.3.8 in general, equality does occur in some examples (see Theorem 5.3.14).

Maximum Likelihood Duality

In this section, we extend ML duality for matrix models when u contains zero entries. We let $X \subset \mathbb{P}^{mn-1}$ be the variety of $m \times n$ matrices of rank less than or equal to r and we let $Y \subset \mathbb{P}^{mn-1}$ be the variety of $m \times n$ matrices of rank less than or equal to $m - r + 1$ where $m \leq n$. In Chapter 3, it is shown that $\mathrm{MLdegree}\, X = \mathrm{MLdegree}\, Y$ by considering critical points of l_u on subvarieties of the algebraic torus; a bijection between said critical points is

also given. Translating these results into the language of determining critical points of $\ell_u(p)$ on subvarieties of projective space, we are able to talk about sampling zeros and model zeros. As a consequence, we can use maximum likelihood duality to gain computational advantages by exploiting that sampling zeros are dual to model zeros.

Proposition 5.4.3. *Let X and Y be defined as above so that they are ML dual varieties. Let $S \subset [n]$ and $u \in U_S$. If $P \in \mathbb{C}^{mn}$ is a solution to $\mathrm{LL}(X, u)$, then there exists a $Q \in \mathbb{C}^{mn}$ such that Q is a solution to $\mathrm{LL}(Y, u)$ and*

$$P \star Q = \Omega_U \tag{5.7}$$

$$\text{where } \Omega_U = \begin{bmatrix} u \\ \hline u_{++} \end{bmatrix} \star \begin{bmatrix} u_{i+}u_{+j} \\ \hline u_{++}^2 \end{bmatrix} \tag{5.8}$$

Proof. Let $\mathcal{D} \subset \mathbb{P}^{nm-1} \times \mathbb{P}^{nm-1} \times \mathbb{P}^{nm-1}$ be the set of all points (p, q, u) such that $(p, u) \in \mathcal{L}_X$, $(q, u) \in \mathcal{L}_Y$ and

$$u_{++}^3 p_{ij} q_{ij} - p_{++} q_{++} u_{i+} u_{ij} u_{+j} = 0 \text{ for } 0 \leq i \leq m,\ 0 \leq i \leq n.$$

The set $\mathcal{D}$ is a projective variety, thus, if we consider the projection

$$\begin{aligned} \phi : \mathbb{P}^n \times \mathbb{P}^n \times \mathbb{P}^n &\to \mathbb{P}^n \times \mathbb{P}^n \\ (p, q, u) &\mapsto (p, u), \end{aligned}$$

the image of $\mathcal{D}$ under ϕ is a variety. By 3.2.4, we know that a dense open subset of $\mathcal{L}_X$ is contained in $\phi(\mathcal{D})$, therefore, $\mathcal{L}_X \subseteq \phi(\mathcal{D})$ and the statement of the theorem follows. □

Theorem 5.4.4. *Let X and Y be defined as in Lemma 5.4.3. Fix $S \subset [m] \times [n]$ and generic $u \in U_S$. Then a solution to the maximum likelihood estimation problem $ML_{R,S}(u)$ is dual to a solution to the maximum likelihood estimation problem $ML_{R',S}(u)$, with $(S \setminus R) \subset R'$.*

When $\mid S \mid = 1$, the theorem says that a sampling zero critical point is dual to a model zero critical point. We also believe that the converse, model zero critical points are dual to sampling zero critical points is true, and that in general, $(S \setminus R) \subset R'$ is actually an equality in the theorem. Nonetheless, because computing model zeros is heuristically easier than computing sampling zeros, we make computational gains with Theorem 5.4.4.

In Example 5.3.10, we see that a column of the ML table is symmetric. This is because the variety of 3×3 matrices of rank 2 is ML self dual. Other examples of varieties that are ML self dual include $m \times n$ matrices of rank $\frac{m+1}{2}$ with m being odd. We conclude this subsection with an ML table of 4×4 matrices of rank 2 and 3 respectively. An ongoing project is to give a recursive formula for the ML table of $m \times n$ matrices of rank r similar to what was done in Theorem 5.3.14.

Example 5.4.5. ML table of 4×4 rank 2 matrices:

$R\backslash S$	$\{\}$	$\{11\}$	$\{11,44\}$	$\{11,22,44\}$	$\{11,22,33,44\}$
$\{\}$	191	118	76	51	35
$\{11\}$		73	42	25	16
$\{22\}$			–	25	16
$\{33\}$			–	–	16
$\{44\}$			42	25	16
$\{11,22\}$			–	17	9
$\{11,33\}$			–	–	9
$\{11,44\}$			31	17	9
$\{22,33\}$				–	9
$\{22,44\}$				17	9
$\{33,44\}$				–	9
$\{11,22,33\}$				–	8
$\{11,22,44\}$				14	8
$\{11,33,44\}$					8
$\{22,33,44\}$					8
$\{11,22,33,44\}$					6

ML table of 4×4 rank 3 matrices:

$R\backslash S$	$\{\}$	$\{11\}$	$\{11,44\}$	$\{11,22,44\}$	$\{11,22,33,44\}$
$\{\}$	191	73	31	14	6
$\{11\}$		118	42	17	8
$\{22\}$			–	17	8
$\{33\}$			–	–	8
$\{44\}$			42	17	8
$\{11,22\}$			–	25	9
$\{11,33\}$			–	–	9
$\{11,44\}$			76	25	9
$\{22,33\}$				–	9
$\{22,44\}$				25	9
$\{33,44\}$				–	9
$\{11,22,33\}$				–	16
$\{11,22,44\}$				51	16
$\{11,33,44\}$					16
$\{22,33,44\}$					16
$\{11,22,33,44\}$					35

Tensors

Let T be the set of $2 \times 2 \times 2 \times 2$ tensors with border rank ≤ 2. The ML degree of this variety is unknown. The variety is defined by the 3×3 minors of all possible flattenings. This is an overdetermined system of equations with codimension 6. We choose 6 of the equations to be $h_1, \ldots, h_6$ for the Lagrange likelihood equations. For the model zero variety

with $p_{1111} = p_{2222} = 0$ we find 3 solutions for a generic $u \in U_S$ with $S = \{1111, 2222\}$. When we solve the Lagrange likelihood equations for $R = \{1111\}$, we find 52 solutions with $p \in X$.

Theorem 5.4.6. *Let T be the set of $2 \times 2 \times 2 \times 2$ tensors with border rank ≤ 2.*

$$\text{MLdegree}(T) \geq 52$$

In this example, we also see that when we have data with zeros the number of critical points can drop significantly as we introduce more model zeros.

Grassmannians

Let the ideal $I_{2,n}$ be generated by the quadrics

$$p_{ij}p_{kl} - p_{ik}p_{jl} + p_{il}p_{jk}, \quad 1 \leq i < j < k < l \leq n.$$

Then the variety of $I_{2,n}$ is the Grassmannian $\text{Gr}_{2,n} \subset \mathbb{P}^{\binom{n}{2}-1}$. The Grassmannian $\text{Gr}_{2,n}$ parameterizes lines in the projective space $\mathbb{P}^{n-1}$. Below we have a table of computations. The top line of numbers are ML degrees of Grassmannians while the next line are ML degrees of a model zero variety for Grassmannians. The bottom line has ML degrees of sampling zeros.

	$\text{Gr}_{2,4}$	$\text{Gr}_{2,5}$	$\text{Gr}_{2,6}$
MLdegree X	4	22	156
MLdegree $\left(X_{\{12\}}, \{12\}\right)$	1	4	22
MLdegree $\left(X_{\emptyset}, \{12\}\right)$	3	18	134

These computations were performed by choosing $c = \text{codim}\, X$ generators of $I_{2,n}$ to be $h_1 \dots h_c$ for $\text{LL}(X, u)$. We used the numerical software `bertini` and symbolic packages available in `Macauay2` [16]. From this data we make the following conjecture to motivate the pursuit of a recursive formula for ML degrees of Grassmannians.

Conjecture 5.4.7. *For $n \geq 4$ we conjecture*

$$\text{MLdegree}\, \text{Gr}_{2,n} = \text{MLdegree}(\text{Gr}_{2,n+1} \cap \{p_{12} = 0\}).$$

5.5 Conclusion

Understanding model and sampling zeros gives us insights into the maximum likelihood degree for a given model. When the data vector contains a zero entry, we see that critical points to the likelihood function partition into two groups: critical points for the sampling zero problem and critical points for the model zero problem. This split can help us obtain bounds for the ML degree and provides interesting directions for further research within the study of likelihood geometry, for example, determining which varieties yield an equality in Corollary 5.3.8. Furthermore, model zeros can help with the computational problem of finding all the solutions to a set of likelihood equations. This chapter illustrates some of the advantages of working with model zeros, as seen by the lower bound obtained on the set of $2 \times 2 \times 2 \times 2$ tensors of border rank ≤ 2.

Chapter 6

Bertini for Macaulay2

The content of this chapter has been submitted to the *Journal of Software for Algebra and Geometry* as an article of the same title with minor changes throughout for consistency with other chapters. It is joint work with Daniel J. Bates, Elizabeth Gross, and Anton Leykin.

6.1 Numerical algebraic geometry

Numerical algebraic geometry (numerical AG) refers to a set of methods for finding and manipulating the solution sets of systems of polynomial equations. Said differently, given $f : \mathbb{C}^N \to \mathbb{C}^n$, numerical algebraic geometry provides facilities for computing numerical approximations to isolated solutions of $V(f) = \{z \in \mathbb{C}^N | f(z) = 0\}$, as well as numerical approximations to generic points on positive-dimensional components. The book [41] provides a good introduction to the field, while the newer book [3] provides a simpler introduction as well as a complete manual for the software package `Bertini` [5].

`Bertini` is a free, open source software package for computations in numerical algebraic geometry. The purpose of this chapter is to present a `Macaulay2` package *Bertini* that provides an interface to `Bertini`. This package uses basic datatypes and service routines for computations in numerical AG provided by the package *NAGtypes*. It also fits the framework of *NumericalAlgebraicGeometry* package [30], a native `Macaulay2` implementation of a collection of numerical AG algorithms: most of the core functions of *NumericalAlgebraicGeometry* have an option of using `Bertini` instead of the native solver.

In the remainder of this section, we very briefly describe a few fundamental concepts of the field. In the subsequent sections, we describe the various run modes of `Bertini` that have been implemented in this interface. We conclude with Section 5, which describes how to use *Bertini* within *NumericalAlgebraicGeometry*.

Finding isolated solutions

The core computational engine within `Bertini` is *homotopy continuation*. This is a three-stage process for finding a superset of all isolated solutions in $V(f)$. Given a polynomial system $f(z)$, the three steps are as follows:

1. Choose an easily-solved polynomial system $g(z)$ that reflects the structure of $f(z)$, and solve it. Call this set of solutions S.

2. Form the homotopy
$$H(z,t) = (1-t)f(z) + \gamma t g(z),$$
with $\gamma \in \mathbb{C}$ a random complex number. Notice that $H(z,1) = \gamma g(z)$, the solutions of which are known, and $H(z,0) = f(z)$, for which we seek the solutions.

3. There is a real curve extending from each solution $z \in S$. Use predictor-corrector methods, adaptive precision, and endgames to track along all of these paths as t goes from 1 to 0.

Assuming $g(z)$ is constructed in one of several canonical ways [41], there is a probability one guarantee that this procedure will result in a superset of all isolated solutions of $f(z) = 0$.

There are many variations of this general technique, and there are many minor issues to consider when implementing this method. However, due to space limitations, we leave the reader to explore the references for more information on this powerful method.

Finding irreducible components

Given an irreducible algebraic set X of dimension k, it is well known that X will intersect almost any linear space of codimension k in a finite set of points. In fact, there is a Zariski open subset of the set of all linear spaces of codimension k for which intersection with X yields some fixed number of points, called the *degree* of X, $\deg X$.

This fundamental fact underlies the computation of positive-dimensional irreducible components in numerical algebraic geometry. Suppose algebraic set Z decomposes into irreducible components $Z_{i,j}$,
$$Z = \cup_{i=0}^{\dim Z} \cup_{j \in \Lambda_i} Z_{i,j},$$
where i is the dimension of $Z_{i,j}$ and j is just the index of component $Z_{i,j}$ in dimension i, stored in finite indexing set Λ_i.

In numerical algebraic geometry, the representation W of an algebraic set Z consists of representations $W_{i,j}$ for each irreducible component $Z_{i,j}$ of Z. In particular, *witness set* $W_{i,j}$ is a triple $(f, L_{i,j}, \widehat{W}_{i,j})$, consisting of polynomial system f, linear functions $L_{i,j}$ corresponding to a linear space of codimension i, and *witness point set* $\widehat{W}_{i,j} = Z_{i,j} \cap V(L_{i,j})$.

There are a variety of ways to compute W, many of which are described in detail in [3]. Most of these methods can be accessed through the package *Bertini* by using optional inputs to specify the desired algorithm.

6.2 Solving zero-dimensional systems

In the following sections we outline and give examples of the different `Bertini` run modes implemented in the interface package *Bertini*.

Finding solutions to zero-dimenstional systems

The method `bertiniZeroDimSolve` calls `Bertini` to solve a polynomial system and returns solutions as a list of `Points` using the data types from *NAGtypes*. Diagnostic information, such as the residuals and the condition number, are stored with the coordinates of the solution and can be viewed using `peek`.

```
i1 : R=CC[x,y];
i2 : f = {x^2+y^2-1,(x-1)^2+y^2-1};
i3 : solutions=bertiniZeroDimSolve(f)
o3 = {{.5, .866025}, {.5, -.866025}}
i4 : peek solutions_0
o4 = Point{ConditionNumber => 88.2015     }
           Coordinates => {.5, .866025}
           CycleNumber => 1
           FunctionResidual => 3.66205e-15
           LastT => .000390625
           MaximumPrecision => 52
           NewtonResidual => 4.27908e-15
           SolutionNumber => 3
```

Users can specify to use regeneration, an equation-by-equation solving method, by setting the option `USEREGENERATION` to 1.

```
i5 : solutions=bertiniZeroDimSolve(f, USEREGENERATION=>1);
```

In common applications, one would like to classify solutions, e.g. separate real solutions from non-real solutions, and, thus, recomputing solutions to a higher accuracy becomes important. The method `bertiniRefineSols` calls the sharpening module of `Bertini` and sharpens a list of solutions to a desired number of digits using Newton's method.

```
i6 : refinedSols=bertiniRefineSols(f, solutions, 20);
i6 : (coordinates refinedSols_0)_1
o6 = .86602540378443859659+3.57969487611345073 51e-83*ii
```

Parameter homotopies

Many fields, such as statistics, physics, chemical biology, and engineering contain applications that require solving a large number of systems from a parameterized family of polynomial systems. In such situations, computational time can be decreased by using parameter homotopies. For an example illustrating how parameter homotopies can be used in statistics see [20].

The method `bertiniParameterHomotopy` calls `Bertini` to run both stages of a parameter homotopy. First, `Bertini` assigns a random complex number to each specified parameter and solves the resulting system, then, after this initial phase, `Bertini` computes solutions for every given choice of parameters using a number of paths equal to the exact root count.

```
i7: i68 : R=CC[a,b,c][x,y];
i8 : f={a*x^2+b*y^2-c, y};
i9 : bertiniParameterHomotopy(f,{a,b,c},{{1,1,1},{2,3,4}})
o9 = {{{-1, 0}, {1, 0}}, {{-1.41421, 0}, {1.41421, 0}}}
```

User-defined homotopies

A user may define their own homotopy to solve a square system of polynomial equations. If the homotopy H consists of n polynomials in n unknowns and the path variable t, then the method `bertiniTrackHomotopy` calls `Bertini` to compute solutions to H when $t = 0$. But to do so, the user must also input start points of the homotopy, which are solutions to the system H when $t = 1$.

```
i10 : R=CC_200[x,y,t]
i11 : H = { (x^2-y^2)*t +(2*x^2-3*x*y-5*y^2)*(1-t),(y-1)*t+(x+2*y-3)*(1-t)}
i12 : sol1=    point{{1,1}}, sol2=    point{{ -1,1}}
i13 : S0={sol1,sol2}
i14 : S1=bertiniTrackHomotopy(  H,t,S0)
o14 : {{1.66667, .666667}, {-3, 3}}
```

6.3 Solving positive-dimensional systems

Given a positive-dimensional system f, the method `bertiniPosDimSolve` calls `Bertini` to compute a numerical irreducible decomposition, this decomposition is assigned the type `NumericalVariety` in `Macaulay2`. In the default settings, `Bertini` uses a classical cascade homotopy to find witness supersets in each dimension, removes extra points using a membership test or local dimension test, deflates singular witness points, then factors using a combination of monodromy and a linear trace test.

```
i10 :    R = CC[x,y,z];
i11 :    f = {(y^2+x^2+z^2-1)*x, (y^2+x^2+z^2-1)*y};
i12 :    NV = bertiniPosDimSolve f

o12 = A variety of dimension 2 with components in
      dim 1:  [dim=1,deg=1]
      dim 2:  [dim=2,deg=2].

o12 : NumericalVariety
```

Once the solution set to a system, i.e. the variety V, is computed and stored as a `NumericalVariety`,
`bertiniComponentMemberTest` can be used to test numerically whether a set of points p lie on the variety V. For every point in p, `bertiniComponentMemberTest` returns the components to which that point belongs. As for sampling, `bertiniSample` will sample from a

witness set W. These methods call the membership testing and sampling options in `Bertini` respectively.

```
i13 : p={{0,0,0}};
i14 : bertiniComponentMemberTest (NV, p)
o14 = {{[dim=1,deg=1]}}

i15 : component=NV#1_0
i16 : bertiniSample(component,1)
o16 = {{0, -8.49385e-20+7.48874e-20*ii, -.148227-.269849*ii}}
```

6.4 Solving homogeneous systems

The package *Bertini* includes functionality to solve a homogenous system that defines a projective variety. In `Bertini`, the numerical computations are performed on a generic affine chart to compute representatives of projective points. To solve homogeneous equations, set the option `ISPROJECTIVE` to 1. If the user inputs a square system of n homogeneous equations in $n+1$ unknowns, then the method `bertiniZeroDimSolve` outputs a list of projective points.

```
i35 : R = CC[x,y,z];
i36 : f = {y^2-4*z^2,16*x^2-y^2};
i37 : bertiniZeroDimSolve(f,ISPROJECTIVE=>1);
o37 = {{.251411+.456072*ii, 1.00564+1.82429*ii, .502821+.912143*ii},
 {.106019+.160896*ii, .424078+.643585*ii, -.212039-.321792*ii},
 {-.15916-.12286*ii, .636639+.49144*ii, -.318319-.24572*ii},
 {-.48005-.092532*ii, 1.9202+.370128*ii, .960101+.185064*ii}}
```

If f is a positive-dimensional homogeneous system of equations, then the method `bertiniPosDimSolve` calls `Bertini` to compute a numerical irreducible decomposition of the projective variety defined by f.

```
i48 : R = CC[x,y,z];
i49 : f = {(x^2+y^2-z^2)*(z-x),(x^2+y^2-z^2)*(z+y)};
i50 : NV = bertiniPosDimSolve(f,ISPROJECTIVE=>1)
o50 : = A projective variety with components in projective dimension:
      dim 0:  [dim=0,deg=1]
      dim 1:  [dim=1,deg=2]
```

6.5 Algebraic Statistics Example

In this section, we use the interface to solve the Lagrange likelihood equations (Proposition 5.2.2) for a statistical model defined by f.

```
i48 : R=CC[p0,p1,p2,p12,L];
i49 : f=2*p0*p1*p2+p1^2*p2+p1*p2^2-p0^2*p12+p1*p2*p12;
i50 : pList={p0,p1,p2,p12};
i51 : uList={75,1,5,7};
i52 : uList=1_RR/sum uList*uList;
i53 : gradLF=matrix{for i to #pList-1 list uList_i-sum(uList)*pList_i};
i54 : jacF=matrix{for i in pList list L*i*diff(i,f)};
i55 : likelihoodEquations=ideal(gradLF+jacF)+ideal f;
i56 : likelihoodEquations=flatten entries gens likelihoodEquations;
i57 : criticalPoints=bertiniZeroDimSolve(likelihoodEquations,
                                    MPTYPE=>2,USEREGENERATION=>1)
o57 : = A list of critical points:
  {{.788947, .0702306,  .117171, .023651, 3.84771},
  {-.92006, .722138, 1.12145, .0764765, 1.09425},
  {-1.16434, 18.58, 13.1081, -29.5238,  .00414006}}
```

6.6 Using *Bertini* from *NumericalAlgebraicGeometry*

The *Bertini* package depends on the *NAGtypes* package, a collection of basic datatypes and service routines common to all `Macaulay2` packages for numerical AG: e.g., an interface package to another polynomial homotopy continuation solver, *PHCpack* [18], also has this dependence.

While independent from the *NumericalAlgebraicGeometry* package, our interface provides a valuable option for this package: the user can set `Bertini` as a default solver for homotopy continuation tasks.

```
i1 : needsPackage "NumericalAlgebraicGeometry";


i2 : setDefault(Software=>BERTINI)
```

An alternative way is to specify the `Software` option in a particular command:

```
i3 : CC[x,y]; system = {x^2+y^2-1,2*x+3*y+5};
i4 : sols = solveSystem(system, Software=>M2engine)
o4 =  {{-.769231-.799408*ii, -1.15385+.532939*ii}, {-.769231+.799408*ii, ...
i5 : refsols = refine(system, sols, Bits=>99, Software=>BERTINI);
i6 : first coordinates first refsols
o6 = -.769230769230769273470116331737+.799408065031789516474702850246*ii
o6 : CC (of precision 100)
```

The unified framework for various implementations of numerical AG algorithms should be particularly convenient to a `Macaulay2` user doing numerical computations with tools from many packages.

Bibliography

[1] D. J. Bates, J. D. Hauenstein, T. M. McCoy, C. Peterson, and A. J. Sommese. Recovering exact results from inexact numerical data in algebraic geometry. *Exp. Math.*, 22(1):38–50, 2013.

[2] D. J. Bates, J. D. Hauenstein, C. Peterson, and A. J. Sommese. Numerical decomposition of the rank-deficiency set of a matrix of multivariate polynomials. In *Approximate commutative algebra*, Texts Monogr. Symbol. Comput., pages 55–77. SpringerWienNewYork, Vienna, 2009.

[3] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler. *Numerically solving polynomial systems with Bertini*, volume 25 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.

[4] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler, II. Adaptive multiprecision path tracking. *SIAM J. Numer. Anal.*, 46(2):722–746, 2008.

[5] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wimpier. Bertini: Software for numerical algebraic geometry. Available at https://bertini.nd.edu/.

[6] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. The M.I.T. Press, Cambridge, Mass.-London, 1977. With the collaboration of Richard J. Light and Frederick Mosteller, Third printing.

[7] M.-L. G. Buot, S. Hoşten, and D. S. P. Richards. Counting and locating the solutions of polynomial systems of maximum likelihood equations. II. The Behrens-Fisher problem. *Statist. Sinica*, 17(4):1343–1354, 2007.

[8] M.-L. G. Buot and D. S. P. Richards. Counting and locating the solutions of polynomial systems of maximum likelihood equations. I. *J. Symbolic Comput.*, 41(2):234–244, 2006.

[9] F. Catanese, S. Hoşten, A. Khetan, and B. Sturmfels. The maximum likelihood degree. *Amer. J. Math.*, 128(3):671–697, 2006.

[10] A. Distler. Radiroot: roots of a polynomial as radicals – a GAP package. Available at http://www.icm.tu-bs.de/ag_algebra/software/radiroot/.

[11] J. Draisma and J. Rodriguez. Maximum likelihood duality for determinantal varieties. *International Mathematics Research Notices*, 2013.

[12] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel, 2009.

[13] S. E. Fienberg, P. Hersh, A. Rinaldo, and Y. Zhou. Maximum likelihood estimation in latent class models for contingency table data. In *Algebraic and geometric methods in statistics*, pages 27–62. Cambridge Univ. Press, Cambridge, 2010.

[14] J. Franecki and M. Kapranov. The Gauss map and a noncompact Riemann-Roch formula for constructible sheaves on semiabelian varieties. *Duke Math. J.*, 104(1):171–180, 2000.

[15] T. Gao, T. Y. Li, and M. Wu. Algorithm 846: MixedVol: a software package for mixed-volume computation. *ACM Trans. Math. Software*, 31(4):555–560, 2005.

[16] D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[17] E. Gross, M. Drton, and S. Petrović. Maximum likelihood degree of variance component models. *Electron. J. Stat.*, 6:993–1016, 2012.

[18] Elizabeth, S. Petrović, and J. Verschelde. Interfacing with phcpack. *J. Softw. Algebra Geom.*, 5:20–25, 2013.

[19] J. Harris. *Algebraic geometry*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. A first course, Corrected reprint of the 1992 original.

[20] J. Hauenstein, J. Rodriguez, and B. Sturmfels. Maximum likelihood for matrices with rank constraints. *To appear in the Journal of Algebraic Statistics*, 2013.

[21] J. D. Hauenstein, A. J. Sommese, and C. W. Wampler. Regenerative cascade homotopies for solving polynomial systems. *Appl. Math. Comput.*, 218(4):1240–1246, 2011.

[22] J. D. Hauenstein and F. Sottile. Algorithm 921: alphaCertified: certifying solutions to polynomial systems. *ACM Trans. Math. Software*, 38(4):Art. ID 28, 20, 2012.

[23] S. Hoşten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[24] S. Hoşten and S. Ruffa. Introductory notes to algebraic statistics. *Rend. Istit. Mat. Univ. Trieste*, 37(1-2):39–70, 2005.

[25] S. Hoşten and S. Sullivant. The algebraic complexity of maximum likelihood estimation for bivariate missing data. In *Algebraic and geometric methods in statistics*, pages 123–133. Cambridge Univ. Press, Cambridge, 2010.

[26] J. Huh. The maximum likelihood degree of a very affine variety. *Compos. Math.*, 149(8):1245–1266, 2013.

[27] J. Huh and B. Sturmfels. Likelihood geometry in *Combinatorial Algebraic Geometry Lecture Notes in Mathematics 2108* pages 63-117, Springer 2014.

[28] K. Kubjas, E. Robeva, and B. Sturmfels. Fixed points of the em algorithm and nonnegative rank boundaries. 2013.

[29] J. M. Landsberg and J. Weyman. On the ideals and singularities of secant varieties of Segre varieties. *Bull. Lond. Math. Soc.*, 39(4):685–697, 2007.

[30] A. Leykin. Numerical algebraic geometry. *J. Softw. Algebra Geom.*, 3:5–10, 2011.

[31] A. Moitra. A singly-exponential time algorithm for computing nonnegative rank. arXiv:1205.0044, 2012.

[32] D. Mond, J. Smith, and D. van Straten. Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 459(2039):2821–2845, 2003.

[33] A. Morgan and A. Sommese. A homotopy for solving general polynomial systems that respects m-homogeneous structures. *Appl. Math. Comput.*, 24(2):101–113, 1987.

[34] A. P. Morgan and A. J. Sommese. Coefficient-parameter polynomial continuation. *Applied Mathematics and Computation*, 29(2):123 – 160, 1989.

[35] L. Pachter and B. Sturmfels. Statistics. In *Algebraic statistics for computational biology*, pages 3–42. Cambridge Univ. Press, New York, 2005.

[36] F. Rapallo. Markov bases and structural zeros. *Journal of Symbolic Computation*, 41(2):164 – 172, 2006. Computational Algebraic Statistics Computational Algebraic Statistics.

[37] P. Rostalski and B. Sturmfels. Dualities in convex algebraic geometry. *Rend. Mat. Appl. (7)*, 30(3-4):285–327, 2010.

[38] P. Rostalski and B. Sturmfels. Dualities. In *Semidefinite optimization and convex algebraic geometry*, volume 13 of *MOS-SIAM Ser. Optim.*, pages 203–249. SIAM, Philadelphia, PA, 2013.

[39] A. Sommese, J. Verschelde, and C. Wampler. Using monodromy to decompose solution sets of polynomial systems into irreducible components. In C. Ciliberto, F. Hirzebruch, R. Miranda, and M. Teicher, editors, *Applications of Algebraic Geometry to Coding Theory, Physics and Computation*, volume 36 of *NATO Science Series*, pages 297–315. Springer Netherlands, 2001.

[40] A. J. Sommese, J. Verschelde, and C. W. Wampler. Symmetric functions applied to decomposing solution sets of polynomial systems. *SIAM J. Numer. Anal.*, 40(6):2026–2046 (2003), 2002.

[41] A. J. Sommese and C. W. Wampler, II. *The numerical solution of systems of polynomials*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005. Arising in engineering and science.

[42] S. Steidel. Gröbner bases of symmetric ideals. *J. Symbolic Comput.*, 54:72–86, 2013.

[43] W. Stein et al. *Sage Mathematics Software (Version x.y.z)*. The Sage Development Team. `http://www.sagemath.org`.

[44] B. Sturmfels. Open problems in algebraic statistics. In *Emerging applications of algebraic geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 351–363. Springer, New York, 2009.

[45] C. Uhler. Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.*, 40(1):238–261, 2012.

[46] J. Verschelde. Algorithm 795: Phcpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Trans. Math. Softw.*, 25(2):251–276, June 1999.

[47] J. Verschelde and R. Cools. Symbolic homotopy construction. *Appl. Algebra Engrg. Comm. Comput.*, 4(3):169–183, 1993.

[48] M. Zhu, G. Jiang, and S. Gao. Solving the 100 Swiss francs problem. *Math. Comput. Sci.*, 5(2):195–207, 2011.