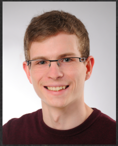
Invariant theory and scaling algorithms for maximum likelihood estimation

Kathlén Kohn
KTH Stockholm

joint with

Carlos Améndola
TU Munich

Philipp Reichenbach
TU Berlin

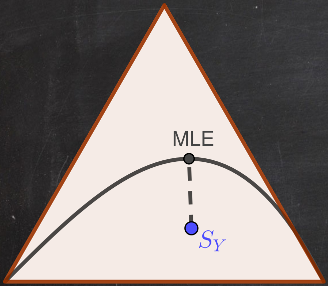


Anna Seigal
University of Oxford

September 10, 2020

Global picture

Statistics

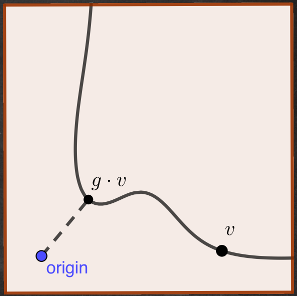


Given: statistical model
sample data S_Y

Task: find **maximum likelihood estimate (MLE)**

= point in model that best fits S_Y

Invariant theory



Given: orbit $G \cdot v = \{g \cdot v \mid g \in G\}$

Task: compute **capacity**

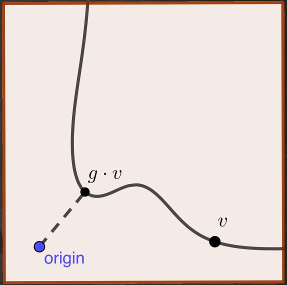
= closest distance of orbit to origin

Invariant theory

Stability notions

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

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



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

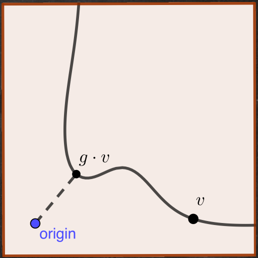
Null cone membership testing

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

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



Observation: $\mathrm{cap}_G(v) = 0$ 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

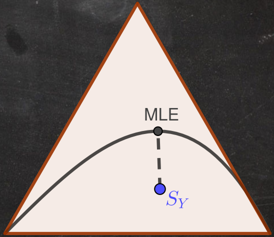
Given:

- $\mathscr{M}$: 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 $\mathscr{M}$ that best fits the empirical data Y

Approach: maximize the **likelihood function**

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



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

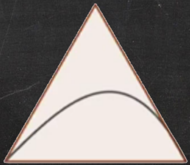
Discrete statistical models

A probability distribution on m states is determined by is **probability mass function** ρ, where ρ_j is the probability that the j-th state occurs.

ρ is a point in the **probability simplex**

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

A **discrete statistical model** $\mathscr{M}$ is a subset of the simplex Δ_{m-1}.



Discrete statistical models

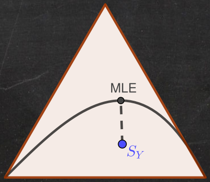
maximum likelihood estimation

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 $\rho \in \mathscr{M}$.

An **MLE** is a point in model $\mathscr{M}$ 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

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

We assume that $\mathbb{1} := (1,\dots,1) \in \mathrm{rowspan}(A)$ (i.e., uniform distribution in $\mathscr{M}_A$).

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

$g \in (\mathbb{C}^\times)^d$ acts on $x \in \mathbb{C}^m$ by left multiplication with

$$\begin{bmatrix} g^{a_1} & & \\ & \ddots & \\ & & g^{a_m} \end{bmatrix}, \quad \text{where } g^{a_j} = g_1^{a_{1j}} \dots g_d^{a_{dj}}.$$

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

Example

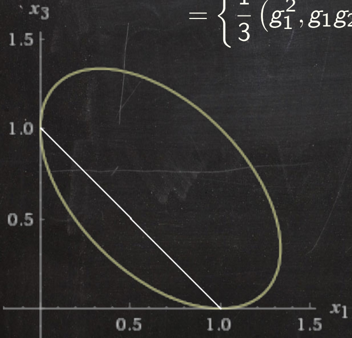
$$\mathscr{M}_A = \{\rho \in \Delta_{m-1} \mid \log \rho \in \mathrm{rowspan}(A)\}. \qquad A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

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

$$\mathscr{M}_A = ((\mathbb{C}^\times)^2 \cdot \frac{1}{3}\mathbb{1}) \cap \Delta_2 \cap \mathbb{R}^3_{>0}$$

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

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

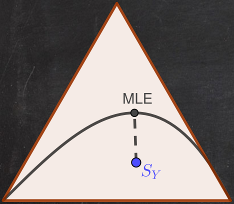


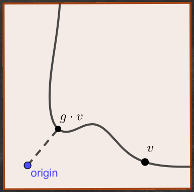
other examples: independence model, graphical models, hierarchical models, . . .

Combining both worlds

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

MLE given Y exists in $\mathscr{M}_A$ $\Leftrightarrow$ $\mathbb{1} \in \mathbb{C}^m$ is polystable under the action of $(\mathbb{C}^\times)^d$ given by the matrix $[na_1 - AY|\dots|na_m - AY]$





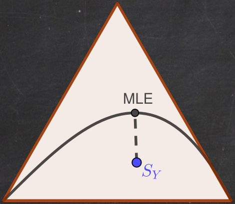
attains its maximum $\Leftrightarrow$ attains its minimum

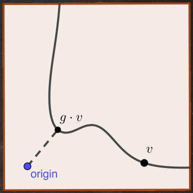
How are the two optimal points related?

Theorem (cont'd)
If $x \in \mathbb{C}^m$ is a point of minimal norm in the orbit $(\mathbb{C}^\times)^d \cdot \mathbb{1}$, then the MLE is

$$\frac{x^{(2)}}{\|x\|^2}, \qquad \text{where } x^{(2)} \text{ is the vector with } j\text{-th entry } |x_j|^2.$$

Algorithmic consequences





algorithms for finding MLE, e.g. iterative proportional scaling (IPS) $\leftrightarrow$ scaling algorithms to compute capacity

maximize likelihood $\Leftrightarrow$ minimize KL divergence

minimize ℓ_2-norm

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

orbit lives in $\mathbb{C}^m$

Gaussian statistical models

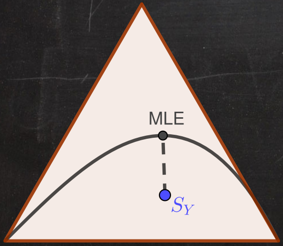
The density function of an m-dimensional Gaussian with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$ is

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

The **concentration matrix** $\Psi = \Sigma^{-1}$ is symmetric and positive definite.
A **Gaussian model** $\mathscr{M}$ 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

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



likelihood L_Y can be unbounded from above

MLE might not exist

MLE might not be unique

Combining both worlds

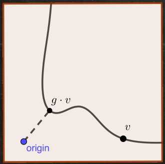
Invariant theory classically over $\mathbb{C}$ – can also define Gaussian models over $\mathbb{C}$

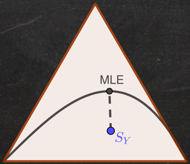
The **Gaussian group model** of a group $G \subset \mathrm{GL}_m(\mathbb{C})$ is $\mathscr{M}_G := \{g^* g \mid g \in G\}$.

Theorem (Améndola, Kohn, Reichenbach, Seigal)
Let $Y = (Y_1, \ldots, Y_n)$ with $Y_i \in \mathbb{C}^m$ and $G \subset \mathrm{GL}_m(\mathbb{C})$ 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 $\mathscr{M}_G$ relates to the action by $G \cap \mathrm{SL}_m(\mathbb{C})$ as follows:

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





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 $G \subset \mathrm{GL}_m(\mathbb{R})$ be a linearly reductive group which is closed under non-zero scalar multiples.
ML estimation for $\mathcal{M}_G$ relates to the action by $G \cap \mathrm{SL}_m(\mathbb{R})$ as follows:

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

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

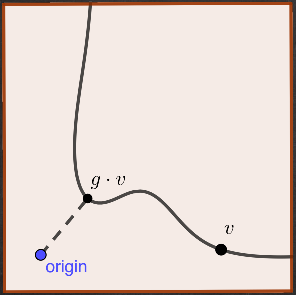
Harm Derksen, Visu Makam: computed ML thresholds using our dictionary! (arXiv:2007.10206)

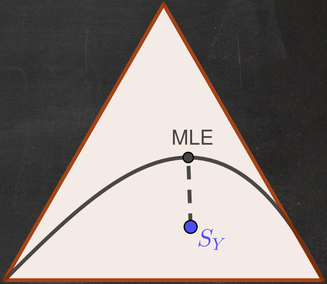
Theorem (Améndola, Kohn, Reichenbach, Seigal)
Let $Y = (Y_1, \ldots, Y_n)$ with $Y_i \in \mathbb{R}^m$, and let $G \subset \mathrm{GL}_m(\mathbb{R})$ be a group which is closed under non-zero scalar multiples, but not necessarily linearly reductive.
ML estimation for $\mathcal{M}_G$ relates to the action by $G \cap \mathrm{SL}^{\pm}_m(\mathbb{R})$ as follows:

(a)	Y unstable	$\Leftrightarrow$	ℓ_Y not bounded from above
(b)	Y semistable	$\Leftrightarrow$	ℓ_Y bounded from above
(c)	Y polystable	$\Rightarrow$	MLE exists

Example: **Gaussian graphical models**

Summary





	Invariant theory		**Statistics**
historical progression ↓	describe null cone	×	algorithms to find MLE
	algorithmic null cone membership testing		convergence analysis