Maximum Likelihood for Matrices with Rank Constraints

Bernd Sturmfels
UC Berkeley

Based on a joint paper with
Jon Hauenstein (NCSU) and Jose Rodriguez (Berkeley)
and a subsequent paper by Jose and Jan Draisma (Eindhoven)
Joint Probability Distributions

Consider two random variables having $m$ and $n$ states. Their joint probability distribution is an $m \times n$-matrix

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

whose entries are non-negative and sum to 1.

Let $\mathcal{M}_r$ be the manifold of rank $r$ matrices in the simplex $\Delta_{mn-1}$. Matrices $P$ in $\mathcal{M}_1$ represent independent distributions.

The model $\mathcal{M}_r$ comprises mixtures of $r$ independent distributions. Its elements $P$ represent conditionally independent distributions.
Does Watching Soccer Cause Hair Loss?

296 British subjects aged 40 to 50 were interviewed about their hair length and how many hours per week they watch soccer on TV. The data are summarized in a $3 \times 3$ contingency table:

\[
U = \begin{pmatrix}
\text{lots of hair} & \text{medium hair} & \text{little hair} \\
\leq 2 \text{ hrs} & 51 & 45 & 33 \\
2-6 \text{ hrs} & 28 & 30 & 29 \\
\geq 6 \text{ hrs} & 15 & 27 & 38
\end{pmatrix}
\]

Is there a correlation between watching soccer and hair loss?
Does Watching Soccer Cause Hair Loss?

296 British subjects aged 40 to 50 were interviewed about their hair length and how many hours per week they watch soccer on TV. The data are summarized in a $3 \times 3$ contingency table:

$$
\begin{pmatrix}
\text{lots of hair} & \text{medium hair} & \text{little hair} \\
\leq 2 \text{ hrs} & 51 & 45 & 33 \\
2–6 \text{ hrs} & 28 & 30 & 29 \\
\geq 6 \text{ hrs} & 15 & 27 & 38 \\
\end{pmatrix}
$$

Is there a correlation between watching soccer and hair loss?

Not really. There is a hidden random variable, namely gender. The table is the sum of a table for 126 males and one for 170 females:

$$
U = \begin{pmatrix}
3 & 9 & 15 \\
4 & 12 & 20 \\
7 & 21 & 35 \\
\end{pmatrix} + \begin{pmatrix}
48 & 36 & 18 \\
24 & 18 & 9 \\
8 & 6 & 3 \\
\end{pmatrix}.
$$

Both tables have rank 1, hence $U$ has rank 2. We cannot reject

$H_0 : \text{Soccer on TV and Hair Growth are Independent Given Gender}.$
The Likelihood Function

Suppose i.i.d. samples are drawn from an unknown distribution. We summarize these data also in a matrix

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

The entries of \( U \) are non-negative integers whose sum is \( u_{++} \). The likelihood function of these data is the monomial

\[ \ell_U = \prod_{i=1}^{m} \prod_{j=1}^{n} p_{ij}^{u_{ij}}. \]

Our problem: Maximize \( \ell_U(P) \) subject to \( P \in M_r \).

The solution \( \hat{P} \) is a rank \( r \) matrix. This is the maximum likelihood estimate for \( U \).
Rank One

For $r = 1$, the maximum likelihood estimate $\hat{P}$ is obtained from the data matrix $U$ as follows. Multiply the vector of row sums with the vector of column sums and divide by the sample size:

$$\hat{P} = \frac{1}{(u_{++})^2} \cdot \begin{pmatrix} u_1+ \\ u_2+ \\ \vdots \\ u_m+ \end{pmatrix} \cdot \begin{pmatrix} u_{+1} & u_{+2} & \cdots & u_{+n} \end{pmatrix}$$

This analytic solution is very simple.

More accurately:

This algebraic solution is very simple.

- The MLE $\hat{P}$ is a rational function of the data $U$.
- What does “rational function” mean?
- The function $U \mapsto \hat{P}$ is an algebraic function of degree 1.
3 \times 3\text{-}Matrices of Rank 2

Optimization Problem:

Maximize \( p_{11}^{u_{11}} p_{12}^{u_{12}} p_{13}^{u_{13}} p_{21}^{u_{21}} p_{22}^{u_{22}} p_{23}^{u_{23}} p_{31}^{u_{31}} p_{32}^{u_{32}} p_{33}^{u_{33}} \) subject to

\[
\det(P) = p_{11} p_{22} p_{33} - p_{11} p_{23} p_{32} - p_{12} p_{21} p_{33} + p_{12} p_{23} p_{31} + p_{13} p_{21} p_{32} - p_{13} p_{22} p_{31} = 0 \quad \text{and} \quad p_{++} = p_{11} + p_{12} + p_{13} + p_{21} + p_{22} + p_{23} + p_{31} + p_{32} + p_{33} = 1.
\]

Equations for the Critical Points:

\[
\det(P) = 0 \quad \text{and} \quad p_{++} = 1
\]

and the maximal minors of this 3 \times 9-matrix are zero:

\[
\begin{bmatrix}
  u_{11} & u_{12} & u_{13} & u_{21} & u_{22} & u_{23} & u_{31} & u_{32} & u_{33} \\
  p_{11} & p_{12} & p_{13} & p_{21} & p_{22} & p_{23} & p_{31} & p_{32} & p_{33} \\
  p_{11} a_{11} & p_{12} a_{12} & p_{13} a_{13} & p_{21} a_{21} & p_{22} a_{22} & p_{33} a_{33} & p_{31} a_{31} & p_{32} a_{32} & p_{33} a_{33}
\end{bmatrix}
\]

where \( a_{ij} = \frac{\partial \det(P)}{\partial p_{ij}} \). These equations have 10 complex solutions.
Determinantal Varieties

We are interested in the critical points of the likelihood function on the semialgebraic set $\mathcal{M}_r$ in $\Delta_{mn-1}$.

It is \textit{easier} to study the variety $\mathcal{V}_r$ of complex matrices of rank $\leq r$. In algebra language, $\mathcal{V}_r$ is the Zariski closure of $\mathcal{M}_r$ in $\mathbb{P}^{mn-1}$.

The codimension of $\mathcal{V}_r$ is $(m - r)(n - r)$. How about the degree?

\begin{example}
Consider the determinantal varieties $\mathcal{V}_r$ defined by $4 \times 5$-matrices $P = \begin{pmatrix}
p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\
p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\
p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\
p_{41} & p_{42} & p_{43} & p_{44} & p_{45}
\end{pmatrix}$

$\mathcal{V}_1$ has dimension 7 and degree 35 in $\mathbb{P}^{19}$, $\mathcal{V}_2$ has dimension 13 and degree 50 in $\mathbb{P}^{19}$, $\mathcal{V}_3$ has dimension 17 and degree 10 in $\mathbb{P}^{19}$.
\end{example}
Determinantal Varieties

We are interested in the critical points of the likelihood function on the semialgebraic set $\mathcal{M}_r$ in $\Delta_{mn-1}$.

It is easier to study the variety $\mathcal{V}_r$ of complex matrices of rank $\leq r$. In algebra language, $\mathcal{V}_r$ is the Zariski closure of $\mathcal{M}_r$ in $\mathbb{P}^{mn-1}$.

The codimension of $\mathcal{V}_r$ is $(m - r)(n - r)$. How about the degree?

Example

Consider the determinantal varieties $\mathcal{V}_r$ defined by $4 \times 5$-matrices

$$P = \begin{pmatrix}
p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\
p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\
p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\
p_{41} & p_{42} & p_{43} & p_{44} & p_{45}
\end{pmatrix}$$

- $\mathcal{V}_1$ has dimension 7 and degree 35 in $\mathbb{P}^{19}$,
- $\mathcal{V}_2$ has dimension 13 and degree 50 in $\mathbb{P}^{19}$,
- $\mathcal{V}_3$ has dimension 17 and degree 10 in $\mathbb{P}^{19}$.

Secant varieties of the Segre variety $\mathbb{P}^{m-1} \times \mathbb{P}^{n-1}$
ML Degree

The ML degree of a statistical model (or a projective variety) is the number of critical points of the likelihood function for generic data.

Theorem

The known values for the ML degrees of the rank varieties $\mathcal{V}_r$ are

$$ (m, n) = (3, 3) \quad (3, 4) \quad (3, 5) \quad (4, 4) \quad (4, 5) \quad (4, 6) \quad (5, 5) $$

\begin{array}{ccccccccc}
\hline
r=1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hline
r=2 & 10 & 26 & 58 & 191 & 843 & 3119 & 6776 & \\
\hline
r=3 & 1 & 1 & 1 & 191 & 843 & 3119 & 61326 & \\
\hline
r=4 & 1 & 1 & 1 & 6776 & & & & \\
\hline
r=5 & 1 & & & & & & & 1 \\
\hline
\end{array}

The work of June Huh identifies the ML degree with a certain topological invariant of the very affine variety $\mathcal{V}_r \{ p_{++} \cdot \prod p_{ij} = 0 \}$. 
Geometric Formulation

An \( m \times n \)-matrix \( P \) is a regular point in \( \mathcal{V}_r \) iff \( \text{rank}(P) = r \).

The tangent space \( T_P \) is a subspace of dimension \( rn + rm - r^2 \) in \( \mathbb{C}^{m \times n} \). Its orthogonal complement \( T_P^{\perp} \) has dimension \( (m-r)(n-r) \).

We lift the \textit{log-likelihood function} \( \log(\ell_U) \) to \( \mathbb{P}^{mn-1} \).

The partial derivatives of this function on \( \mathbb{P}^{mn-1} \) are

\[
\frac{\partial \log(\ell_U)}{\partial p_{ij}} = \frac{u_{ij}}{p_{ij}} - \frac{u_{++}}{p_{++}}.
\]

Proposition

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 matrix

\[
\begin{bmatrix}
\frac{u_{ij}}{p_{ij}} - \frac{u_{++}}{p_{++}} \\
\end{bmatrix}_{i=1,\ldots,m \atop j=1,\ldots,n}
\]
Linear Algebra Formulation

Assume $m \leq n$. Let $P_1, R_1, L_1$ and $\Lambda$ be matrices of unknowns of formats $r \times r$, $r \times (n-r)$, $(m-r) \times r$, and $(n-r) \times (m-r)$. Set

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

where $I_{m-r}$ and $I_{n-r}$ are identity matrices.

Proposition

Fix a general $m \times n$ data matrix $U$. The polynomial system

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

consists of $mn$ equations in $mn$ unknowns. It has finitely many complex solutions $(P_1, L_1, R_1, \Lambda)$. The resulting $m \times n$-matrices $P$ are precisely the critical points of the likelihood function $\ell_U$ on the determinantal variety $\mathcal{V}_r$. 
Symmetric Matrices

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

Theorem

The known values for ML degrees of rank \( r \) symmetric matrices (mixtures of identically distributed random variables) are

\[
\begin{array}{c|cccc}
  n & 3 & 4 & 5 & 6 \\
r = 1 & 1 & 1 & 1 & 1 \\
r = 2 & 6 & 37 & 270 & 2341 \\
r = 3 & 1 & 37 & 1394 & \\
r = 4 & 1 & 270 & \\
r = 5 & 1 & 2341 & 
\end{array}
\]
A Symmetric $3 \times 3$-Matrix

Consider the symmetric matrix model 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 are real and positive:

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

The first three points are local maxima in $\Delta_5$ and the last three points are local minima. These six points define an extension of degree 6 over $\mathbb{Q}$. Can the coordinates be written in radicals?
The Joy of Arithmetic

**Yes**, the Galois group of the $3 \times 3$ symmetric problem is a subgroup of $S_6$ that is isomorphic to the solvable group $S_4$.

The minimal polynomial for the last coordinate equals

$$
9528773052286944p_{33}^6 - 4125267629399052p_{33}^5 \\
+ 713452955656677p_{33}^4 - 6334941958182p_{33}^3 \\
+ 3049564842009p_{33}^2 - 75369770028p_{33} + 744139872 = 0.
$$

We solve this equation in radicals as follows:

$$
p_{33} = \frac{16427}{227664} + \frac{1}{12} (\zeta - \zeta^2) \omega_2 - \frac{66004846384302}{19221271018849} \omega_2^2 + \\
\left( \frac{14779904193}{211433981207339} \zeta^2 - \frac{14779904193}{211433981207339} \zeta \right) \omega_1 \omega_2 + \frac{1}{2} \omega_3
$$

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

$$
\begin{align*}
\omega_2^3 &= \left( \frac{5992589425361}{15097277084532208} \zeta - \frac{5992589425361}{15097277084532208} \zeta^2 \right) + \frac{97163}{40083040181952} \omega_1, \\
\omega_3^2 &= \frac{5006721709}{1248260766912} + \left( \frac{212309132509}{4242035935404} \zeta - \frac{212309132509}{4242035935404} \zeta^2 \right) \omega_2 - \frac{2409}{20272573168} \omega_1 \omega_2 \\
&\quad - \frac{158808750548335}{76885084075396} \omega_2^2 + \left( \frac{17063004159}{422867962414678} \zeta^2 - \frac{17063004159}{422867962414678} \zeta \right) \omega_1 \omega_2.
\end{align*}
$$
Duality

Theorem (Draisma and Rodriguez)

Let \( m \leq n \) and consider either general or symmetric matrices. Then the ML degrees for rank \( r \) and for rank \( m-r+1 \) coincide.

Given a data matrix \( U \) of format \( m \times n \), we write \( \Omega_U \) for the \( m \times n \)-matrix whose \((i,j)\) entry equals

\[
\frac{u_{ij} \cdot u_{i+} \cdot u_{+j}}{(u_{++})^3}.
\]

Theorem (Draisma and Rodriguez)

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 \( \ell_U \) on \( \mathcal{V}_r \) and the complex critical points \( Q_1, Q_2, \ldots, Q_s \) of \( \ell_U \) on \( \mathcal{V}_{m-r+1} \) such that

\[
P_1 \star Q_1 = P_2 \star Q_2 = \cdots = P_s \star Q_s = \Omega_U.
\]

Thus, this bijection preserves reality, positivity, and rationality.
We had tested these conjectures for thousands of instances. All our computational results were obtained using the software Bertini.

Bertini is numerical software, based on homotopy continuation, for finding all complex solutions to a system of polynomial equations (and much more). The developers are Daniel Bates, Jonathan Hauenstein, Andrew Sommese, Charles Wampler.

They have a new textbook.

Numerical Algebraic Geometry will be one of the topics featured in the Fall 2014 special research semester on Algorithms and Complexity in Algebraic Geometry at Berkeley’s Simons Institute for the Theory of Computing.
Preprocessing versus Solving

Numerical algebraic geometry for MLE has big advantages over symbolic computation. After solving the equations once, for one generic $U_0$, all subsequent computations for other $U$ are very fast.

Homotopy continuation starts from the critical points of $\ell_{U_0}$ and transforms them into the critical points of $\ell_U$.

Geometrically speaking, for a fixed statistical model $\mathcal{M}$, the homotopy amounts to changing the data.

Here are times (in seconds) needed to preprocess our equations and times needed to solve subsequent instances, using Bertini on a 64-bit Linux cluster with 160 processors:

<table>
<thead>
<tr>
<th>$(m, n, r)$</th>
<th>$(4, 4, 2)$</th>
<th>$(4, 4, 3)$</th>
<th>$(4, 5, 2)$</th>
<th>$(4, 5, 3)$</th>
<th>$(5, 5, 2)$</th>
<th>$(5, 5, 4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>257</td>
<td>427</td>
<td>1938</td>
<td>2902</td>
<td>348555</td>
<td>146952</td>
</tr>
<tr>
<td>Solving</td>
<td>4</td>
<td>4</td>
<td>20</td>
<td>20</td>
<td>83</td>
<td>83</td>
</tr>
</tbody>
</table>
What Statisticians Really Do

The *mixture model* \( \text{Mix}_r \) is the set of \( m \times n \) matrices

\[
P = A \cdot \Lambda \cdot B,
\]

where \( A \) is a non-negative \( m \times r \)-matrix whose rows sum to 1, \( \Lambda \) is a non-negative \( r \times r \) diagonal matrix whose diagonal sums to 1, and \( B \) is a non-negative \( r \times n \)-matrix whose columns sum to 1.

These \( P \) are the matrices in \( \Delta_{mn-1} \) of *non-negative rank* at most \( r \).

**Proposition**

*Our rank-constrained model* \( \mathcal{M}_r \) *is the Zariski closure of the mixture model* \( \text{Mix}_r \) *inside the simplex* \( \Delta_{mn-1} \).

*If* \( r \leq 2 \) *then* \( \text{Mix}_r = \mathcal{M}_r \). *If* \( r \geq 3 \) *then* \( \text{Mix}_r \subsetneq \mathcal{M}_r \).

Statisticians seek to maximize likelihood function \( \ell_U \) on \( \text{Mix}_r \) by running the *expectation-maximization* (EM) algorithm in the space \( (\Delta_{m-1})^r \times \Delta_{r-1} \times (\Delta_{n-1})^r \) of parameters \( (A, \Lambda, B) \).
Further Reading


