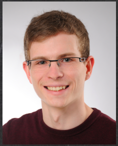
Invariant Theory for Maximum Likelihood Estimation

Carlos Améndola & Kathlén Kohn

joint with

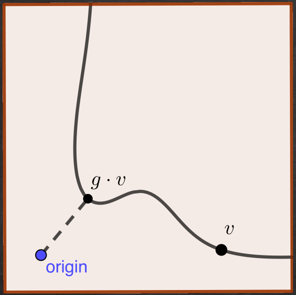
Philipp Reichenbach
TU Berlin

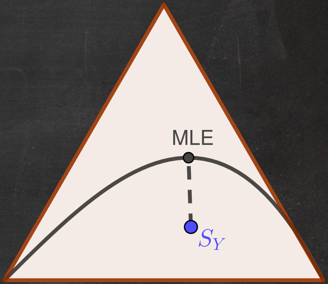
Anna Seigal
University of Oxford



May 6, 2020

Global picture





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

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

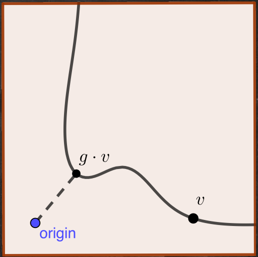
Null cone membership testing

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

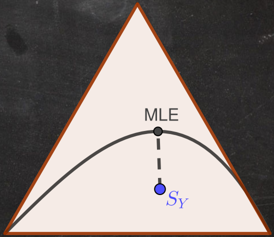
Given:

- $\mathcal{M}$: 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**

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



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

Maximum likelihood estimation

Gaussian models

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

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

The **concentration matrix** $\Psi = \Sigma^{-1}$ is positive definite.
A **Gaussian model** $\mathcal{M}$ is a set of concentration matrices, i.e. a subset of the cone of $m \times m$ positive definite matrices. Given data $Y = (Y_1, \ldots, Y_n)$, the likelihood is

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

The **Gaussian group model** of a group G with a representation $G \xrightarrow{\varphi} \mathrm{GL}_m$ on $\mathbb{R}^m$ is

$$\mathcal{M}_G := \left\{ \Psi_g = \varphi(g)^T \varphi(g) \mid g \in G \right\}.$$

We want to find an MLE, i.e. a maximizer of

$$\log L_Y(\Psi_g) = \frac{1}{2} \underbrace{\left(n \log\det \Psi_g - \|g \cdot Y\|_2^2\right)}_{\ell_Y(\Psi_g)} - \frac{nm}{2} \log(2\pi) \qquad \text{for } g \in G.$$

Combining both worlds

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

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

$$\sup_{g \in G} \ell_Y(\Psi_g) = - \inf_{\tau \in \mathbb{R}_{>0}} \left(\tau \left(\inf_{h \in G \cap \mathrm{SL_m}} \|h \cdot Y\|_2^2 \right) - nm \log \tau \right).$$

The MLEs for the Gaussian group model $\mathcal{M}_G$, if they exist, are the matrices $\tau h^* h$, where $h \in G \cap \mathrm{SL}_m(\mathbb{C})$ s.t. $\|h \cdot Y\|_2^2 = \mathrm{cap}_{G \cap \mathrm{SL}}(Y)$, and

$\tau \in \mathbb{R}_{>0}$ is the unique value minimizing $\tau \, \mathrm{cap}_{G \cap \mathrm{SL}}(Y) - nm \log \tau$.

Theorem (Améndola, Kohn, Reichenbach, Seigal)
Let Y and G as above. If G is linearly reductive,
ML estimation for $\mathcal{M}_G$ 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 $Y = (Y_1, \dots, Y_n)$ 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

Theorem (Améndola, Kohn, Reichenbach, Seigal)
Let $Y = (Y_1, \dots, Y_n)$ 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 $G \cap \mathrm{SL}_m^{\pm}(\mathbb{R})$ 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 model defined by transitive DAG

Gaussian graphical models

Directed acyclic graphs

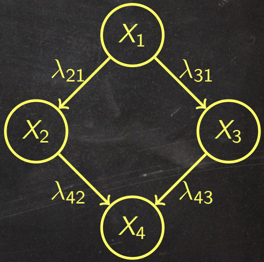
Important family of statistical models that represent interaction structures between several random variables.

- Consider a directed acyclic graph (DAG) $\mathcal{G}$ with m nodes.
- Each node j represents a random variable X_j (e.g., Gaussian).
- Each edge $j \to i$ encodes (conditional) dependence: X_j 'causes' X_i.
- The parents of i are $\mathrm{pa}(i) = \{j \mid j \to i\}$.

The model is defined by the recursive linear equation:

$$X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ij} X_j + \varepsilon_i$$

where λ_{ij} is the edge coefficient and ε_i is Gaussian error.



It can be written as $\boldsymbol{X} = \boldsymbol{\Lambda X} + \boldsymbol{\varepsilon}$ where $\Lambda \in \mathbb{R}^{m \times m}$ satisfies $\lambda_{ij} = 0$ for $j \not\to i$ in $\mathcal{G}$ and $\varepsilon \sim N(0, \Omega)$ with Ω diagonal, positive definite.

Gaussian graphical models

coming from groups

From $X = \Lambda X + \varepsilon$, we rewrite

$$X = (I - \Lambda)^{-1}\varepsilon$$

so that $X \sim N(0, \Sigma)$ with

$$\Sigma = (I - \Lambda)^{-1}\Omega(I - \Lambda)^{-T} \quad \& \quad \Psi = (I - \Lambda)^{T}\Omega^{-1}(I - \Lambda).$$

The **Gaussian graphical model** $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ consists of concentration matrices Ψ of this form. Consider the set

$$G(\mathcal{G}) = \{g \in \mathrm{GL}_m \mid g_{ij} = 0 \text{ for } i \neq j \text{ with } j \not\to i \text{ in } \mathcal{G}\}.$$

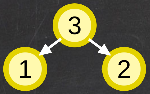
Proposition
The set of matrices $G(\mathcal{G})$ is a group if and only if $\mathcal{G}$ is a **transitive** directed acyclic graph (TDAG), i.e., $k \to j$ and $j \to i$ in $\mathcal{G}$ imply $k \to i$. In this case,

$$\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_{G(\mathcal{G})}.$$

TDAG group models

Example

Let $\mathcal{G}$ be the TDAG



.

The corresponding group $G(\mathcal{G}) \subseteq \mathrm{GL}_3$ consists of invertible matrices g of the form

$$g = \begin{bmatrix} * & 0 & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix}.$$

The Gaussian graphical model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is a 5-dimensional linear subspace of the cone of symmetric positive definite 3×3 matrices:

$$\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \{g^T g \mid g \in G(\mathcal{G})\} = \{\Psi \in \mathrm{PD}_3 \mid \psi_{12} = \psi_{21} = 0\}.$$

Note that $G(\mathcal{G})$ is not reductive!

The MLE is known to be unique if it exists. So when does it exist?

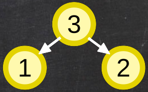
Null cone of TDAGs

Theorem (Améndola, Kohn, Reichenbach, Seigal)
Let $Y \in \mathbb{R}^{m \times n}$ be a tuple of n samples. If some row of Y corresponding to vertex i is in the linear span of the rows corresponding to the parents of i,

- then Y is unstable under the action by $G(\mathcal{G}) \cap \mathrm{SL}_m$, i.e. the likelihood is unbounded;
- otherwise, Y is polystable, i.e. the MLE exists.

(by our main theorem in the real non-reductive case)

Example Let $n = 2$ in



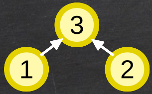
and consider three different pairs of samples:

$$Y^1 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad Y^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 4 \end{pmatrix}, \quad Y^3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 2 \end{pmatrix}.$$

Using the theorem, we see that Y^1 and Y^2 are unstable and Y^3 is polystable.
The null cone has two components: $\langle y_{11}y_{32} - y_{12}y_{31} \rangle \cap \langle y_{21}y_{32} - y_{22}y_{31} \rangle$.

Null cones of TDAGs

Corollary Let $\mathcal{G}$ be a TDAG with m nodes and n samples.
Each irreducible component of the Zariski closure of the null cone under the action of $G(\mathcal{G}) \cap \mathrm{SL}_m$ on $\mathbb{R}^{m\times n}$ is defined by the maximal minors of the submatrix whose rows are a childless node and its parents.

Example
Let $\mathcal{G}$ be the TDAG 

.

- The null cone is **not** Zariski closed for $n \geq 2$.
  Its Zariski closure is the variety of matrices of rank at most two.
- For $n = 2$, Y is not in the null cone but in its Zariski closure ($= \mathbb{R}^{3\times 2}$):

$$Y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

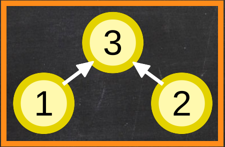
  Hence, the MLE given Y exists. What is it?

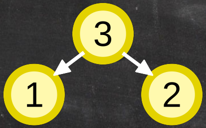
  Y is of minimal norm in its orbit, so the MLE given Y is λI_3, where λ minimizes $\frac{3}{2}\lambda - 3\log(\lambda)$. Hence $\lambda = 2$.

Undirected Graphical Models

Which TDAGs have Zariski closed null cones?

Corollary Let $\mathcal{G}$ be a TDAG with m nodes. The null cone under the action of $G(\mathcal{G}) \cap \mathrm{SL}_m$ on $\mathbb{R}^{m \times n}$ is Zariski closed for every n iff $\mathcal{G}$ has no unshielded colliders.



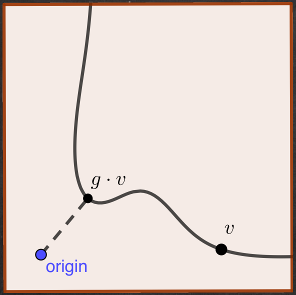


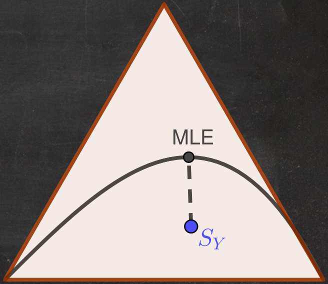
An **unshielded collider** of $\mathcal{G}$ is a subgraph $j \to i \leftarrow k$ with no edge between j and k.

This is a very interesting condition in statistics! $\mathcal{G}$ has no unshielded colliders if and only if it has the same graphical model as its underlying undirected graph.

Summary

Invariant Theory and Scaling Algorithms for Maximum Likelihood Estimation
arXiv:2003.13662





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