Maximum Likelihood

Bernd Sturmfels

February 14, 2023

Our first lecture was concerned with minimizing the distance from a given data point u to a model X that is described by polynomial equations. In the second lecture we considered the analogous problem in the setting of algebraic statistics [23], where the model X represents a family of probability distributions, and we used the Wasserstein metric to measure the distances from u to X. In this lecture we stay in the setting of statistical models but we now use Kullback-Leibler divergence and likelihood inference instead of Wasserstein distance.

The two scenarios of most interest for statisticians are Gaussian models and discrete models. We start with discrete models, where we take the state space is the finite set $\{0, 1, \ldots, n\}$. The simplex of all probability distributions on this state space equals

$$\Delta_n \;=\; \bigl\{ p = (p_0, p_1, \ldots, p_n) \in \mathbb{R}^{n+1} \,:\, p_0 + p_1 + \cdots + p_n = 1 \text{ and } p_0, p_1, \ldots, p_n > 0 \bigr\}. \quad (1)$$

Given two distributions q and p in Δ_n, the *Kullback-Leibler (KL) divergence* is defined as

$$D_{\mathrm{KL}}(\, q \,||\, p \,) \;=\; \sum_{i=0}^{n} q_i \cdot \log(q_i/p_i). \quad (2)$$

This function is not symmetric in its two arguments, i.e. we have $D_{\mathrm{KL}}(\, q \,||\, p \,) \neq D_{\mathrm{KL}}(\, p \,||\, q \,)$ in general. Nevertheless, we interpret KL divergence as a kind of metric on the simplex Δ_n.

Lemma 1. *The KL divergence is nonnegative and it is zero if and only if the two distributions agree. In symbols, $D_{\mathrm{KL}}(\, q \,||\, p \,) \geq 0$ for all $p, q \in \Delta_n$, and equality holds if and only if $p = q$.*

Proof. We use the calculus fact that the function $x \mapsto (x-1) - \log(x)$ is nonnegative for $x \in \mathbb{R}_{>0}$ and its only zero occurs at $x = 1$. Hence sum in (2) is bounded below as follows:

$$D_{\mathrm{KL}}(\, q \,||\, p \,) \;=\; -\sum_{i=0}^{n} q_i \cdot \log(p_i/q_i) \;\geq\; -\sum_{i=0}^{n} q_i \cdot \bigl(p_i/q_i - 1\bigr) \;=\; \sum_{i=0}^{n} p_i \,-\, \sum_{i=0}^{n} q_i \;=\; 1 - 1 \;=\; 0.$$

Moreover, equality holds if and only if $p_i/q_i = 1$ for all indices i. □

Our model is a subset X of Δ_n defined by polynomial equations. As before, for venturing beyond linear algebra, we identify X with its Zariski closure in complex projective space $\mathbb{P}^n$.

We shall present the algebraic approach to maximum likelihood estimation (MLE). See [8, 10, 16, 17, 14, 23] and references therein. Suppose we are given N i.i.d. samples. These are

summarized in the *data vector* $u = (u_0, u_1, \ldots, u_n)$ where u_i is the number of samples that were in state i. Note that $N = u_0 + \cdots + u_n$. The associated log-likelihood function equals

$$\ell_u : \Delta_n \to \mathbb{R}\,,\ p \mapsto u_0 \cdot \log(p_0) + u_1 \cdot \log(p_1) + \cdots + u_n \cdot \log(p_n).$$

Performing MLE for the model X means solving the following optimization problem:

$$\text{Maximize } \ell_u(p) \text{ subject to } p \in X. \tag{3}$$

Viewed through the lens of metric algebraic geometry, this problem amounts to minimizing a certain distance, namely KL divergence, to the variety X. Namely, given a data vector u with $u_i > 0$ for all i, we write $q = \frac{1}{N}u$ for the corresponding empirical distribution in Δ_n

Remark 2. The maximum likelihood estimation problem (3) is equivalent to:

$$\text{Minimize } D_{\mathrm{KL}}(\, q \,||\, p\,) \text{ subject to } p \in X. \tag{4}$$

This holds because the KL divergence can be rewritten as the entropy of the empirical distribution q minus the log-likelihood function: $D_{\mathrm{KL}}(\, q \,||\, p\,) = \sum_{i=0}^{n} q_i \log(q_i) - \frac{1}{N}\,\ell_u(p)$.

As in previous lectures, we identify the model X with a projective variety in $\mathbb{P}^n$. The objective function in the optimization problem (3) involves logarithms and it is not an algebraic function. However, each of its partial derivatives is a rational function, and therefore we can study this problem using algebraic geometry.

We define the *ML degree* of the given projective variety X to be the number of complex critical points for generic data u. The optimal solution is denoted $\hat{p}$ and called the *maximum likelihood estimate* of the model X for the data u. Thus ML degree is the analogue to ED degree, when now KL divergence replaces Euclidean distance.

The critical equations for (3) are similar to those of ED problem. Let $I_X = \langle f_1, \ldots, f_k \rangle$ be the homogeneous ideal of the model X. In addition, we consider the inhomogeneous linear polynomial $f_0 := p_0 + p_1 + \cdots + p_n - 1$. Let $\mathcal{J} = \big(\partial f_i/\partial p_j\big)$ denote the Jacobian matrix of size $(k+1) \times (n+1)$ for these polynomials, and set $c = \mathrm{codim}(X)$. The augmented Jacobian $\mathcal{A}\mathcal{J}$ is obtained by prepending one more row, namely the gradient of the objective function

$$\nabla \ell_u \;=\; \big(\, u_0/p_0,\, u_1/p_1,\, \ldots\,,\, u_n/p_n \,\big).$$

To obtain the critical equations, enlarge I_X by the $(c+2)\times(c+2)$ minors of the $(k+2)\times(n+1)$ matrix $\mathcal{A}\mathcal{J}$, then clear denominators, and remove extraneous components by saturation.

Example 3 (Space curves)**.** Let $n = 3$ and X the curve in Δ_3 defined by two general polynomials f_1 and f_2 of degrees d_1 and d_2 in p_0, p_1, p_2, p_3. The augmented Jacobian matrix is

$$\mathcal{A}\mathcal{J} \;=\; \begin{pmatrix} u_0/p_0 & u_1/p_1 & u_2/p_2 & u_3/p_3 \\ 1 & 1 & 1 & 1 \\ \partial f_1/\partial p_0 & \partial f_1/\partial p_1 & \partial f_1/\partial p_2 & \partial f_1/\partial p_3 \\ \partial f_2/\partial p_0 & \partial f_2/\partial p_1 & \partial f_2/\partial p_2 & \partial f_2/\partial p_3 \end{pmatrix}. \tag{5}$$

Clearing denominators amounts to multiplying the ith column by p_i, so the determinant contributes a polynomial of degree d_1+d_2+1 to the critical equations. Here the codimension equals $c = 2$, we need to take the 4×4 minors of $\mathcal{A}\mathcal{J}$. Since the generators of I_X have degrees d_1 and d_2, we conclude that the ML degree of X equals $d_1 d_2 (d_1 + d_2 + 1)$.

The following general upper bound on the ML degree is established in [14, Theorem 5].

Proposition 4. *Let X be a model of codimension c in Δ_n whose ideal I_X is generated by polynomials $f_1, f_2, \ldots, f_c, \ldots, f_k$ of degrees $d_1 \geq d_2 \geq \cdots \geq d_c \geq \cdots \geq d_k$. Then*

$$MLdegree(X) \quad \leq \quad d_1 d_2 \cdots d_c \cdot \sum_{i_1+i_2+\cdots+i_c \leq n-c} d_1^{i_1} d_2^{i_2} \cdots d_c^{i_c}. \tag{6}$$

Equality holds when X is a generic complete intersection of codimension c (hence $c = k$).

We next present a more precise formula. For the ED degree, the polar degrees in $\mathbb{P}^n$ were used to give such a formula. For the ML degree, we shall use the Euler characteristic.

Given our variety X in the complex projective space $\mathbb{P}^n$, and let X^o be the open subset of X that is obtained by removing $\{p_0 p_1 \cdots p_n (\sum_{i=0}^n p_i) = 0\}$. We recall from [15, 16] that a *very affine variety* is a closed subvariety of an algebraic torus $(\mathbb{C}^*)^r$. Thus X^o is a very affine variety, with $r = n + 2$. The following formula works for any very affine variety.

Theorem 5. *Suppose that the very affine variety X^o is non-singular. The ML degree of the model X equals the signed Euler characteristic $(-1)^{dim(X)} \cdot \chi(X^o)$ of the manifold X^o.*

Proof and Discussion. This was proved under additional assumptions in [8, Theorem 19], and in full generality in [15, Theorem 1]. If X^o is singular then the Euler characteristic can be replaced by the Chern-Schwartz-MacPherson class, as shown in [15, Theorem'2]. □

The optimal solution of (3)-(4) in the statistical model $X^o \cap \Delta_n = X \cap \Delta_n$ is denoted $\hat{p}$. This point is called the *maximum likelihood estimate (MLE)* for the data u and the model X. The ML degree measures the algebraic complexity of the MLE. Theorem 5 says that the ML degree is a topological invariant. Varieties X for which the ML degree is equal to one are of special interest, both statistically and geometrically. ML degree one means that the MLE $\hat{p}$ is a rational function of the data u. Here are two natural examples where this happens.

Example 6 ($n = 3$)**.** The independence model for two binary random variables is a quadratic surface X in the tetrahedron Δ_3. This model is described by the constraints

$$\det \begin{bmatrix} p_0 & p_1 \\ p_2 & p_3 \end{bmatrix} = 0 \quad \text{and} \quad p_0 + p_1 + p_2 + p_3 = 1 \quad \text{and} \quad p_0, p_1, p_2, p_3 > 0.$$

Consider data $u = \begin{bmatrix} u_0 & u_1 \\ u_2 & u_3 \end{bmatrix}$ of *sample size* $|u| = u_0+u_1+u_2+u_3$. The ML degree of the surface X equals one because the MLE $\hat{p}$ is a rational function of the data, namely

$$\begin{aligned} \hat{p}_0 &= |u|^{-2}(u_0+u_1)(u_0+u_2), \quad & \hat{p}_1 &= |u|^{-2}(u_0+u_1)(u_1+u_3), \\ \hat{p}_2 &= |u|^{-2}(u_2+u_3)(u_0+u_2), \quad & \hat{p}_3 &= |u|^{-2}(u_2+u_3)(u_1+u_3). \end{aligned} \tag{7}$$

In words, we multiply the row sums with the column sums in the empirical distribution $\frac{1}{|u|}u$.

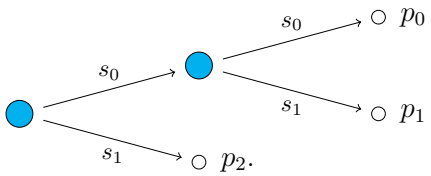

Figure 1: Probability tree that describes the coin toss model in Example 7.

Example 7 ($n = 2$)**.** Given a biased coin, we perform the following experiment: *Flip a biased coin. If it shows heads, flip it again.* The outcome is the number of heads: 0, 1 or 2.

If s is the bias of the cone, then the model is the parametric curve X given by

$$(0,1) \to X \subset \Delta_2\,,\ \ s \mapsto \big(s^2, s(1-s), 1-s\big).$$

This model is the conic $X = V(p_0p_2 - (p_0+p_1)p_1) \subset \mathbb{P}^2$. The MLE is given by the formula

$$(\hat{p}_0, \hat{p}_1, \hat{p}_2) \ = \ \left(\frac{(2u_0+u_1)^2}{(2u_0+2u_1+u_2)^2}\,,\ \frac{(2u_0+u_1)(u_1+u_2)}{(2u_0+2u_1+u_2)^2}\,,\ \frac{u_1+u_2}{2u_0+2u_1+u_2}\right). \tag{8}$$

Since the coordinates of $\hat{p}$ are rational functions, the ML degree of X is equal to one.

The following theorem explains what we saw in equations (7) and (8):

Theorem 8. *If $X \subset \Delta_n$ is a model of ML degree one, so $\hat{p}$ is a rational function of u, then each coordinate $\hat{p}_i$ is an alternating product of linear forms with positive coefficients.*

Proof and Discussion. This was shown in the setting of arbitrary complex very affine varieties by Huh in [16]. It was adapted to real algebraic geometry and statistical models in [10]. These articles offer precise statements via Horn uniformization for A-discriminants [12], i.e. hypersurfaces dual to toric varieties. For more information see [17, Corollary 3.12]. □

Models given by rank constraints on matrices and tensors are particularly important in applications, since these represent conditional independence. Consider two random variables, having n_1 and n_2 states respectively, which are conditionally independent, given a hidden random variable with r states. In algebraic geometry, this model is the variety X_r in $\mathbb{P}^{n_1n_2-1}$ that is defined by the $(r+1) \times (r+1)$ minors of an $n_1 \times n_2$ matrix (p_{ij}). The ML degree of this rank r model was first studied by Hauenstein, Rodriguez and Sturmfels in [13], who obtained the following results using methods from numerical algebraic geometry.

Proposition 9. *For small values of n_1 and n_2, the ML degrees of low rank models X_r are*

$(n_1, n_2) =$	$(3,3)$	$(3,4)$	$(3,5)$	$(4,4)$	$(4,5)$	$(4,6)$	$(5,5)$
$r=1$	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

(9)

Every entry in the $r = 1$ row is 1 because the MLE for the independence model is a rational function in the data (u_{ij}). One finds $\hat{p} = (\hat{p}_{ij})$ by multiplying the column vector of row sums of u with the row vector of column sums of u, and then dividing by $|u|^2$, as shown in (7) for $n_1 = n_2 = 2$. The other entries are more interesting, and they give precise information on the algebraic complexity of minimizing the Kullback-Leibler distance from a given data matrix u to the conditional independence model X_r. Here is a matrix from [13].

Example 10 ($n_1 = n_2 = 5$)**.** Following [13, Example 7], we consider the data

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

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 11 below. 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.

The columns of the table in (9) exhibit an obvious symmetry. This was conjectured in [13], and it was proved by Draisma and Rodriguez in their article [9] on maximum likelihood duality. Given an $n_1 \times n_2$ matrix u, we write Ω_u for the matrix whose (i, j) entry equals

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

In the following theorem, the symbol $\star$ denotes the Hadamard product (or entrywise product) of two matrices. All matrices p_i and q_i have format $n_1 \times n_2$ and they have complex entries.

Theorem 11. *Fix $n_1 \leq n_2$ and u an $n_1 \times n_2$-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 for u on X_r and the complex critical points $q_1, q_2, \ldots, q_s$ on X_{n_1-r+1} such that*

$$p_1 \star q_1 \;=\; p_2 \star q_2 \;=\; \cdots \;=\; p_s \star q_s \;=\; \Omega_u. \tag{10}$$

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

This result represents a multiplicative version of the duality we encountered in our study of ED degrees. Recall that the ED degree of any projective variety X equals that of its dual variety $X^\vee$. Under some genericity assumption, this common ED degree is the sum of the polar degrees, which arises from the conormal variety $N_X = N_{X^\vee}$. By "multiplicative" we mean that u_i/p_i instead of $u_i - p_i$ appears in the first row of the augmented Jacobian matrix.

It is an interesting challenge in intersection theory and singularity theory to find general formulas for the ML degrees encountered in Proposition 9. This problem was solved for $r = 2$ by Rodriguez and Wang in [20]. They give a recursive formula in [20, Theorem 4.1], and they present impressive values in [20, Table 1]. They unravel the recursion, and they obtain the explicit formulas for the ML degree of conditional independence in many cases. In particular, they obtain the following result which had been stated as a conjecture in [13].

Theorem 12 (Rodriguez-Wang [20]). *Consider the variety $X_2 \subset \mathbb{P}^{3n-1}$ whose points are the $3 \times n$ matrices of rank ≤ 2. The ML degree of this variety equals $2^{n+1} - 6$.*

We now turn to a connection between algebraic statistics and particle physics that was developed in [21]. The physics context is scattering amplitudes, where the critical equations for (3)-(4) are known as the *scattering equations*. We consider the *CEGM model*, due to Cachazo and his collaborators [6, 7]. The role of the data vector u is played by the *Mandelstam invariants*. This theory rests on the space X^o of m labeled points in general position in $\mathbb{P}^{k-1}$, up to projective transformations. Consider the Grassmannian $\mathrm{Gr}(k,m)$ in its Plücker embedding into $\mathbb{P}^{\binom{m}{k}-1}$. The torus $(\mathbb{C}^*)^m$ acts on $\mathrm{Gr}(k,m)$ by scaling the columns of $k \times m$ matrices representing subspaces. Let $\mathrm{Gr}(k,m)^o$ be the open Grassmannian where all Plücker coordinates are nonzero. The CEGM model is the $(k-1)(m-k-1)$-dimensional manifold

$$X^o \quad = \quad \mathrm{Gr}(k,m)^o/(\mathbb{C}^*)^m. \tag{11}$$

Example 13 ($k = 2$). For $k = 2$, the very affine variety in (11) has dimension $m-3$, and it is the moduli space of m distinct labeled points on the complex projective line $\mathbb{P}^1$. This space is ubiquitous in algebraic geometry where it is known as $\mathcal{M}_{0,m}$. The point of our discussion here is to interpret $\mathcal{M}_{0,m}$ as a statistical model, and to argue that its ML degree is equal to $(m-3)!$. For instance, if $m = 4$ then $X^o = \mathcal{M}_{0,4}$ is the Riemann sphere $\mathbb{P}^1$ with three points removed. The signed Euler characteristic of this surface is one, and Theorem 8 applies.

Proposition 14. *The variety X^o in (11) is very affine, with coordinates given by the $k \times k$ minors of the following $k \times m$ matrix, which we denote by $M_{k,m}$:*

$$\begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & (-1)^k & 1 & 1 & 1 & \cdots & 1 \\
0 & 0 & 0 & \cdots & (-1)^{k-1} & 0 & 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m-k-1} \\
\vdots & \vdots & \vdots & \iddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & -1 & \cdots & 0 & 0 & 1 & x_{k-3,1} & x_{k-3,2} & \cdots & x_{k-3,m-k-1} \\
0 & 1 & 0 & \cdots & 0 & 0 & 1 & x_{k-2,1} & x_{k-2,2} & \cdots & x_{k-2,m-k-1} \\
-1 & 0 & 0 & \cdots & 0 & 0 & 1 & x_{k-1,1} & x_{k-1,2} & \cdots & x_{k-1,m-k-1}
\end{bmatrix}.$$

To be precise, the coordinates on $X^o \subset (\mathbb{C}^*)^{\binom{m}{k}}$ are the non-constant minors $p_{i_1 i_2 \cdots i_k}$.

Following [1, equation (4)], the antidiagonal matrix in the left $k \times k$ block of $M_{k,m}$ is chosen so that each unknown $x_{i,j}$ is precisely equal to $p_{i_1 i_2 \cdots i_k}$ for some $i_1 < i_2 < \cdots < i_k$. The *scattering potential* for the CEGM model is the following multivalued function on X^o:

$$\ell_u \quad = \quad \sum_{i_1 i_2 \cdots i_k} u_{i_1 i_2 \cdots i_k} \cdot \log(p_{i_1 i_2 \cdots i_k}). \tag{12}$$

The critical point equations, known as *scattering equations* [1, equation (7)], are given by

$$\frac{\partial \ell_u}{\partial x_{i,j}} = 0 \qquad \text{for } 1 \leq i \leq k-1 \text{ and } 1 \leq j \leq m-k-1. \tag{13}$$

These are equations of rational functions. Solving these equations is the agenda in [6, 7, 21].

Corollary 15. *The number of complex solutions to (13) is the ML degree of the CEGM model X^o. This number equals the signed Euler characteristic $(-1)^{(k-1)(m-k-1)} \cdot \chi(X^o)$.*

Example 16 $(k = 2, m = 6)$**.** The very affine threefold $X^o = \mathcal{M}_{0,6}$ is embedded in $(\mathbb{C}^*)^9$ via

$$p_{24} = x_1,\ p_{25} = x_2,\ p_{26} = x_3,\ p_{34} = x_1 - 1,\ p_{35} = x_2 - 1,$$
$$p_{36} = x_3 - 1,\ p_{45} = x_2 - x_1,\ p_{46} = x_3 - x_1,\ p_{56} = x_3 - x_2.$$

These nine coordinates on $X^o \subset (\mathbb{C}^*)^9$ are the non-constant 2×2 minors of our matrix

$$M_{2,6} \ = \ \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & x_1 & x_2 & x_3 \end{bmatrix}.$$

The scattering potential is the analogue to the log-likelihood function in statistics:

$$\ell_u \ = \ u_{24}\log(p_{24}) + u_{25}\log(p_{25}) + \cdots + u_{56}\log(p_{56}).$$

This function has six critical points in X^o. Hence $\text{MLdegree}(X^o) = -\chi(X^o) = 6$.

We now examine the number of critical points of the scattering potential (12).

Theorem 17. *The known values of the ML degree for the CEGM model (11) are as follows. For $k = 2$, the ML degree equals $(m-3)!$ for all $m \geq 4$. For $k = 3$, it equals $2, 26, 1272, 188112, 74570400$ for $m = 5, 6, 7, 8, 9$, and for $k = 4, m = 8$ it equals 5211816.*

Proof. We refer to [1, Example 2.2], [1, Theorem 5.1] and [1, Theorem 6.1] for $k = 2, 3, 4$. □

Knowing these ML degrees helps in solving the scattering equations reliably. We demonstrated in [1, 21] how this can be done in practice with `HomotopyContinuation.jl` [5, 4]. For instance, we see in [21, Table 1] that the $10! = 3628800$ solutions for $k = 2, m = 13$ are found in under one hour. See [1, Section 6] for the solution in the challenging case $k = 4, m = 8$.

We now change topic by turning to models for Gaussian random variables. Let PD_n denote the open convex cone of positive-definite symmetric $n \times n$ matrices. This cone now plays the role which was played by the simplex Δ_n when we discussed discrete models above.

Given a mean vector $\mu \in \mathbb{R}^n$ and a covariance matrix $\Sigma \in \text{PD}_n$, the associated *Gaussian distribution* is supported on $\mathbb{R}^n$. Its density has the familiar "bell shape"; it is the function

$$f_{\mu,\Sigma}(x) \ := \ \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

We fix a model $Y \subset \mathbb{R}^n \times \text{PD}_n$ that is defined by polynomial equations in (μ, Σ). Suppose we are given N samples $U^{(1)}, \ldots, U^{(N)}$ in $\mathbb{R}^n$. These are summarized in the *sample mean* $\bar{U} = \frac{1}{N}\sum_{i=1}^N U^{(i)}$ and in the *sample covariance matrix* $S = \frac{1}{N}\sum_{i=1}^N (U^{(i)} - \bar{U})(U^{(i)} - \bar{U})^T$. Given these data, the log-likelihood is the following function in the unknowns (μ, Σ):

$$\ell(\mu, \Sigma) \quad = \quad -\frac{N}{2} \cdot \left[\log\det\Sigma \ + \ \text{trace}(S\Sigma^{-1}) \ + \ (\bar{U}-\mu)^T\Sigma^{-1}(\bar{U}-\mu)\right]. \qquad (14)$$

The task of likelihood inference is to minimize this function subject to $(\mu, \Sigma) \in Y$.

There are two extreme cases. First, consider a model where Σ is fixed to be the identity matrix Id_n. Then $Y = X \times \{\mathrm{Id}_n\}$ and we are supposed to minimize the Euclidean distance from the sample mean $\bar{U}$ to the variety X in $\mathbb{R}^n$. This is precisely the earlier ED problem.

We instead focus on the second case, the family of *centered Gaussians*, where μ is fixed at zero. The model has the form $\{0\} \times X$, where X is a variety in the space $\mathrm{Sym}_2(\mathbb{R}^n)$ of symmetric $n \times n$ matrices. Following [23, Proposition 7.1.10], our task is now as follows:

$$\text{Minimize the function } \ \Sigma \ \mapsto \ \log\det\Sigma \ + \ \mathrm{trace}(S\,\Sigma^{-1}) \quad \text{subject to } \Sigma \in X. \tag{15}$$

Using the concentration matrix $K = \Sigma^{-1}$, we can write this equivalently as follows:

$$\text{Maximize the function } \ K \ \mapsto \ \log\det K \ - \ \mathrm{trace}(S\,K) \quad \text{subject to } K \in X^{-1}. \tag{16}$$

Here the variety X^{-1} is the Zariski closure of the set of inverses of all matrices in X.

Remark 18. The optimization problem (15)-(16) has a metric interpretation as in (4). Namely, we can define the KL divergence between two probability distributions on $\mathbb{R}^n$ by replacing the sum in (2) by the corresponding integral over $\mathbb{R}^n$. For two Gaussians we obtain a certain kind of distance between the unknown Σ and the sample covariance matrix S.

The critical equations of the optimization problem (15)-(16) can be written as polynomials, since the partial derivatives of the logarithm are rational functions. These equations have finitely many complex solutions. Their number is the *ML degree* of the model X^{-1}.

In the remainder of this lecture we focus on Gaussian models that are described by linear constraints on either the covariance matrix or its inverse, which is the concentration matrix. Let $\mathcal{L} \subset \mathrm{Sym}_2(\mathbb{R}^n)$ be a linear space of symmetric matrices (LSSM), whose general element is assumed to be invertible. We are interested in the models $X^{-1} = \mathcal{L}$ and $X = \mathcal{L}$. It is convenient to use primal-dual coordinates (Σ, K) to write the respective critical equations.

Proposition 19. *Fix an LSSM $\mathcal{L}$ and its orthogonal complement $\mathcal{L}^\perp$ for the inner product $\langle X, Y\rangle = \mathrm{trace}(XY)$. The critical equations for the* linear concentration model $X^{-1} = \mathcal{L}$ *are*

$$K \in \mathcal{L} \ \text{ and } \ K\Sigma = Id_n \ \text{ and } \ \Sigma - S \in \mathcal{L}^\perp. \tag{17}$$

The critical equations for the linear covariance model $X = \mathcal{L}$ *are*

$$\Sigma \in \mathcal{L} \ \text{ and } \ K\Sigma = Id_n \ \text{ and } \ KSK - K \in \mathcal{L}^\perp. \tag{18}$$

Proof. This is well-known in statistics. For proofs see [22, Propositions 3.1 and 3.3]. □

The system (17) is linear in K, but the last group of equations in (18) is quadratic in K. The numbers of complex solutions are the *ML degree* of $\mathcal{L}$ and the *reciprocal ML degree* of $\mathcal{L}$. The former is smaller than the latter, and (17) is easier to solve than (18).

Example 20. Let $n = 4$ and $\mathcal{L}$ a generic LSSM of dimension k. Our degrees are as follows:

$k = \dim(\mathcal{L})$:	2	3	4	5	6	7	8	9
ML degree :	3	9	17	21	21	17	9	3
reciprocal ML degree :	5	19	45	71	81	63	29	7

These numbers and many more appear in [22, Table 1].

ML degrees and the reciprocal ML degrees have been studied intensively in the recent literature, both for generic and special spaces $\mathcal{L}$. See [2, 3, 11] and the references therein. We now present an important result due to Manivel, Michałek, Monin, Seynnaeve, Vodička and Wiśniewski. Theorem 21 paraphrases highlights from their articles [18, 19].

Theorem 21. *The ML degree of a generic linear subspace $\mathcal{L}$ of dimension k in $Sym_2(\mathbb{R}^n)$ is the number of quadrics in $\mathbb{P}^{n-1}$ that pass through $\binom{n+1}{2} - k$ general points and are tangent to $k-1$ general hyperplanes. For fixed k, this number is a polynomial in n of degree $k-1$.*

Proof. The first statement is [19, Corollary 2.6 (4)], here interpreted classically in terms of Schubert calculus. For a detailed discussion see the introduction of [18]. The second statement appears in [18, Theorem 1.3 and Corollary 4.13]. □

Example 22 ($n = 4$)**.** Fix $10 - k$ points and $k - 1$ planes in $\mathbb{P}^3$. We are interested in all quadratic surfaces that contain the points and are tangent to the planes. This points and planes impose 9 constraints on $\mathbb{P}(\mathrm{Sym}_2(\mathbb{C}^4)) \simeq \mathbb{P}^9$. Passing through a point is a linear equation. Being tangent to a plane is a cubic constraint on $\mathbb{P}^9$. Bézout's Theorem suggests that there could be 3^{k-1} solutions. This is correct for $k \leq 3$ but it overcounts for $k \geq 4$. Indeed, in Example 20 we see $17, 21, 21, \ldots$ instead of $27, 81, 243, \ldots$.

The intersection theory approeach in [19, 18] leads to formulas for the ML degrees of linear Gaussian models. From this we obtain provably correct numerical methods for maximum likelihood estimation. Namely, after computing critical points as in [22], we can certify them as in [4]. Since the ML degree is known, one can be sure that all solutions have been found.

References

[1] D. Agostini, T. Brysiewicz, C. Fevola, L. Kühne, B. Sturmfels and S. Telen: Likelihood degenerations, *Advances in Mathematics*, to appear.

[2] C. Améndola, L. Gustafsson, K. Kohn, O. Marigliano, and A. Seigal: The maximum likelihood degree of linear spaces of symmetric matrices, *Matematiche (Catania)* **76** (2021), 535-?557.

[3] T. Boege, J. Coons, C. Eur and F. Röttger: Reciprocal maximum likelihood degrees of Brownian motion tree models, *Matematiche (Catania)* **76** (2021), 535-?557.

[4] P. Breiding, K. Rose and S. Timme: Certifying zeros of polynomial systems using interval arithmetic, *Trans. Math. Software*, to appear.

[5] P. Breiding and S. Timme: HomotopyContinuation.jl: A Package for Homotopy Continuation in Julia, *Math. Software – ICMS 2018*, 458–465, Springer, 2018.

[6] F. Cachazo, N. Early, A. Guevara and S. Mizera: Scattering equations: from projective spaces to tropical Grassmannians, *J. High Energy Phys.* (2019), no. 6, 039.

[7] F. Cachazo, B. Umbert and Y. Zhang: Singular solutions in soft limits, *J. High Energy Phys.* (2020), no. 5, 148.

[8] F. Catanese, S. Hoşten, A. Khetan and B. Sturmfels: The maximum likelihood degree, *American Journal of Mathematics* **128** (2006) 671–697.

[9] J. Draisma and J. Rodriguez: Maximum likelihood duality for determinantal varieties, *Int. Math. Res. Not. IMRN* **20** (2014) 5648-–5666.

[10] E. Duarte, O. Marigliano and B. Sturmfels: Discrete statistical models with rational maximum likelihood estimator, *Bernoulli* **27** (2021) 135–154.

[11] C. Eur, T. Fife, J. Samper and T. Seynnaeve: Reciprocal maximum likelihood degrees of diagonal linear concentration models, *Matematiche (Catania)* **76** (2021), 447–459.

[12] I. Gel'fand, M. Kapranov and A. Zelevinsky: *Discriminants, Resultants and Multidimensional Determinants*, Birkhäuser, Boston, 1994.

[13] J. Hauenstein, J. Rodriguez and B. Sturmfels: Maximum likelihood for matrices with rank constraints, *J. Algebr. Stat.* **5** (2014) 18–38.

[14] S. Hoşten, A. Khetan and B. Sturmfels: Solving the likelihood equations, *Found. Comput. Math.* **5** (2005) 389–407.

[15] J. Huh: The maximum likelihood degree of a very affine variety, *Compositio Math.* **149** (2013) 1245–1266.

[16] J. Huh: Varieties with maximum likelihood degree one, *J. Algebraic Statistics* **5** (2014) 1–17.

[17] J. Huh and B. Sturmfels: Likelihood geometry, *Combinatorial Algebraic Geometry*. Lecture Notes in Mathematics **2108**, Springer Verlag, 63–117, 2014.

[18] L. Manivel, M. Michałek, L. Monin, T. Seynnaeve and M. Vodička: Complete quadrics: Schubert calculus for Gaussian models and semidefinite programming, *Journal of the European Mathematical Society*, to appear.

[19] M. Michałek, L. Monin and J. Wiśniewski: Maximum likelihood degree, complete quadrics, and $\mathbb{C}^*$-action, *SIAM J. Appl. Algebra Geom.* **5** (2021) 60–85.

[20] J. Rodriguez and B. Wang: The maximum likelihood degree of mixtures of independence models, *SIAM J. Appl. Algebra Geom.* **1** (2017) 484-–506.

[21] B. Sturmfels and S. Telen: Likelihood equations and scattering amplitudes, *Algebraic Statistics* **12** (2021) 167–186.

[22] B. Sturmfels, S. Timme and P. Zwiernik: Estimating linear covariance models with numerical nonlinear algebra, *Algebraic Statistics* **1** (2020) 31–52.

[23] S. Sullivant: *Algebraic Statistics*, Graduate Studies in Mathematics 194, American Mathematical Society, Providence, 2018.