Invariant theory and scaling algorithms for maximum likelihood estimation

Kathlén Kohn
KTH Stockholm

joint with

Carlos Améndola
TU Munich

Philipp Reichenbach
TU Berlin

Anna Seigal
University of Oxford

September 10, 2020
Given: statistical model
    sample data $S_Y$
Task: find maximum likelihood estimate (MLE)
    = point in model that best fits $S_Y$

Given: orbit $G \cdot v = \{ g \cdot v \mid g \in G \}$
Task: compute capacity
    = closest distance of orbit to origin
Invariant theory

Stability notions

The orbit of a vector \( v \) in a vector space \( V \) under an action by a group \( G \) is

\[
G \cdot v = \{ g \cdot v \mid g \in G \} \subset V.
\]

- \( v \) is \textbf{unstable} iff \( 0 \in \overline{G \cdot v} \) (i.e. \( v \) can be scaled to 0 in the limit)
- \( v \) \textbf{semistable} iff \( 0 \notin \overline{G \cdot v} \)
- \( v \) \textbf{polystable} iff \( v \neq 0 \) and its orbit \( G \cdot v \) is closed
- \( v \) \textbf{is stable} iff \( v \) is polystable and its stabilizer is finite

The null cone of the action by \( G \) is the set of unstable vectors \( v \).
Invariant theory

Null cone membership testing

Classical and often hard question: Describe null cone
(essentially equivalent to finding generators for the ring of polynomial invariants)

Modern approach: Provide a test to determine if a vector \( v \) lies in null cone

The **capacity** of \( v \) is

\[
\text{cap}_G(v) := \inf_{g \in G} \|g \cdot v\|_2^2.
\]

**Observation:** \( \text{cap}_G(v) = 0 \) iff \( v \) lies in null cone

Hence: Testing null cone membership is a minimization problem.

\( \rightsquigarrow \) algorithms: [series of 3 papers in 2017 – 2019 by Bürgisser, Franks, Garg, Oliveira, Walter, Wigderson]
Maximum likelihood estimation

Given:

- $\mathcal{M}$: a statistical **model** = a set of probability distributions
- $Y = (Y_1, \ldots, Y_n)$: $n$ samples of observed **data**

Goal: find a distribution in the model $\mathcal{M}$ that best fits the empirical data $Y$

**Approach:** maximize the **likelihood function**

$$L_Y(\rho) := \rho(Y_1) \cdots \rho(Y_n), \quad \text{where } \rho \in \mathcal{M}.$$

A **maximum likelihood estimate (MLE)** is a distribution in the model $\mathcal{M}$ that maximizes the likelihood $L_Y$. 
Discrete statistical models

A probability distribution on \( m \) states is determined by its probability mass function \( \rho \), where \( \rho_j \) is the probability that the \( j \)-th state occurs.

\( \rho \) is a point in the probability simplex

\[
\Delta_{m-1} = \{ q \in \mathbb{R}^m \mid q_j \geq 0 \text{ and } \sum q_j = 1 \}.
\]

A discrete statistical model \( M \) is a subset of the simplex \( \Delta_{m-1} \).
Discrete statistical models

maximum likelihood estimation

Given data is a vector of counts $Y \in \mathbb{Z}^m_{\geq 0}$, where $Y_j$ is the number of times the $j$-th state occurs.

The empirical distribution is $S_Y = \frac{1}{n} Y \in \Delta_{m-1}$, where $n = Y_1 + \ldots + Y_m$.

The likelihood function takes the form $L_Y(\rho) = \rho_1^{Y_1} \cdots \rho_m^{Y_m}$, where $\rho \in \mathcal{M}$.

An MLE is a point in model $\mathcal{M}$ that maximizes the likelihood $L_Y$ of observing $Y$. 
Log-linear models

= set of distributions whose logarithms lie in a fixed linear space.

Let $A \in \mathbb{Z}^{d \times m}$, and define

$$
\mathcal{M}_A = \{ \rho \in \Delta_{m-1} \mid \log \rho \in \text{rowspan}(A) \}.
$$

We assume that $1 := (1, \ldots, 1) \in \text{rowspan}(A)$ (i.e., uniform distribution in $\mathcal{M}_A$).

Matrix $A = [a_1 | a_2 | \ldots | a_m]$ also defines an action by the torus $(\mathbb{C}^\times)^d$ on $\mathbb{C}^m$:

$$
g \in (\mathbb{C}^\times)^d \text{ acts on } x \in \mathbb{C}^m \text{ by left multiplication with } \begin{bmatrix} g^{a_1} & & \\ & \ddots & \\ & & g^{a_m} \end{bmatrix}, \quad \text{where } g^{a_j} = g_1^{a_{1j}} \ldots g_d^{a_{dj}}.
$$

$\mathcal{M}_A$ is the orbit of the uniform distribution in $\Delta_{m-1} \cap \mathbb{R}_{>0}^m$. 

VII - XIV
Example

\[ M_A = \{ \rho \in \Delta_{m-1} \mid \log \rho \in \text{rowspan}(A) \} \]

\[ A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \]

\[ g \in (\mathbb{C}^\times)^2 \text{ acts on } x \in \mathbb{C}^3 \text{ by } \begin{bmatrix} g^{a_1} \\ g^{a_2} \\ g^{a_3} \end{bmatrix} = \begin{bmatrix} g_1^2 \\ g_1 g_2 \\ g_2^2 \end{bmatrix}. \]

\[ M_A = \left( (\mathbb{C}^\times)^2 \cdot \frac{1}{3} \mathbb{I} \right) \cap \Delta_2 \cap \mathbb{R}^3_{>0} \]

\[ = \left\{ \frac{1}{3} (g_1^2, g_1 g_2, g_2^2) \mid g_1, g_2 > 0, \ g_1^2 + g_1 g_2 + g_2^2 = 3 \right\} \]

\[ = \{ \rho \in \mathbb{R}^3_{>0} \mid \rho_2^2 = \rho_1 \rho_3, \ \rho_1 + \rho_2 + \rho_3 = 1 \} \]

other examples: independence model, graphical models, hierarchical models, ...
Combining both worlds

**Theorem** (Améndola, Kohn, Reichenbach, Seigal)
Let \( A = [a_1|...|a_m] \in \mathbb{Z}^{d \times m} \) and \( Y \in \mathbb{Z}^m \) be a vector of counts with \( n = \sum Y_j \).

MLE given \( Y \) exists in \( M_A \) \iff \( 1 \in \mathbb{C}^m \) is polystable under the action of \((\mathbb{C}^\times)^d\) given by the matrix \([na_1 - AY|...|na_m - AY]\) attains its maximum \iff \( x \) attains its minimum

How are the two optimal points related?

**Theorem** (cont’d)
If \( x \in \mathbb{C}^m \) is a point of minimal norm in the orbit \((\mathbb{C}^\times)^d \cdot 1\), then the MLE is \( x^{(2)} = \frac{x^{(2)}}{\|x\|^2} \), where \( x^{(2)} \) is the vector with \( j \)-th entry \( |x_j|^2 \).
Algorithmic consequences

algorithms for finding MLE, e.g. iterative proportional scaling (IPS)

⇔ scaling algorithms to compute capacity

maximize likelihood ⇔ minimize KL divergence

model lives in $\Delta_{m-1} \cap \mathbb{R}_{>0}^m$

minimize $\ell_2$-norm

orbit lives in $\mathbb{C}^m$
Gaussian statistical models

The density function of an \( m \)-dimensional Gaussian with mean zero and covariance matrix \( \Sigma \in \mathbb{R}^{m \times m} \) is

\[
\rho_{\Sigma}(y) = \frac{1}{\sqrt{\det(2\pi \Sigma)}} \exp \left( -\frac{1}{2} y^T \Sigma^{-1} y \right), \quad \text{where } y \in \mathbb{R}^m.
\]

The concentration matrix \( \Psi = \Sigma^{-1} \) is symmetric and positive definite. A Gaussian model \( \mathcal{M} \) is a set of concentration matrices, i.e. a subset of the cone of \( m \times m \) symmetric positive definite matrices.

Given data \( Y = (Y_1, \ldots, Y_n) \), the likelihood is

\[
L_Y(\Psi) = \rho_{\Psi^{-1}}(Y_1) \cdots \rho_{\Psi^{-1}}(Y_n), \quad \text{where } \Psi \in \mathcal{M}.
\]

The likelihood \( L_Y \) can be unbounded from above. MLE might not exist. MLE might not be unique.
Combining both worlds

Invariant theory classically over $\mathbb{C}$ – can also define Gaussian models over $\mathbb{C}$.

The **Gaussian group model** of a group $G \subset \text{GL}_m(\mathbb{C})$ is $\mathcal{M}_G := \{g^* g \mid g \in G\}$.

**Theorem** (Améndola, Kohn, Reichenbach, Seigal)

Let $Y = (Y_1, \ldots, Y_n)$ with $Y_i \in \mathbb{C}^m$ and $G \subset \text{GL}_m(\mathbb{C})$ be a group closed under non-zero scalar multiples (i.e., $g \in G, \lambda \in \mathbb{C}, \lambda \neq 0 \Rightarrow \lambda g \in G$).

If $G$ is linearly reductive, ML estimation for $\mathcal{M}_G$ relates to the action by $G \cap \text{SL}_m(\mathbb{C})$ as follows:

(a) $Y$ unstable $\iff$ $L_Y$ not bounded from above
(b) $Y$ semistable $\iff$ $L_Y$ bounded from above
(c) $Y$ polystable $\iff$ MLE exists
(d) $Y$ stable $\iff$ finitely many MLEs exist $\iff$ unique MLE

![Diagram](image)
Combining both worlds

Real examples

**Theorem** (Améndola, Kohn, Reichenbach, Seigal)

Let \( Y = (Y_1, \ldots, Y_n) \) with \( Y_i \in \mathbb{R}^m \), and let \( G \subseteq \text{GL}_m(\mathbb{R}) \) be a linearly reductive group which is closed under non-zero scalar multiples.

ML estimation for \( \mathcal{M}_G \) relates to the action by \( G \cap \text{SL}_m(\mathbb{R}) \) as follows:

(a) \( Y \) unstable \( \iff \ell_Y \) not bounded from above
(b) \( Y \) semistable \( \iff \ell_Y \) bounded from above
(c) \( Y \) polystable \( \iff \) MLE exists
(d) \( Y \) stable \( \implies \) finitely many MLEs exist \( \iff \) unique MLE

Examples: full Gaussian model, independence model, matrix normal model

---

Summary

**Invariant theory**
- describe null cone
- algorithmic null cone
- membership testing

**Statistics**
- algorithms to find MLE
- convergence analysis

historical progression