SIAM REVIEW
Vol. 66, No. 4, pp. 721–747

© 2024 SIAM. Published by SIAM under the terms
of the Creative Commons 4.0 license

A Bridge between Invariant Theory and Maximum Likelihood Estimation*

Carlos Améndola[†]
Kathlén Kohn[‡]
Philipp Reichenbach[†]
Anna Seigal[§]

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

Abstract. We uncover connections between maximum likelihood estimation in statistics and norm minimization over a group orbit in invariant theory. We present a dictionary that relates notions of stability from geometric invariant theory to the existence and uniqueness of a maximum likelihood estimate. Our dictionary holds for both discrete and continuous statistical models: we discuss log-linear models and Gaussian models, including matrix normal models and directed Gaussian graphical models. Our approach reveals promising consequences of the interplay between invariant theory and statistics. For instance, algorithms from statistics can be used in invariant theory, and vice versa.

Key words. maximum likelihood estimation, group actions, Gaussian models, graphical models, log-linear models

MSC codes. 14L24, 14P05, 20G45, 62F10, 62H22, 62R01

DOI. 10.1137/24M1661753

Contents

*Published electronically November 7, 2024. This paper originally appeared in *SIAM Journal on Applied Algebra and Geometry*, Volume 5, Number 2, 2021, pages 304–337, under the title “Invariant Theory and Scaling Algorithms for Maximum Likelihood Estimation.”
https://doi.org/10.1137/24M1661753

Funding: The work of the second author was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The work of the third author was funded by the European Research Council (ERC) under the European Horizon 2020 research and innovation programme grant 787840. The work of the fourth author was partially supported by NSF grant DMR-2011754.

[†] Institut für Mathematik, TU Berlin, Berlin, 10623 Germany (amendola@math.tu-berlin.de, ph-reichenbach@t-online.de).

[‡] Department of Mathematics, KTH, Stockholm, 10044 Sweden (kathlen@kth.se).

[§] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (aseigal@seas.harvard.edu).

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

1. Introduction.

Fitting a model to data is fundamental in statistics. A widespread approach is to maximize the likelihood of observing the data as we range over the model. A point that maximizes the likelihood is called a *maximum likelihood estimate* (MLE). There are close connections between statistical models and group actions, dating back to the translation and scaling actions considered by Fisher [21]. In this paper, we develop such connections to build a bridge between invariant theory and maximum likelihood (ML) estimation.

Invariant theory studies actions of groups. An important concept is the orbit of a point, the set of points that differ from the original point by a transformation in the group. In this paper, we establish connections between minimizing the norm over an orbit and computing the MLE; see Figure 1.

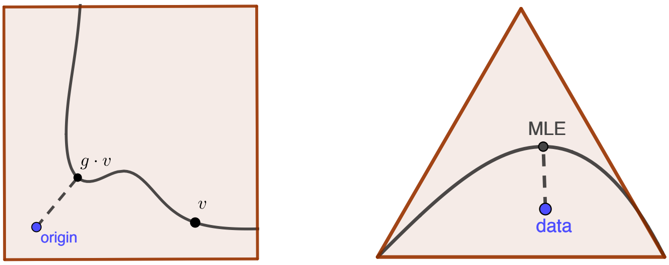

Fig. 1 *Sketch of minimizing the norm over a group orbit (left) and maximizing the likelihood over a statistical model (right). We connect these two pictures.*

The main invariant theory tool we use to establish the connection in Figure 1 is the Kempf–Ness theorem [30]. It can be viewed as an optimization duality result,

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

which gives information on the optimization landscapes of norm minimization over a group orbit. The Kempf–Ness theorem gives a first order criterion (the vanishing of the moment map) for a point being of minimal norm in an orbit. It also describes the set of all points of minimal norm. The Kempf–Ness theorem has been applied in a number of contexts, including quantum information theory, complexity theory, and analytic inequalities [10]. Here, we use it for ML estimation.

We relate notions of stability from invariant theory to the existence and uniqueness of an MLE. The *capacity* of a point is the infimal norm along its orbit. If the orbit is closed, the capacity is attained; otherwise the capacity is attained only on the orbit closure. Points with zero capacity are called *unstable*; they form the null cone, a classical object in invariant theory dating back to Hilbert [26], which is of particular interest for moduli spaces of algebraic objects. Points that are not unstable are called *semistable*, with *polystable* and *stable* points being two successive specializations of semistable points.

Main Contributions. An MLE is usually computed via optimization approaches that find a local maximum [34, 38]. There is growing interest in understanding when such algorithms are guaranteed to work, and under which conditions an MLE exists or is unique [42]. For several classes of statistical model, we show that finding an MLE can be cast as a norm minimization problem. When the model is parametrized by a group, this is a capacity problem, minimizing the norm over a group orbit. This connection allows us to build a dictionary between notions of stability from invariant theory and MLE properties:

$$\left\{\begin{matrix}\text{unstable}\\ \text{semistable}\\ \text{polystable}\\ \text{stable}\end{matrix}\right\} \qquad \longleftrightarrow \qquad \left\{\begin{matrix}\text{likelihood unbounded from above}\\ \text{likelihood bounded from above}\\ \text{MLE exists}\\ \text{MLE exists uniquely}\end{matrix}\right\}$$

For some statistical models, there is exact equivalence between the four notions of stability on the left and the four properties of ML estimation on the right. We call this the *full correspondence*. For other models, we obtain partial correspondences.

This bridge between invariant theory and ML estimation has enabled invariant theorists to answer questions from statistics. For instance, the paper [13] computes a formula for the number of data samples needed such that the MLE exists and is unique almost surely in a matrix normal model, a generalization of standard multivariate normal distribution to matrix-valued random variables. The result was extended to tensor normal models [14].

Our Statistical Models. A statistical model is a set of probability distributions. Statistical models that exhibit symmetry, or on which a group acts, are of longstanding statistical interest; see [4, 6, 15, 20, 33]. In this paper, we consider both continuous and discrete statistical models.

On the continuous side, we study *Gaussian models*, multivariate normal distributions, including matrix normal models, tensor normal models, and Gaussian graphical models. We study these within the framework of Gaussian group models. Our full correspondence holds for complex Gaussian group models on reductive groups (Theorem 5.3) and directed Gaussian graphical models (Theorem 6.5). For other models, the correspondence depends on whether the group is reductive or nonreductive. While invariant theory traditionally focuses on reductive groups, Gaussian group models are natural for both reductive and nonreductive groups. For nonreductive groups we have

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

a *weak correspondence* (Theorem 2.12) and for reductive groups a *strong correspondence* holds (Theorem 3.3).

On the discrete side, we study *log-linear models.* These play a fundamental role in categorical data analysis and include, for example, independence models and discrete graphical models. Our correspondence for these models is Theorem 7.4.

Algorithmic Consequences. Our bridge between invariant theory and ML estimation shows that one can compute the MLE from a vector of minimal norm in an orbit (Theorems 2.8 and 7.6). As a consequence, algorithms in invariant theory can be used in ML estimation, and vice versa. In statistics, many iterative algorithms for finding the MLE are well known, e.g., [19]. A more recent question is to understand when they converge, i.e., when an MLE exists, and when convergence is to a unique solution, i.e., when the MLE is unique [42]. The historical progression is the opposite in invariant theory: the distinction between different types of stability is classical. More recently, algorithmic approaches to stability questions have been taken, with a focus on null cone membership [1, 10, 11, 12, 23, 29].

Organization. We study multivariate Gaussian models whose MLE can be obtained via norm minimization in section 2. We define *Gaussian group models* in section 3. These are multivariate Gaussian models whose concentration matrices are of the form $g^\mathsf{T} g$, where g lies in a group. We discuss the special case of matrix normal models in section 4. We recall the Kempf–Ness theorem and use it to prove our main results in section 5. We study Gaussian models on directed acyclic graphs in section 6; such models are Gaussian group models when the underlying graph is transitive. We study log-linear models in section 7 and give an outlook in section 8.

2. Gaussian Models.

We recap ML estimation for Gaussian models in section 2.1 and explain how it can be viewed as a norm minimization problem in section 2.2. We describe stability notions from invariant theory and their interplay with ML estimation in section 2.3.

2.1. Maximum Likelihood Estimation.

We consider m-dimensional multivariate Gaussian distributions with mean zero. Their density functions have the form

$$f_\Sigma \colon \mathbb{R}^m \to \mathbb{R}, \quad f_\Sigma(y) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} y^\mathsf{T} \Sigma^{-1} y\right),$$

where Σ is the covariance matrix, which lives in the cone of $m \times m$ symmetric positive definite matrices, denoted by PD_m. We parametrize our models via the *concentration matrix* $\Psi = \Sigma^{-1}$ and make the following definition.

Definition 2.1. A *Gaussian model* is a set of *concentration matrices* $\mathcal{M} \subseteq \mathrm{PD}_m$.

We now explain parameter estimation. We observe a tuple $Y = (Y_1, \ldots, Y_n) \in (\mathbb{R}^m)^n$ of n samples that are independent and identically distributed (i.i.d.). The *likelihood function* given data Y is

$$L_Y \colon \mathcal{M} \to \mathbb{R}_{\geq 0}, \quad L_Y(\Psi) = \prod_{i=1}^{n} f_{\Psi^{-1}}(Y_i).$$

It measures how likely it is to observe Y under the distribution given by $f_{\Psi^{-1}}$. Maximum likelihood estimation seeks a point in the model $\mathcal{M}$ that maximizes L_Y. Such a point $\hat{\Psi}$ is called a *maximum likelihood estimate* (MLE) given Y.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

It is convenient to work with the *log-likelihood function* $\ell_y = \log L_y$, which has the same maximizers as L_Y. The log-likelihood can be written, up to additive and multiplicative constants, as

$$\ell_Y(\Psi) = \log\det(\Psi) - \operatorname{tr}(\Psi S_Y), \tag{2.1}$$

where $S_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^{\mathsf{T}}$ is the sample covariance matrix, an $m \times m$ positive semidefinite matrix. Here we view Y_i as column vectors, which canonically writes Y as a matrix in $\mathbb{R}^{m\times n}$.

Example 2.2. For $\mathcal{M} = \mathrm{PD}_m$, it is well known that the unique maximizer of the likelihood is $\hat{\Psi} = S_Y^{-1}$, if S_Y is invertible. If S_Y is not invertible, the likelihood function is unbounded and the MLE does not exist; see [42, Proposition 5.3.7].

The following questions are central to ML estimation. We say that a property holds almost surely if it holds with probability one.

1. Is the likelihood bounded from above? Does an MLE given Y exist (uniquely)?
2. Can we provide an algorithm to compute an MLE?
3. Which sample size almost surely guarantees affirmative answers to the questions in item 1?

A key message of this paper is that, for several models, these three questions may be tackled via invariant theory. For the third question, one considers the following concepts.

Definition 2.3. *To a Gaussian model $\mathcal{M} \subseteq \mathrm{PD}_m$, we associate three* ML thresholds*:*

- $\mathrm{mlt_b}(\mathcal{M})$ *is the minimum number of samples n needed for the likelihood to be bounded almost surely; i.e., for almost all $Y \in (\mathbb{R}^m)^n$.*
- $\mathrm{mlt_e}(\mathcal{M})$ *is the minimum number of samples needed for an MLE to exist almost surely.*
- $\mathrm{mlt_u}(\mathcal{M})$ *is the minimum number of samples needed for almost sure existence of a unique MLE.*

A model $\mathcal{M}$ satisfies $\mathrm{mlt_b}(\mathcal{M}) \leq \mathrm{mlt_e}(\mathcal{M}) \leq \mathrm{mlt_u}(\mathcal{M})$. The study of ML thresholds is an active area [9, 16].

Example 2.4. The sample covariance matrix S_Y is invertible if and only if $Y \in \mathbb{R}^{m\times n}$ has full row rank. This holds almost surely if $m \leq n$ and cannot hold if $m > n$. Thus, we have $\mathrm{mlt_b}(\mathrm{PD}_m) = \mathrm{mlt_e}(\mathrm{PD}_m) = \mathrm{mlt_u}(\mathrm{PD}_m) = m$, by Example 2.2.

Remark 2.5. ML estimation for multivariate Gaussians with arbitrary mean can be translated to our mean zero setup; see [40, Remark 6.3.7]. This process shifts the ML thresholds by one.

2.2. MLE via Norm Minimization. We link ML estimation and norm minimization via the following parametrization of a Gaussian model. We denote the $m \times m$ real invertible matrices by $\mathrm{GL}_m(\mathbb{R})$.

Definition 2.6. *For a subset $E \subseteq \mathrm{GL}_m(\mathbb{R})$, we define*

$$\mathcal{M}_E := \left\{ e^{\mathsf{T}} e \mid e \in E \right\} \subseteq \mathrm{PD}_m. \tag{2.2}$$

Different sets E may give rise to the same model. For example, we have $\mathcal{M}_E = \mathrm{PD}_m$ whenever E contains all invertible upper triangular matrices. Many Gaussian models of interest are closed under positive scalar multiples. The next proposition follows from the Cholesky decomposition; see [40, Propositions 8.1.2 and 8.1.5].

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

PROPOSITION 2.7. *Let $\mathcal{M} \subseteq \mathrm{PD}_m$ be a Gaussian model. Then*

(i) *there exists a set $E \subseteq \mathrm{GL}_m(\mathbb{R})$ such that $\mathcal{M} = \mathcal{M}_E$;*

(ii) *$\mathcal{M}$ is closed under positive scalar multiples if and only if there is some set $E \subseteq \mathrm{GL}_m(\mathbb{R})$, closed under nonzero scalar multiples, with $\mathcal{M} = \mathcal{M}_E$.*

Our motivation for $\mathcal{M}_E$ is to link ML estimation to norm minimization via the following key observation. We can rewrite the log-likelihood (2.1) as

$$\ell_Y(e^\mathsf{T} e) = \log\det(e^\mathsf{T} e) - \frac{1}{n}\|e \cdot Y\|^2, \tag{2.3}$$

as follows. For $e \in \mathrm{GL}_m(\mathbb{R})$ and $Y = (Y_1, \ldots, Y_n) \in (\mathbb{R}^m)^n$, we have

$$\|e \cdot Y\|^2 = \sum_{i=1}^{n} (eY_i)^\mathsf{T}(eY_i) = \sum_{i=1}^{n} \operatorname{tr}\left((eY_i)^\mathsf{T}(eY_i)\right) = n \operatorname{tr}(e^\mathsf{T} e S_Y). \tag{2.4}$$

We always consider the standard Euclidean norm. For matrices, this is the Frobenius norm. We define

$$E_{\mathrm{SL}}^{\pm} := \{e \in E \mid \det(e) = \pm 1\} \qquad \text{and} \qquad E_{\mathrm{SL}} := \{e \in E \mid \det(e) = 1\}. \tag{2.5}$$

The following result shows that ML estimation is equivalent to minimizing the norm of $h \cdot Y$, where h ranges over $H := E_{\mathrm{SL}}^{\pm}$ or over $H := E_{\mathrm{SL}}$ under slightly stronger assumptions; see Remark 2.13. In invariant theory, this infimal norm squared is known as the *capacity* of Y under H:

$$\operatorname{cap}_H(Y) := \inf_{h \in H} \|h \cdot Y\|^2. \tag{2.6}$$

THEOREM 2.8 (MLE via norm minimization). *Let $E \subseteq \mathrm{GL}_m(\mathbb{R})$ be closed under nonzero scalar multiples and let $Y \in (\mathbb{R}^m)^n$. Then the supremum of the log-likelihood ℓ_Y over $\mathcal{M}_E$ can be computed as a double infimum:*

$$\sup_{e \in E} \ell_Y\left(e^\mathsf{T} e\right) = -\inf_{x \in \mathbb{R}_{>0}} \left(\frac{x}{n} \left(\inf_{h \in E_{\mathrm{SL}}^{\pm}} \|h \cdot Y\|^2 \right) - m \log(x) \right). \tag{2.7}$$

The MLEs, if they exist, are the matrices $\lambda h^\mathsf{T} h$, where $h \in E_{\mathrm{SL}}^{\pm}$ minimizes the inner infimum and $\lambda \in \mathbb{R}_{>0}$ is the unique minimum of the outer infimum.

Proof. Maximizing ℓ_Y over $\mathcal{M}_E$ is equivalent to minimizing

$$\varphi(e) = \frac{1}{n}\|e \cdot Y\|^2 - \log\det(e^\mathsf{T} e)$$

over E, where the MLE is $e^\mathsf{T} e$, by (2.3). We can write $e \in E$ as $e = \tau h$, where $\tau \in \mathbb{R}_{>0}$ and $h \in E_{\mathrm{SL}}^{\pm}$, since E is closed under nonzero scalar multiples. Then $e^\mathsf{T} e = \tau^2 h^\mathsf{T} h$ and, setting $x := \tau^2$, we have

$$\varphi(e) = \frac{x}{n}\|h \cdot Y\|^2 - m \log(x).$$

The convex function $x \mapsto xC/n - m\log(x)$ has minimum value $m(1 - \log(mn) + \log(C))$ for $C > 0$, which increases as C increases. For $C = 0$, the function is unbounded from

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

below. Hence, to minimize φ, we first find the minimal norm under $E_{\mathrm{SL}}^{\pm}$ and then minimize a univariate function. That is,

$$\inf_{e\in E}\varphi(e)=\inf_{x\in\mathbb{R}_{>0}}\left(\frac{x}{n}\left(\inf_{h\in E_{\mathrm{SL}}^{\pm}}\|h\cdot Y\|^2\right)-m\log x\right).$$

An MLE is a matrix in $\mathcal{M}_E$ that maximizes $\ell_Y(\Psi)$. We see that the MLEs are the matrices $\hat{\Psi}=e^{\mathsf{T}}e=\lambda h^{\mathsf{T}}h$, where $e=\sqrt{\lambda}h$, and h and λ minimize the inner and outer infima, respectively. ∎

ML estimation given Y is equivalent to minimizing the Kullback–Leibler (KL) divergence to the sample covariance matrix S_Y. The KL divergence is not a metric. Nevertheless, Theorem 2.8 shows the connection between minimizing the KL divergence and minimizing the Frobenius norm.

2.3. Weak Correspondence between Invariant Theory and ML Estimation.

Theorem 2.8 opens up the study of ML estimation in Gaussian models to the setting of invariant theory, as we now describe. Fix a set $E\subseteq \mathrm{GL}_m(\mathbb{R})$ and sample matrix $Y\in\mathbb{R}^{m\times n}$. By analogy to an orbit and stabilizer under a group action, we define the *orbit* and *stabilizer*[1] under the set E to be, respectively,

$$E\cdot Y:=\{eY\mid e\in E\},\qquad E_Y:=\{e\in E\mid eY=Y\}.$$

The following stability notions are motivated by invariant theory. We will see how they relate to ML estimation in Theorems 2.12, 3.3, 5.3, 6.5, and 7.4.

DEFINITION 2.9. *The matrix $Y\in\mathbb{R}^{m\times n}$, under the set E, is*

(i) unstable *if zero is contained in the Euclidean closure $\overline{E\cdot Y}$, i.e.,* $\mathrm{cap}_E(Y)=0$*;*
(ii) semistable *if Y is not unstable, i.e., $0\notin\overline{E\cdot Y}$;*
(iii) polystable *if $Y\neq 0$ and the set $E\cdot Y$ is Euclidean closed;*
(iv) stable *if Y is polystable and E_Y is finite.*

Remark 2.10. Invariant theory studies group actions on vector spaces. The above notions are usually studied for E a subgroup of GL_m. The group $E=G$ acts on the vector space $\mathbb{R}^{m\times n}$ via left multiplication. More generally, one can define the capacity and stability notions for the linear action of a group G on a normed vector space V. For details, see [40, section 1.4], which also explains the connection between the above topological definition and stability notions defined via invariants [37, page 41]. Stability notions are defined for sets, as above, in [35, Definition A.1].

Example 2.11. Any $Y\in\mathbb{R}^{m\times n}$ is unstable under $\mathrm{GL}_m(\mathbb{R})$: we have $(\varepsilon I_m)\cdot Y\to 0$ as $\varepsilon\to 0$. It is more interesting to consider the special linear group $E=\mathrm{SL}_m(\mathbb{R})$. Let $m=n$. Then Y is stable if it is invertible, and unstable otherwise. This can be seen by using Gaussian elimination to create a zero row in the noninvertible case, and observing that $\mathrm{SL}_m\cdot Y=\{X\in\mathbb{R}^{m\times m}\mid \det(X)=\det(Y)\}$ if Y is invertible.

THEOREM 2.12 (weak correspondence). *Let $E\subseteq\mathrm{GL}_m(\mathbb{R})$ be closed under nonzero scalar multiples. There is a correspondence between stability under $E_{\mathrm{SL}}^{\pm}$ and ML estimation in the model $\mathcal{M}_E$ given sample matrix $Y\in\mathbb{R}^{m\times n}$:*

$$\begin{array}{lccc}\text{(a)} & Y\ \textit{unstable} & \Leftrightarrow & \textit{log-likelihood}\ \ell_Y\ \textit{unbounded from above,}\\ \text{(b)} & Y\ \textit{semistable} & \Leftrightarrow & \textit{log-likelihood}\ \ell_Y\ \textit{bounded from above,}\\ \text{(c)} & Y\ \textit{polystable} & \Rightarrow & \textit{MLE given}\ Y\ \textit{exists.}\end{array}$$

[1] Properties from group actions need not hold; e.g., the sets $E\cdot Y$ may not partition $\mathbb{R}^{m\times n}$.

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Proof. We use the expression (2.7). Let $C = \inf_{h \in E_{\mathrm{SL}}^{\pm}} \|h \cdot Y\|^2$. For $C \geq 0$, the convex function $x \mapsto xC/n - m\log(x)$ is not bounded from below if and only if $C = 0$; see the proof of Theorem 2.8. Hence the log-likelihood is unbounded from above if and only if Y is unstable. This gives parts (a) and (b).

If Y is polystable, then $E_{\mathrm{SL}}^{\pm} \cdot Y$ is Euclidean closed. Hence, the capacity C is attained on the compact set $(E_{\mathrm{SL}}^{\pm} \cdot Y) \cap \{Z \in \mathbb{R}^{m \times n} \mid \|Z\|^2 \leq C+1\}$, i.e., there exists $h \in E_{\mathrm{SL}}^{\pm}$ with $C = \|h \cdot Y\|^2$. This implies $C > 0$, since $Y \neq 0$. Hence, there is a unique $\lambda \in \mathbb{R}_{>0}$ minimizing the outer infimum in (2.7) and $\lambda h^{\mathsf{T}} h$ is an MLE given Y, by Theorem 2.8. This gives (c). □

Remark 2.13. We can replace $E_{\mathrm{SL}}^{\pm}$ by E_{SL} in Theorems 2.8 and 2.12 if we assume that for any $e \in E$ there is an orthogonal matrix $o = o(e)$ such that $o^{\mathsf{T}} e \in E$ and $\det(o^{\mathsf{T}} e) > 0$. Indeed, we may then rewrite any $e \in E$ as $e = \tau o h$, where $\tau = \sqrt[m]{|\det(e)|}$ and $h = \tau^{-1} o^{\mathsf{T}} e$. Note that $h \in E_{\mathrm{SL}}$, since E is closed under nonzero scalar multiples. The same proof for Theorem 2.8 applies, using orthogonality of o.

3. Gaussian Group Models.

In this section, we study models $\mathcal{M}_E$ when the set E is a group, and we write $G = E$. We call the resulting statistical model the *Gaussian group model* given by G:

$$\mathcal{M}_G = \{g^{\mathsf{T}} g \mid g \in G\}.$$

ML estimation in the Gaussian group model $\mathcal{M}_G$ given a sample matrix $Y \in \mathbb{R}^{m \times n}$ is equivalent to norm minimization along the orbit of Y under the action by the group $H := G_{\mathrm{SL}}^{\pm}$, by Theorem 2.8. One can use the group $H := G_{\mathrm{SL}}$ whenever the additional assumption in Remark 2.13 holds. The infimal norm squared along its orbit is the capacity of Y under H, as defined in (2.6).

If we strengthen our assumptions on the group G, we can use tools from invariant theory to extend the weak correspondence Theorem 2.12.

DEFINITION 3.1. *Let G be a subgroup of GL_m. We call G* Zariski closed *if it is the zero locus of a set of polynomials in the matrix entries. The group G is* self-adjoint *if $g \in G$ implies $g^{\mathsf{T}} \in G$.*

On the statistics side, self-adjointness implies that the set of concentration matrices in the Gaussian group model $\mathcal{M}_G$ is equal to the set of covariance matrices in the model. Under the assumptions in Definition 3.1, we can use the Kempf–Ness theorem from invariant theory (see section 5) to obtain a *strong correspondence* between ML estimation over Gaussian group models and norm minimization over group orbits. Moreover, with these assumptions we are in the setting of [10], so we can use their algorithms to compute the capacity in order to find an MLE.

Remark 3.2. For a Zariski closed self-adjoint group $G \subseteq \mathrm{GL}_m(\mathbb{R})$ that is closed under nonzero scalar multiples, we can work with G_{SL} instead of $G_{\mathrm{SL}}^{\pm}$. Indeed, either there are no matrices in G of determinant -1, in which case $G_{\mathrm{SL}}^{\pm} = G_{\mathrm{SL}}$, or G contains an orthogonal matrix o of determinant -1 [2, Lemma 3.8]. In the latter case, we can invoke Remark 2.13, since, for any $g \in G$ of negative determinant, we have $o^{\top} g \in G$ and $\det(o^{\top} g) > 0$.

THEOREM 3.3 (strong correspondence). *Let $G \subseteq \mathrm{GL}_m$ be a Zariski closed self-adjoint group that is closed under nonzero scalar multiples. The stability under the action of G_{SL} on $(\mathbb{R}^m)^n$ relates to ML estimation for the Gaussian group model $\mathcal{M}_G$*

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

given a tuple of samples $Y \in (\mathbb{R}^m)^n$, *as follows:*

$$\begin{array}{llcccc}
\text{(a)} & Y \textit{ unstable} & \Leftrightarrow & \ell_Y \textit{ not bounded from above,} & & \\
\text{(b)} & Y \textit{ semistable} & \Leftrightarrow & \ell_Y \textit{ bounded from above,} & & \\
\text{(c)} & Y \textit{ polystable} & \Leftrightarrow & \textit{MLE exists,} & & \\
\text{(d)} & Y \textit{ stable} & \Rightarrow & \textit{finitely many MLEs exist} & \Leftrightarrow & \textit{unique MLE exists.}
\end{array}$$

We give a proof in section 5.3. All four stability conditions can occur; see Example 4.2. The converse of Theorem 3.3(d) does not hold generally over $\mathbb{R}$ [2, Example 4.2]. That G is self-adjoint is important for the equivalence in (d); without it, there can be finitely many nonunique MLEs [2, Example 3.12].

Remark 3.4. The concept of self-adjoint groups is closely related to the classical invariant-theoretic notion of linearly reductive groups [36, Theorems 7.1 and 7.2]; see [40, Theorem 1.3.10].

4. Matrix Normal Models.

Consider a multivariate Gaussian of dimension $m = m_1 m_2$. A *matrix normal model* consists of covariance matrices that factor as a Kronecker product[2] $\Sigma_1 \otimes \Sigma_2$, where $\Sigma_i \in \mathrm{PD}_{m_i}$. Setting $\Psi_1 := \Sigma_1^{-1}$ and $\Psi_2 := \Sigma_2^{-1}$, we can write the log-likelihood function (2.1) for the matrix normal model as

$$\ell_Y(\Psi_1, \Psi_2) = m_2 \log\det(\Psi_1) + m_1 \log\det(\Psi_2) - \frac{1}{n}\mathrm{tr}\left(\Psi_1 \sum_{i=1}^{n} Y_i \Psi_2 Y_i^{\mathsf{T}}\right). \tag{4.1}$$

An MLE is a concentration matrix $\hat{\Psi}_1 \otimes \hat{\Psi}_2 \in \mathrm{PD}_{m_1} \otimes \mathrm{PD}_{m_2}$ that maximizes (4.1).

4.1. Maximum Likelihood Estimation via Norm Minimization.

We specialize our results for general Gaussian group models to matrix normal models. Consider the left-right action of $\mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2}$ on $(\mathbb{R}^{m_1 \times m_2})^n$ given by

$$g \cdot Y := (g_1 Y_1 g_2^{\mathsf{T}}, \ldots, g_1 Y_n g_2^{\mathsf{T}}), \tag{4.2}$$

where $Y = (Y_1, \ldots, Y_n)$ is a tuple in $(\mathbb{R}^{m_1 \times m_2})^n$ and $g = (g_1, g_2) \in \mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2}$. The left-right action induces the representation

$$\varrho\colon\ \mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2} \to \mathrm{GL}_{m_1 m_2}, \quad (g_1, g_2) \mapsto g_1 \otimes g_2. \tag{4.3}$$

The concentration matrices in the Gaussian group model $\mathcal{M}_G$ of $G := \varrho(\mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2})$ are those of the form

$$(g_1 \otimes g_2)^{\mathsf{T}}(g_1 \otimes g_2) = g_1^{\mathsf{T}} g_1 \otimes g_2^{\mathsf{T}} g_2, \tag{4.4}$$

a Kronecker product of an $m_1 \times m_1$ concentration matrix and an $m_2 \times m_2$ concentration matrix. Thus, $\mathcal{M}_G$ is the matrix normal model described above.

This group $G \subseteq \mathrm{GL}_m$ is Zariski closed, self-adjoint, and closed under nonzero scalar multiples. Therefore, our results from the previous section apply to the action of G_{SL}. However, it is more convenient to work with the left-right action of $\mathrm{SL}_{m_1} \times \mathrm{SL}_{m_2}$. The following theorem (see [2, Theorem 4.1]) makes this precise.

[2] Given matrices $A_k \in \mathbb{R}^{m_k \times m_k}$, the Kronecker product $A_1 \otimes A_2$ is an $m_1 m_2 \times m_1 m_2$ matrix. Its rows are indexed by (i_1, i_2) and its columns by (j_1, j_2), where i_k and j_k range from 1 to m_k. The entry of $A_1 \otimes A_2$ at position $((i_1, i_2), (j_1, j_2))$ is $(A_1)_{i_1 j_1}(A_2)_{i_2 j_2}$.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

THEOREM 4.1. *Let $Y \in (\mathbb{R}^{m_1 \times m_2})^n$ be a matrix tuple. The supremum of the log-likelihood ℓ_Y in* (4.1) *over* $\mathrm{PD}_{m_1} \times \mathrm{PD}_{m_2}$ *is given by the double infimum*

$$- \inf_{\lambda \in \mathbb{R}_{>0}} \left(\frac{\lambda}{n} \left(\inf_{h \in \mathrm{SL}_{m_1} \times \mathrm{SL}_{m_2}} \|h \cdot Y\|^2 \right) - m_1 m_2 \log \lambda \right). \tag{4.5}$$

The MLEs, if they exist, are the matrices of the form $\lambda h_1^\mathsf{T} h_1 \otimes h_2^\mathsf{T} h_2$, where $h = (h_1, h_2)$ minimizes $\|h \cdot Y\|$ under the left-right action of $\mathrm{SL}_{m_1} \times \mathrm{SL}_{m_2}$, and $\lambda \in \mathbb{R}_{>0}$ is the unique value that minimizes the outer infimum. The stability under the left-right action of $\mathrm{SL}_{m_1} \times \mathrm{SL}_{m_2}$ is related to ML estimation via

(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 $\Leftrightarrow$ MLE exists uniquely.*

The converse of Theorem 4.1(d) does not hold over $\mathbb{R}$ [2, Example 4.2]: there are matrix tuples Y with a unique MLE that are polystable, but not stable. The following example shows that all four stability conditions (a)–(d) in Theorem 4.1 can occur.

Example 4.2. We study stability under $\mathrm{SL}_2 \times \mathrm{SL}_2$ on $(\mathbb{R}^{2 \times 2})^n$. Let

$$Y_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad Y_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y_4 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

(a) The matrix Y_4 is unstable and the matrix tuple (Y_4, Y_4) is unstable as well.

(b) The tuple $Y = (Y_1, Y_4)$ is semistable but not polystable, as follows. The orbit of Y is contained in $\{(g, M) \mid g \in \mathrm{SL}_2,\ M \neq 0\}$. Since SL_2 is closed, Y is semistable. Moreover, any $g \in \mathrm{SL}_2$ has Frobenius norm at least $\sqrt{2}$. Indeed, if σ_1 and σ_2 are the singular values of g, then $\|g\|^2 = \sigma_1^2 + \sigma_2^2$, where $\sigma_1 \sigma_2 = 1$. By the arithmetic mean–geometric mean inequality, we have $\|g\|^2 \geq 2$. So any $M \in \mathbb{R}^{2 \times 2} \setminus \{0\}$ satisfies $\|(g, M)\|^2 = \|g\|^2 + \|M\|^2 > 2$. On the other hand, we have

$$\left(\begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}, \begin{pmatrix} \varepsilon^{-1} & 0 \\ 0 & \varepsilon \end{pmatrix} \right) \cdot (Y_1, Y_4) = \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & \varepsilon^2 \\ 0 & 0 \end{pmatrix} \right),$$

which tends to $(Y_1, 0)$ as $\varepsilon \to 0$. Since $\|(Y_1, 0)\|^2 = 2$, the capacity of Y is not attained by an element in the orbit of Y, and Y is not polystable.

(c) The matrix $Y = Y_1$ is polystable. In fact, it is of minimal norm in its orbit, as explained in (b). Thus, an MLE given Y is $\lambda I_2 \otimes I_2$, where λ is the minimizer of the outer infimum in (4.5). Furthermore, Y is not stable, because its stabilizer is $\{(g, g^{-\mathsf{T}}) \mid g \in \mathrm{SL}_2\}$. There are infinitely many MLEs given Y, of the form $\lambda g^\mathsf{T} g \otimes g^{-1} g^{-\mathsf{T}}$ for $g \in \mathrm{SL}_2$; see Proposition 5.5.

(d) The tuple $Y = (Y_1, Y_2, Y_3)$ is stable, as follows. First, any tuple (M_1, M_2, M_3) in the orbit of Y satisfies $M_1, M_2 \in \mathrm{SL}_2$ and $\det(M_3) = -1$. Any 2×2 matrix of determinant ± 1 has Frobenius norm at least $\sqrt{2}$; see (b). Therefore, Y is of minimal norm in its orbit. Hence, an MLE given Y is $\lambda I_2 \otimes I_2$, where λ is the minimizer of the outer infimum in (4.5), and Y is polystable. The stabilizer of Y is the finite set $\{(I_2, I_2), (-I_2, -I_2)\}$.

4.2. Maximum Likelihood Thresholds. ML thresholds are the minimum number of samples needed for the MLE to exist or to be unique; see Definition 2.3. Before

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

our correspondence theorems, determining the ML thresholds of matrix normal models was an open problem in the statistics community, solved only in special cases or partially via lower or upper bounds [17].

Matrix normal models are examples of Gaussian group models on reductive groups. We have used descriptions of the null cone to give improved bounds on the number of samples generically required for a bounded likelihood function; see [2, section 4]. Subsequently, our dictionary between ML estimation and invariant theory was used by Derksen and Makam in [13] to determine the ML thresholds for matrix normal models, by combining Theorem 4.1 with representation theory of quivers. We state their result here.

THEOREM 4.3 ([13, Theorem 1.3]). *Let d be the greatest common divisor of m_1 and m_2, and $r := (m_1^2 + m_2^2 - d^2)/m$. The ML thresholds of the matrix normal model satisfy* $\mathrm{mlt}_{\mathrm{b}} = \mathrm{mlt}_{\mathrm{e}}$*, and the following hold:*

- *If $m_1 = m_2 = 1$, then* $\mathrm{mlt}_{\mathrm{e}} = \mathrm{mlt}_{\mathrm{u}} = 1$.
- *If $m_1 = m_2 > 1$, then* $\mathrm{mlt}_{\mathrm{e}} = 1$ *and* $\mathrm{mlt}_{\mathrm{u}} = 3$.
- *If $m_1 \neq m_2$ and $r \in \mathbb{Z}$, then* $\mathrm{mlt}_{\mathrm{e}} = r$*. If $d = 1$, then* $\mathrm{mlt}_{\mathrm{u}} = r$*; otherwise* $\mathrm{mlt}_{\mathrm{u}} = r + 1$.
- *If $m_1 \neq m_2$ and $r \notin \mathbb{Z}$, then* $\mathrm{mlt}_{\mathrm{e}} = \mathrm{mlt}_{\mathrm{u}} = \lceil (m_1^2 + m_2^2)/m \rceil$.

The invariant-theoretic viewpoint provided a new perspective on the problem of determining the ML thresholds. From statistics, it is natural to fix m_1 and m_2 and let the number of samples n vary. From this viewpoint, it is challenging to spot a pattern among ML thresholds. However, from the invariant theory of quivers, it is more natural to fix n and let m_1 and m_2 vary.

Matrix normal models can be extended to *tensor normal models* by replacing the map in (4.3) with $(g_1, \ldots, g_d) \mapsto g_1 \otimes \cdots \otimes g_d$. Our correspondence theorem led to a complete description of ML thresholds for tensor normal models [14, Theorem 1.1], which includes Theorem 4.3 as a special case.

5. The Kempf–Ness Theorem.

The Kempf–Ness theorem can be thought of as a noncommutative version of linear programming duality. It applies to actions by Zariski closed and self-adjoint groups. Norm minimization along an orbit of such a group action is a nonlinear optimization problem. Nevertheless, the Kempf–Ness theorem states that every critical point of that optimization problem is a global minimum. The underlying reason is *geodesic convexity.*

The Kempf–Ness theorem is classically stated over the complex numbers. We denote by $\mathbb{K}$ either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers and state the Kempf–Ness theorem over both fields. We consider a subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$ that is Zariski closed and self-adjoint. When $\mathbb{K} = \mathbb{C}$, self-adjoint means that

$$g \in G \text{ implies } g^* \in G,$$

where g^* denotes the conjugate transpose of g.

The group G acts on the vector space $\mathbb{K}^m$ via left-multiplication; i.e., a group element $g \in G$ acts on a vector $v \in \mathbb{K}^m$ to give the vector $gv \in \mathbb{K}^m$. For a vector $v \in \mathbb{K}^m$, we wish to minimize the norm along its orbit under this action. In other words, we seek the infimum of

$$\gamma_v \colon G \longrightarrow \mathbb{R}, \quad g \longmapsto \|gv\|^2.$$

The calculus approach to this minimization problem is to compute all critical points

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

and then identify the minima. To find the critical points, we compute the derivative of the map γ_v.

Here, we can use the fact that G is a group: for any vector w in the orbit of v, the orbits $G \cdot w$ and $G \cdot v$ are equal. Therefore, a vector w is of minimal norm in the orbit $G \cdot v$ if and only if it is of minimal norm in its own orbit $G \cdot w$. The latter is equivalent to the identity matrix being a minimizer of the map γ_w. Hence, we do not have to compute the derivative of γ_v at arbitrary group elements $g \in G$. It suffices to compute the derivative of γ_w at the identity matrix for arbitrary $w \in \mathbb{K}^m$.

Since $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is Zariski closed (i.e., defined by polynomial equations), its tangent space $T_I G$ at the $m \times m$ identity matrix I can be computed as a vector subspace of $\mathbb{K}^{m \times m}$. The derivative of the map γ_w at the identity is a linear map

$$D_I \gamma_w \colon T_I G \longrightarrow \mathbb{R}. \tag{5.1}$$

The identity matrix is a critical point of the map γ_w if and only if the derivative $D_I \gamma_w$ is the zero map. To analyze this behavior for all vectors w simultaneously, we study the *moment map* μ that assigns to each vector w the corresponding derivative:

$$\mu \colon \mathbb{K}^m \longrightarrow \mathrm{Hom}_{\mathbb{R}}(T_I G, \mathbb{R}), \quad w \longmapsto D_I \gamma_w.$$

The moment map vanishes at w if and only if the identity matrix I is a critical point of the map γ_w. This is a necessary criterion for I to be a minimizer of γ_w; i.e., for the vector w to be of minimal norm in its orbit. The Kempf–Ness theorem says that it is also sufficient. This can be thought of as a convexity theorem: indeed, this is where geodesic convexity comes into play; see [10, section 3.2].

There are several equivalent formulations of the Kempf–Ness theorem in the literature, some containing further details on the equivalence between critical points and minimizers. We formulate the relevant versions for our correspondence theorems relating invariant theory and ML estimation. The proof over $\mathbb{K} = \mathbb{C}$ is due to [30]; the first proof for $\mathbb{K} = \mathbb{R}$ was given in [41].

THEOREM 5.1 (Kempf–Ness). *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup with moment map μ, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. If $\mathbb{K} = \mathbb{R}$, let K be the set of orthogonal matrices in G. If $\mathbb{K} = \mathbb{C}$, let K be the set of unitary matrices in G. For $v \in \mathbb{K}^m$, we have the following:*

(a) *The vector v is of minimal norm in its orbit if and only if $\mu(v) = 0$.*

(b) *If $\mu(v) = 0$ and $w \in G \cdot v$ is such that $\|v\| = \|w\|$, then $w \in K \cdot v$.*

(c) *If the orbit $G \cdot v$ is closed, then there exists some $w \in G \cdot v$ with $\mu(w) = 0$.*

(d) *If $\mu(v) = 0$, then the orbit $G \cdot v$ is closed.*

(e) *The vector v is polystable if and only if there exists $0 \neq w \in G \cdot v$ with $\mu(w) = 0$.*

(f) *The vector v is semistable if and only if there exists $0 \neq w \in \overline{G \cdot v}$ with $\mu(w) = 0$.*

Part (a) says that finding a vector of minimal norm in an orbit is equivalent to finding a vector in that orbit where the moment map vanishes. Moreover, for nonzero v, we can rewrite part (f) as

$$\mathrm{cap}_G(v) = \inf_{w \in G \cdot v} \|w\|^2 > 0 \qquad \Leftrightarrow \qquad \inf_{w \in G \cdot v} \|\mu(w)\| = 0. \tag{5.2}$$

In other words, norm minimization along the orbit of a vector v has the following dual formulation: find a vector $w \in G \cdot v$ such that $\mu(w) = 0$. This dual problem is

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

sometimes known as a scaling problem [10, Problem 1.9], since it generalizes matrix, operator, and tensor scaling; see [24] and [40, section 3.1]. In the special case that the acting group is a torus, this duality of optimization problems is precisely linear programming duality; see [24, section 3.2.3].

5.1. Illustrative Example: The Conjugation Action.

Consider the *conjugation action* of $G := \mathrm{GL}_m(\mathbb{C})$ on $\mathbb{C}^{m\times m}$. That is, $g \in \mathrm{GL}_m(\mathbb{C})$ acts on a matrix $A \in \mathbb{C}^{m\times m}$ by

$$g \cdot A := gAg^{-1}.$$

This is also called the adjoint action in the literature. After appropriate identification[3] the moment map of the conjugation action is

$$\mu\colon \mathbb{C}^{m\times m} \to \mathbb{C}^{m\times m}, \quad A \mapsto AA^* - A^*A. \tag{5.3}$$

We characterize all stability notions using the Jordan normal form and the Kempf–Ness theorem. Recall from linear algebra that the orbits of the conjugation action can be represented by the Jordan normal form. For $\varepsilon > 0$ and $\lambda \in \mathbb{C}$, consider the conjugation

$$\begin{pmatrix} \varepsilon^{-1} & & \\ & \varepsilon^{-2} & \\ & & \varepsilon^{-3} \end{pmatrix} \begin{pmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{pmatrix} \begin{pmatrix} \varepsilon & & \\ & \varepsilon^{2} & \\ & & \varepsilon^{3} \end{pmatrix} = \begin{pmatrix} \lambda & \varepsilon & \\ & \lambda & \varepsilon \\ & & \lambda \end{pmatrix} \xrightarrow{\varepsilon\to 0} \begin{pmatrix} \lambda & & \\ & \lambda & \\ & & \lambda \end{pmatrix}.$$

The computation generalizes: a Jordan block of size $k \times k$ with eigenvalue λ can be scaled in the limit to $\mathrm{diag}(\lambda, \dots, \lambda)$ via conjugation. A matrix A is *diagonalizable* if there exists a diagonal matrix D in the orbit $G \cdot A$. Since the Jordan normal form is a block-diagonal matrix consisting of Jordan blocks, we deduce the following:

1. A nilpotent matrix, i.e., one having zero as its only eigenvalue, contains the zero matrix in its orbit closure and hence is unstable.
2. Any nondiagonalizable matrix contains in its orbit closure a diagonal matrix, which is a different Jordan normal form. Therefore, the orbit of a nondiagonalizable matrix is not closed; i.e., the matrix cannot be polystable.

These are first steps towards characterizing the stability notions. The Kempf–Ness theorem completes the picture, which is as follows.

PROPOSITION 5.2. *Let $A \in \mathbb{C}^{m\times m}$ and consider the conjugation action by $\mathrm{GL}_m(\mathbb{C})$.*
(i) *A is unstable if and only if A is nilpotent.*
(ii) *A is polystable if and only if $A \neq 0$ is diagonalizable.*
In particular, A is semistable but not polystable if and only if A is nondiagonalizable and nonnilpotent. Moreover, there are no stable matrices under this action.

Proof. (i) Item 1 shows the "if" direction. For the converse, assume A is not nilpotent. Then A can be scaled in the limit to a nonzero diagonal matrix by the above argument on the Jordan normal form. The moment map vanishes on diagonal matrices, by (5.3). Thus, applying Kempf–Ness part (f), we conclude that A is semistable.

(ii) Item 2 shows that if A is polystable, then it must be diagonalizable. For the converse, assume A is nonzero and diagonalizable. Then A contains a nonzero

[3] The derivative of the map $g \mapsto \|gAg^{-1}\|^2$ at the identity sends $\dot{g} \in T_I G = \mathbb{C}^{m\times m}$ to $2\,\mathrm{Re}[\mathrm{tr}(\dot{g}(AA^* - A^*A))]$. This is an $\mathbb{R}$-linear map, i.e., an element in $\mathrm{Hom}_{\mathbb{R}}(T_I G, \mathbb{R})$. We can identify $\mathrm{Hom}_{\mathbb{R}}(T_I G, \mathbb{R})$ with $\mathbb{C}^{m\times m}$ by viewing a matrix $M \in \mathbb{C}^{m\times m}$ as an $\mathbb{R}$-linear map $\dot{g} \mapsto 2\,\mathrm{Re}[\mathrm{tr}(\dot{g}M)]$.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

diagonal matrix in its orbit. Since the moment map vanishes on such matrix, we conclude from Kempf–Ness part (e) that A is polystable.

Finally, note that for any $A \in \mathbb{C}^{m\times m}$ the stabilizer G_A contains the infinite set $\{\alpha I_m \mid \alpha \in \mathbb{C}\backslash\{0\}\}$. Thus, no matrix is stable. □

Parts (a), (b), and (d) of Kempf–Ness recover well-known results from linear algebra. Part (d) shows that a normal matrix is diagonalizable, as follows. Let $A \in \mathbb{C}^{m\times m}$ be a normal matrix, i.e., $AA^* = A^*A$. This is equivalent to $\mu(A) = 0$, by (5.3). The orbit of A is closed, by part (d) of Kempf–Ness. Hence, any nonzero normal matrix A is polystable, and by Proposition 5.2(ii) we conclude that such an A is diagonalizable.

Part (b) shows that any normal matrix is diagonalizable by a unitary matrix. The group K here is the group of unitary matrices. Let A be a normal matrix and $D \in G \cdot A$ diagonal. Since D is also normal, part (a) of Kempf–Ness tells us that A and D are both of minimal norm in their orbits. Applying part (b), we conclude that $D \in G \cdot A$ actually lies in $K \cdot A$.

Part (a) shows that a normal matrix A is of minimal Frobenius norm in its orbit. This can also be deduced using the Schur decomposition. Indeed, let $B \in G \cdot A$ and consider a Schur decomposition $B = URU^*$, where $U \in K$ and R is upper triangular. Since A and B lie in the same orbit they have the same eigenvalues given by diagonal matrix D. Hence, $R = D + N$ with N strictly upper triangular. Then

$$\|B\|^2 = \|R\|^2 = \|D\|^2 + \|N\|^2 \geq \|D\|^2 = \|A\|^2.$$

The conjugation action over $\mathbb{R}$ is more delicate. While still true that $A \in \mathbb{R}^{m\times m}$ is unstable under the conjugation action by $\mathrm{GL}_m(\mathbb{R})$ if and only if A is nilpotent, there exist matrices that are polystable but not diagonalizable over $\mathbb{R}$, such as

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

whose eigenvalues are $\pm i$. This matrix A also satisfies $\mu(A) = 0$ since it is normal, i.e., $AA^\mathsf{T} = A^\mathsf{T}A$, and Kempf–Ness part (a) implies that it is of minimal norm in its orbit under $\mathrm{GL}_m(\mathbb{R})$. Kempf–Ness part (b) allows us to recover the fact that if A is symmetric (and thus has only real eigenvalues), then it is orthogonally diagonalizable.

5.2. Complex Gaussian Models. Motivated by the Kempf–Ness theorem over the complex numbers, we study Gaussian models defined over $\mathbb{C}$ [5, 25, 44]. Given a subset $E \subseteq \mathrm{GL}_m(\mathbb{C})$, we define the associated *complex Gaussian model* to be

$$\mathcal{M}_E = \{e^*e \mid e \in E\}.$$

When the set E is closed under nonzero scalar multiples, then ML estimation over the complex model $\mathcal{M}_E$ given data $Y \in (\mathbb{C}^m)^n$ is equivalent to norm minimization along the orbit of Y under the set $E_{\mathrm{SL}} = \{e \in E \mid \det(e) = 1\}$. In other words, Theorem 2.8 holds analogously after replacing $E_{\mathrm{SL}}^{\pm}$ by E_{SL} (and the transpose by the conjugate transpose). When, in addition, the set E is a group, denoted again by $G = E$, that is Zariski closed and self-adjoint, then the strong correspondence Theorem 3.3 becomes a *full correspondence* in the sense that (a)–(d) hold and (d) becomes an equivalence.

THEOREM 5.3 (full correspondence). *Let $Y \in (\mathbb{C}^m)^n$ be a tuple of samples, and let $G \subseteq \mathrm{GL}_m(\mathbb{C})$ be a Zariski closed self-adjoint group that is closed under nonzero*

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

scalar multiples. The stability under the action of G_{SL} on $(\mathbb{C}^m)^n$ is related to ML estimation for the complex Gaussian group model $\mathcal{M}_G$ as follows:

$$\begin{array}{llcll}
\text{(a)} & Y \textit{ unstable} & \Leftrightarrow & \ell_Y \textit{ not bounded from above}, & \\
\text{(b)} & Y \textit{ semistable} & \Leftrightarrow & \ell_Y \textit{ bounded from above}, & \\
\text{(c)} & Y \textit{ polystable} & \Leftrightarrow & \textit{MLE exists}, & \\
\text{(d)} & Y \textit{ stable} & \Leftrightarrow & \textit{finitely many MLEs exist} & \Leftrightarrow \textit{ unique MLE exists.}
\end{array}$$

We outline the proof in the next subsection. One use case is that of *complex matrix normal models*, whose concentration matrices are of the form

$$(g_1 \otimes g_2)^*(g_1 \otimes g_2) = g_1^* g_1 \otimes g_2^* g_2$$

for $g_1 \in \mathrm{GL}_{m_1}(\mathbb{C})$ and $g_2 \in \mathrm{GL}_{m_2}(\mathbb{C})$. Theorem 4.1 holds analogously over $\mathbb{C}$ (after replacing the transpose by the conjugate transpose in the line after (4.5)) and it becomes a full correspondence with an equivalence in (d). Moreover, all ML thresholds reported in Theorem 4.3 also hold for complex matrix normal models.

5.3. Proof of the Strong and Full Correspondences. We prove Theorems 3.3 and 5.3 for real, respectively, complex Gaussian group models with Zariski closed and self-adjoint group. For this, we establish two propositions. The first one holds for any Gaussian group model and shows that the group action yields symmetries that are meaningful from an MLE perspective.

PROPOSITION 5.4. *Let $G \subseteq \mathrm{GL}_m(\mathbb{R})$ be a subgroup and consider the Gaussian group model $\mathcal{M}_G$ with sample matrix $Y \in \mathbb{R}^{m \times n}$. For $h \in G_{\mathrm{SL}}^{\pm}$, the following hold:*

(i) *The supremum of ℓ_Y equals the supremum of $\ell_{h \cdot Y}$.*
(ii) *There exists an MLE given Y if and only if there exists an MLE given $h \cdot Y$ and, in that case,*

$$\{\textit{MLEs given } h \cdot Y\} = (h^{-1})^{\mathsf{T}} \{\textit{MLEs given } Y\} h^{-1}. \tag{5.4}$$

Proof. Set $\tilde{g} := gh^{-1} \in G$. Then (2.3) and $\log(\det(h^{-1})^2) = 0$ yield

$$\begin{aligned}
\ell_{h \cdot Y}\left(\tilde{g}^{\mathsf{T}} \tilde{g}\right) &= \log\left(\det\left((gh^{-1})^{\mathsf{T}} gh^{-1}\right)\right) - \frac{1}{n}\|(gh^{-1}) \cdot (h \cdot Y)\|^2 \\
&= \log(\det(g^{\mathsf{T}} g)) + \log(\det(h^{-1})^2) - \frac{1}{n}\|g \cdot Y\|^2 = \ell_Y(g^{\mathsf{T}} g).
\end{aligned}$$

This proves (i) and shows that $\tilde{g}^{\mathsf{T}} \tilde{g} = (h^{-1})^{\mathsf{T}}(g^{\mathsf{T}} g) h^{-1}$ is an MLE given $h \cdot Y$ if and only if $g^{\mathsf{T}} g$ is an MLE given Y. Hence, we also obtain (ii). ∎

From now on, we consider Zariski closed self-adjoint groups G. In the upcoming proofs, this serves us in two ways. First, we can apply Theorems 2.8 and 2.12 using G_{SL}; see Remark 3.2. Second, G_{SL} is also Zariski closed (given by the additional polynomial equation $\det(g) = 1$) and self-adjoint, and hence we can apply Kempf–Ness for the action of G_{SL}.

PROPOSITION 5.5. *Let $Y \in \mathbb{R}^{m \times n}$ be a sample matrix. Consider $\mathcal{M}_G$, where $G \subseteq \mathrm{GL}_m(\mathbb{R})$ is a Zariski closed self-adjoint subgroup closed under nonzero scalar multiples. Let Stab_Y be the stabilizer of Y under G_{SL}. If Ψ is an MLE given Y, then*

$$\big\{\textit{MLEs given } Y\big\} = \big\{g^{\mathsf{T}} \Psi g \mid g \in \mathrm{Stab}_Y\big\}.$$

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Proof. We have $\Psi = \lambda h^{\mathsf{T}} h$, where $\lambda > 0$ is unique and h minimizes the norm of Y under the action of G_{SL}, by Theorem 2.8. The MLEs given Y are exactly the matrices of that form. For any $g \in \mathrm{Stab}_Y$, the matrix $hg \in G_{\mathrm{SL}}$ also minimizes the norm of Y under the action of G_{SL}. Therefore, $\lambda(hg)^{\mathsf{T}} hg = g^{\mathsf{T}}(\lambda h^{\mathsf{T}} h)g$ is another MLE. Conversely, if $\lambda(h')^{\mathsf{T}} h'$ with $h' \in G_{\mathrm{SL}}$ is another MLE, then

$$\|h' \cdot Y\|^2 = \inf_{\tilde{h} \in G_{\mathrm{SL}}} \|\tilde{h} \cdot Y\|^2 = \|h \cdot Y\|^2.$$

Applying Kempf–Ness part (b) to the G_{SL}-action, there is an orthogonal matrix $o \in G_{\mathrm{SL}}$ with $o \cdot (h \cdot Y) = h' \cdot Y$. Hence, $g := h^{-1}\, o^{-1}\, h' \in \mathrm{Stab}_Y$ and, using $h' = ohg$, we deduce that $\lambda(h')^{\mathsf{T}} h' = g^{\mathsf{T}}(\lambda h^{\mathsf{T}} h)g$. □

Equipped with these propositions, we prove the strong correspondence.

Proof of Theorem 3.3. By the weak correspondence Theorem 2.12, it remains to prove the converse implication in (c) and all of (d). For the former, assume an MLE given Y exists. Then $Y \neq 0$ and there is $h \in G_{\mathrm{SL}}$ such that $h \cdot Y$ has minimal norm in the orbit of Y under G_{SL}; see Theorem 2.8. Hence, the orbit $G_{\mathrm{SL}} \cdot Y$ is closed by Kempf–Ness part (d), and Y is polystable.

We now prove (d). If Y is stable under G_{SL}, its stabilizer Stab_Y is finite. Thus, there are only finitely many MLEs given Y, by Proposition 5.5. It remains to show that a tuple Y cannot have finitely many MLEs unless it has a unique MLE. A tuple Y with finitely many MLEs is polystable, by (c). Moreover, (5.4) holds and we can relate the stabilizers of Y and $h \cdot Y$ by $\mathrm{Stab}_{h \cdot Y} = h\, \mathrm{Stab}_Y\, h^{-1}$. Hence, to study the stabilizer and MLEs of a polystable Y, we can assume that Y is of minimal norm in its G_{SL}-orbit. One of the MLEs given Y is then λI_m, where $\lambda > 0$ minimizes the outer infimum in Theorem 2.8.

We show that the set $\{g^{\mathsf{T}} g \mid g \in \mathrm{Stab}_Y\}$ is either the identity matrix or infinite. This implies that Y has either a unique MLE or infinitely many MLEs, by Proposition 5.5. The group Stab_Y is self-adjoint, by [43, Corollary 2.25]. If it is contained in the set of orthogonal matrices, then $\{g^{\mathsf{T}} g \mid g \in \mathrm{Stab}_Y\} = \{I_m\}$. Otherwise, let $h \in \mathrm{Stab}_Y$ be nonorthogonal. Then $h^{\mathsf{T}} \in \mathrm{Stab}_Y$ and hence $h^{\mathsf{T}} h \in \mathrm{Stab}_Y$, and this positive definite matrix is not equal to the identity matrix. The matrix $h^{\mathsf{T}} h$ has infinite order, since the eigenvalues of $(h^{\mathsf{T}} h)^N$ are the Nth powers of the eigenvalues of $h^{\mathsf{T}} h$, and there exist eigenvalues that are not equal to one. Since $(h^{\mathsf{T}} h)^N \in \mathrm{Stab}_Y$ and $((h^{\mathsf{T}} h)^N)^{\mathsf{T}}((h^{\mathsf{T}} h)^N) = (h^{\mathsf{T}} h)^{2N}$, the set $\{g^{\mathsf{T}} g \mid g \in \mathrm{Stab}_Y\}$ is infinite. □

We now turn to complex Gaussian group models. Statements for real Gaussian group models hold analogously in the complex case after replacing transposes by conjugate transposes.

Proof of Theorem 5.3. Theorems 2.8 and 2.12 using G_{SL} hold over $\mathbb{C}$. Similarly, Propositions 5.4 and 5.5, the strong correspondence Theorem 3.3, and their proofs carry over to the complex case.

We prove the reverse implication in (d): assuming there exists a unique MLE given Y, we show that Y is stable. Such a Y is polystable, by part (c). To show that Stab_Y is finite we can assume as in the real setting that Y is of minimal norm in its G_{SL}-orbit. Then λI_m is the unique MLE given Y, by the complex analogue of Theorem 2.8. Hence, Stab_Y is contained in the group of unitary matrices in G, by the complex analogue of Proposition 5.5. In particular, Stab_Y is $\mathbb{C}$-compact. As the stabilizer Stab_Y is Zariski closed (defined by the equations $gY = Y$), we conclude that Stab_Y is finite. □

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

The above proof fails over $\mathbb{R}$, since there are infinite sets that are compact and Zariski closed, such as the orthogonal group. See [2, Example 4.2].

6. Directed Gaussian Graphical Models.

In this section, we study directed Gaussian graphical models on directed acyclic graphs, also known as DAG models. We show that the full correspondence between the four notions of stability and ML estimation holds for DAG models. These results are from [2, section 5] combined with the follow-up work [35], which extended the results from transitive DAGs to all DAGs.[4] Though this section studies directed graphs, the results also cover certain undirected graphical models; see [2, Remark 5.9] and [35, section 3].

6.1. DAG Models.

Let $\mathcal{G}$ be a directed graph on m nodes. We label the nodes by $[m] = \{1, 2, \ldots, m\}$. We denote an edge from j to i by $j \to i$; otherwise, if there is no such edge, we write $j \not\to i$. A directed acyclic graph (DAG) is a directed graph with no directed cycles, i.e., no cycles $i \to j \to \cdots \to k \to i$. Edges $i \to i$ do not appear in a DAG as they are cycles of length one. A DAG $\mathcal{G}$ is *transitive* (and then known as a TDAG) if $k \to j$ and $j \to i$ in $\mathcal{G}$ imply $k \to i$ in $\mathcal{G}$. The *parents* of i in $\mathcal{G}$ comprise the set of vertices

$$\mathrm{pa}(i) = \{j \in [m] \mid (j \to i) \in E\}.$$

We define a statistical model and a set of matrices associated to a DAG $\mathcal{G}$. We demonstrate the link between ML estimation in the model and stability with respect to the set of matrices.

First we define the DAG model. It is the linear structural equation model given by the equation

$$Y = \Lambda Y + \varepsilon, \tag{6.1}$$

where Y is a random vector on $\mathbb{R}^m$, the matrix $\Lambda \in \mathbb{R}^{m \times m}$ satisfies $\Lambda_{ij} = 0$ for $j \not\to i$ in $\mathcal{G}$, and $\varepsilon \sim N(0, \Omega)$ with $\Omega \in \mathbb{R}^{m \times m}$ diagonal and positive definite. The model writes each Y_i as a linear combination of all Y_j with $j \to i$, up to Gaussian error. This consists of multiple instances of classical linear regression, organized by the graph. Each variable is conditionally independent of its nondescendants given its parents (see, e.g., [42, Chapter 13]). This is why such DAG models are also known as Gaussian Bayesian networks.

Solving for Y in (6.1), we have

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

where the acyclicity of $\mathcal{G}$ implies that $(I - \Lambda)$ is invertible. We see that Y is multivariate Gaussian with covariance matrix and concentration matrix, respectively,

$$\Sigma = (I - \Lambda)^{-1}\Omega(I - \Lambda)^{-\mathsf{T}}, \qquad \Psi = (I - \Lambda)^{\mathsf{T}}\Omega^{-1}(I - \Lambda). \tag{6.2}$$

The DAG model $\mathcal{M}_{\mathcal{G}}^{\to}$ consists of the set of concentration matrices Ψ of the form in (6.2) for Λ and Ω defined in terms of $\mathcal{G}$ as above.

There are close connections between DAG models and our Gaussian group models from section 3. We define a set of matrices associated to $\mathcal{G}$:

$$E(\mathcal{G}) = \{e \in \mathrm{GL}_m(\mathbb{R}) \mid e_{ij} = 0 \text{ for } i \neq j \text{ with } j \not\to i \text{ in } \mathcal{G}\}. \tag{6.3}$$

We define $\mathcal{M}_{E(\mathcal{G})} = \{e^{\mathsf{T}}e : e \in E(\mathcal{G})\}$.

[4] And from usual DAGs to statistical models defined on a DAG together with a coloring of its vertices and edges.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

PROPOSITION 6.1. *The set of concentration matrices in* (6.2) *equals the set* $\mathcal{M}_{E(\mathcal{G})}$. *The set of matrices* $E(\mathcal{G})$ *is a group if and only if* $\mathcal{G}$ *is a TDAG.*

Proof. The equality of models follows from reparametrizing $(I-\Lambda)^{\mathsf{T}}\Omega^{-1}(I-\Lambda)$ by $e^{\mathsf{T}}e$, where $e=\Omega^{-\frac{1}{2}}(I-\Lambda)\in E(\mathcal{G})$. The characterization of when $E(G)$ is a group is [2, Proposition 5.1]. □

For a TDAG $\mathcal{G}$ we denote $E(\mathcal{G})$ by $G(\mathcal{G})$ to highlight that it is a group.

Example 6.2. Let $\mathcal{G}$ be the DAG $1 \leftarrow 3 \rightarrow 2$. It is transitive, since it contains no paths of length greater than one. The group $G(\mathcal{G}) \subseteq \mathrm{GL}_3$ consists of invertible matrices

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

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

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

6.2. The Full Correspondence for DAG Models. MLE existence and uniqueness of DAG models are characterized by linear dependence conditions on the sample matrix, as we now describe. For a DAG $\mathcal{G}$ on m nodes and n data samples, the sample matrix is $Y\in\mathbb{R}^{m\times n}$. For a node i in $\mathcal{G}$ we denote by $Y^{(i)}$ the ith row of Y, by $Y^{(\mathrm{pa}(i))}$ the submatrix of Y with rows indexed by the parents of i in $\mathcal{G}$, and by $Y^{(\mathrm{pa}(i)\cup i)}$ the submatrix of Y with rows indexed by node i and its parents. The following conditions relate the MLE to the rows of Y.

THEOREM 6.3. *Consider the DAG model on* $\mathcal{G}$ *with* m *nodes, and fix sample matrix* $Y\in\mathbb{R}^{m\times n}$. *The following possibilities characterize ML estimation given* Y*:*

$$\begin{array}{llcl}
\text{(a)} & \ell_Y \textit{ unbounded from above} & \Leftrightarrow & \exists\, i\in[m]\colon Y^{(i)}\in\operatorname{span}\{Y^{(j)}: j\in\mathrm{pa}(i)\},\\
\text{(b)} & \textit{MLE exists} & \Leftrightarrow & \forall\, i\in[m]\colon Y^{(i)}\notin\operatorname{span}\{Y^{(j)}: j\in\mathrm{pa}(i)\},\\
\text{(c)} & \textit{MLE exists uniquely} & \Leftrightarrow & \forall\, i\in[m]\colon Y^{(\mathrm{pa}(i)\cup i)} \textit{ has full row rank.}
\end{array}$$

Remark 6.4. See [35, Theorem 4.9] for a proof of Theorem 6.3. If Y has a row of zeros, then ℓ_Y is unbounded from above. This satisfies the criterion in the above theorem: a row of zeros at row i is interpreted as a trivial linear combination, independently of whether or not node i has parents in $\mathcal{G}$. The above trichotomy shows that the log-likelihood given Y is bounded from above if and only if the MLE given Y exists.

For DAG models, we prove a full correspondence between the four notions of stability and the four possibilities for ML existence. The set $E(\mathcal{G})$ from (6.3) is closed under nonzero scalar multiples. Consequently, Theorem 2.8 (MLE via norm minimization) and the weak correspondence Theorem 2.12 hold for the DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}=\mathcal{M}_{E(\mathcal{G})}$ using $E(\mathcal{G})_{\mathrm{SL}}$. Here we use Remark 2.13, since the orthogonal matrices $o_1=I_m$ and $o_2:=\operatorname{diag}(-1,1,\ldots,1)$ satisfy $o_k^{\mathsf{T}}e\in E(\mathcal{G})$ for all $e\in E(\mathcal{G})$. Although our previous results only established the full correspondence for complex Gaussian group models on self-adjoint groups, we are able to establish the full correspondence for real DAG models, which may not be Gaussian group models (if the DAG is not transitive) and whose group in question is not self-adjoint.

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Theorem 6.5 (the full correspondence for DAG models). *Consider a DAG model on $\mathcal{G}$ and sample matrix $Y \in \mathbb{R}^{m\times n}$. Then stability under $E(\mathcal{G})_{\mathrm{SL}}$ relates to ML estimation as follows:*

$$\begin{array}{rccc} \text{(a)} & Y \textit{ unstable} & \Leftrightarrow & \ell_Y \textit{ unbounded from above,} \\ \text{(b)} & Y \textit{ semistable} & \Leftrightarrow & \ell_Y \textit{ bounded from above,} \\ \text{(c)} & Y \textit{ polystable} & \Leftrightarrow & \textit{MLE exists,} \\ \text{(d)} & Y \textit{ stable} & \Leftrightarrow & \textit{MLE exists uniquely.} \end{array}$$

Theorem 6.5 can be proved by showing that stability conditions are equivalent to the linear independence conditions of Theorem 6.3. A complete proof is given in [35, Theorem A.2], where it is proved in the more general setting of restricted DAG (RDAG) models. Usual DAG models are the special case where every vertex and every edge has a unique color. The equivalence in (a), (b), and (c) was proved for TDAGs in [2, Theorem 5.3]. The group $G(\mathcal{G})$ associated to a TDAG $\mathcal{G}$ is Zariski closed and closed under nonzero scalar multiples, but it is not self-adjoint. Hence we are not in the setting of Theorem 3.3, where our general results show the strong correspondence, so instead [2, Theorem 5.3] gives a linear algebra based proof.

We relate the MLEs to the stabilizer E_Y; cf. Proposition 5.5.

Proposition 6.6. *Fix the DAG model on $\mathcal{G}$ and set $E := E(\mathcal{G})_{\mathrm{SL}}$. If $\lambda e^{\mathsf{T}} e$ is an MLE given Y, where $e \in E$ and $\lambda > 0$ is a scalar, then the set of MLEs given Y is in bijection with E_Y under mapping $b \in E_Y$ to $\lambda(e+b-I)^{\mathsf{T}}(e+b-I)$.*

See [35, Proposition A.3] for a proof. We now discuss consequences for ML thresholds; see Definition 2.3. The results of Theorems 6.3 and 6.5 characterize MLE existence for any tuple Y. We derive a corollary for generic tuples, regarding the ML thresholds. This is known in the graphical models literature; see [31, section 5.4.1] and [16, Theorem 1]. The *in-degree* of a DAG $\mathcal{G}$ is the maximum number of parents of any node in $\mathcal{G}$.

Corollary 6.7. *For the model $\mathcal{M}_{\mathcal{G}}^{\to}$ of a DAG $\mathcal{G}$, we have*

$$\mathrm{mlt_b}(\mathcal{G}) = \mathrm{mlt_e}(\mathcal{G}) = \mathrm{mlt_u}(\mathcal{G}) = \textit{in-degree}(\mathcal{G}) + 1.$$

Proof. The equivalence of the likelihood thresholds $\mathrm{mlt_b}$ and $\mathrm{mlt_e}$ follows from Theorem 6.3, where we also see that for the MLE to exist generically we need that every row in a generic matrix of samples $Y \in \mathbb{R}^{m\times n}$ is not a linear combination of its parent rows. Generic linear independence is guaranteed if and only if the number of columns n is at least the number of rows involved in a node plus its parents. When the MLE exists it is generically unique, as can be seen from Theorem 6.3. □

Example 6.8. Let $\mathcal{G}$ be the DAG $1 \leftarrow 3 \rightarrow 2$ from Example 6.2. We apply Theorem 6.5 to establish when the MLE given a sample matrix $Y \in \mathbb{R}^{3\times n}$ exists. Node 3 has no parents, while nodes 1 and 2 both have node 3 as their parent. Hence the log-likelihood ℓ_Y is unbounded from above if the first or second row is a scalar multiple of the third row, or if the third row is zero, and otherwise the MLE given Y exists. When $n = 1$, the first and second rows are always scalar multiples of the third row; hence, the log-likelihood is always unbounded from above. With $n = 2$ samples, the scalar multiple condition means

$$y_{11}y_{32} - y_{12}y_{31} = 0 \quad \text{and} \quad y_{21}y_{32} - y_{22}y_{31} = 0.$$

For generic $Y \in \mathbb{R}^{3\times 2}$, these equations do not vanish, and the MLE given Y exists. Hence the ML threshold is $\mathrm{mlt_e}(\mathcal{G}) = \mathrm{mlt_b}(\mathcal{G}) = \mathrm{mlt_u}(\mathcal{G}) = 2$.

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

Example 6.9. Let $\mathcal{G}$ be the TDAG $1 \to 3 \leftarrow 2$. The group $G(\mathcal{G})$ consists of invertible matrices

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

This is the transpose of the group in Examples 6.2 and 6.8, but we observe differences between the two models. Since node 3 has nodes 1 and 2 as parents, Corollary 6.7 tells us that $\mathrm{mlt_e}(\mathcal{G}) = \mathrm{mlt_b}(\mathcal{G}) = \mathrm{mlt_u}(\mathcal{G}) = 2 + 1 = 3$.

When $n = 2$, row 3 is generically a linear combination of rows 1 and 2 and the MLE does not exist. However, for special Y the MLE does exist. For example, let

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

Rows 1 and 2 are nonzero, and row 3 is not a linear combination of rows 1 and 2; hence the MLE given Y exists. Since Y is of minimal norm in its orbit, one MLE is $2I_3$, where $\lambda = 2$ minimizes $\frac{3}{2}\lambda - 3\log(\lambda)$, by Theorem 2.8. In fact, there are infinitely many MLEs, as follows. For any g in the stabilizer of Y, the vector $g \cdot Y$ is also of minimal norm in the orbit. Then $\lambda g^{\mathsf{T}} g$ is also an MLE given Y, where $\lambda = 2$ as before. The stabilizer is

$$\left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t & -t & 1 \end{bmatrix} : t \in \mathbb{R} \right\}; \quad \text{thus} \quad 2I_3 + 2t \begin{bmatrix} t & -t & 1 \\ -t & t & -1 \\ 1 & -1 & 0 \end{bmatrix}, \, t \in \mathbb{R}, \text{ are also MLEs.}$$

We can verify that these are all of the MLEs, using Theorem 2.8 or Proposition 6.6.

7. Log-Linear Models.

In this section, we describe connections between invariant theory and ML estimation for log-linear models. These results are from [3], the companion paper to [2]. Log-linear models are discrete statistical models, so they live in a probability simplex

$$\Delta_{m-1} = \left\{ p \in \mathbb{R}^m \mid p_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^{m} p_j = 1 \right\}. \tag{7.1}$$

Log-linear models have the form

$$\mathcal{M}_A = \{ p \in \Delta_{m-1} \mid \log p \in \mathrm{rowspan}(A) \}, \tag{7.2}$$

where $A \in \mathbb{Z}^{d \times m}$. We assume that the vector $\mathbb{1}$ is in the row span of A. In particular, the uniform distribution $\frac{1}{m}\mathbb{1}$ is in the model $\mathcal{M}_A$. The coordinatewise logarithm $\log p$ applies to p with strictly positive entries, and therefore $\mathcal{M}_A \subseteq \mathrm{relint}(\Delta_{m-1})$. A parametrization of the model $\mathcal{M}_A$ is

$$\begin{array}{cccc} \phi^A : & \mathbb{R}^d_{>0} & \longrightarrow & \Delta_{m-1}, \\ & \theta & \longmapsto & \left(\frac{1}{Z(\theta)} \prod_{i=1}^d \theta_i^{a_{ij}} \right)_{1 \leq j \leq m}, \end{array} \tag{7.3}$$

where Z is a normalization factor.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

Example 7.1. A probability distribution on two ternary random variables is a 3×3 matrix $p = (p_{ij})$ of nonnegative entries that sum to one. Let p_{i+} denote the sum of the ith row of p and p_{+j} the sum of the jth column. A distribution lies in the independence model if

$$p_{ij} = p_{i+}p_{+j} \quad \text{for all} \quad 1 \leq i, j \leq 3.$$

The independence model on a pair of ternary random variables is the log-linear model $\mathcal{M}_A$ for

$$A = \begin{bmatrix} 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \\ 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \end{bmatrix} \in \mathbb{Z}^{6\times 9},$$

where we require that the entries of p are strictly positive. A distribution $p \in \mathcal{M}_A$ has nine states, i.e., $\mathcal{M}_A \subseteq \Delta_8$. We identify $\mathbb{R}^9$ with $\mathbb{R}^{3\times 3}$ to view the nine state random variable as a pair of ternary random variables. Column (i, j) of the matrix A is obtained by concatenating canonical basis vectors e_i and e_j.

7.1. Maximum Likelihood Estimation for Log-Linear Models. For discrete models, data takes the form of a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$, where the coordinate u_j is the number of times that the jth state occurs and $n = u_+ := \sum_{j=1}^m u_j$ is the total number of observations. The corresponding empirical distribution is $\bar{u} = \frac{1}{n}u \in \Delta_{m-1}$, and the likelihood function is

$$L_u(p) = p_1^{u_1} \cdots p_m^{u_m}. \tag{7.4}$$

For example, if the model fills the simplex Δ_{m-1}, the likelihood is maximized uniquely at $\hat{p} = \bar{u}$. An MLE given u is a point in $\mathcal{M}$ that maximizes the likelihood (7.4) or, equivalently, that minimizes the *Kullback–Leibler* (KL) divergence to the empirical distribution $\bar{u}$. This is the discrete analogue to the settings in sections 2.1 and 2.2.

The vector $A\bar{u}$ is a vector of sufficient statistics for the model $\mathcal{M}_A$. A standard result from exponential families is that the MLE, if it exists, is the point $q \in \mathcal{M}_A$ such that

$$Aq = A\bar{u}; \tag{7.5}$$

see, e.g., [18, Proposition 2.1.5] or [42, Corollary 7.3.9]. This is sometimes known as Birch's theorem; see [39, Theorem 1.10].

Since the model $\mathcal{M}_A$ is not closed, the MLE may not exist. A modern necessary and sufficient condition for existence is as follows. The convex hull of the columns $a_j \in \mathbb{Z}^d$ of the matrix A is the polytope

$$P(A) := \operatorname{conv}\{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d.$$

The set $P(A)$ consists of vectors in $\mathbb{R}^d$ of the form Au for some $u \in \Delta_{m-1}$.

Proposition 7.2 ([42, Theorem 8.2.1]). *Let us assume that $A \in \mathbb{Z}^{d\times m}$ is such that $\mathbb{1} \in \operatorname{rowspan}(A)$ and let $\mathcal{M}_A$ be the corresponding log-linear model. Suppose we observe a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$. Then the MLE given u exists in $\mathcal{M}_A$ if and only if $A\bar{u}$ lies in the relative interior of the polytope $P(A)$.*

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

We see that if u has all entries positive, the MLE always exists. However, if u has some entries zero, the MLE may or may not exist. The *extended log-linear model* $\overline{\mathcal{M}_A}$ is the closure of $\mathcal{M}_A$ in the Euclidean topology on $\mathbb{R}^m$; see [31, section 4.2.3]. The extended model allows for distributions with some coordinates zero. The MLE always exists for the extended model, because it is compact and the likelihood function is continuous. If the MLE given u does not exist in $\mathcal{M}_A$, we refer to the MLE given u in the extended model $\overline{\mathcal{M}_A}$ as the *extended MLE* given u. Since the likelihood function (7.4) is strictly concave for log-linear models, the MLE is unique if it exists, and similarly for the extended MLE.

7.2. Correspondence for Log-Linear Models. To relate ML estimation to invariant theory, we need a group action. We consider the d-dimensional complex torus, denoted GT_d. We consider the action of GT_d on a complex projective space $\mathbb{P}_{\mathbb{C}}^{m-1}$, encoded by a $d \times m$ matrix of integers $A = (a_{ij})$. The torus element $\lambda = (\lambda_1, \dots, \lambda_d)$ acts on a point v in $\mathbb{P}_{\mathbb{C}}^{m-1}$ by multiplication by the diagonal matrix

$$\begin{bmatrix} \lambda_1^{a_{11}}\lambda_2^{a_{21}}\cdots\lambda_d^{a_{d1}} & & & \\ & \lambda_1^{a_{12}}\lambda_2^{a_{22}}\cdots\lambda_d^{a_{d2}} & & \\ & & \ddots & \\ & & & \lambda_1^{a_{1m}}\lambda_2^{a_{2m}}\cdots\lambda_d^{a_{dm}} \end{bmatrix}, \tag{7.6}$$

i.e., it acts on the coordinates of a point v via $v_j \mapsto \lambda_1^{a_{1j}} \cdots \lambda_d^{a_{dj}} v_j$.

A *linearization* of the action of GT_d on $\mathbb{P}^{m-1}$ is a corresponding action on the underlying m-dimensional vector space $\mathbb{C}^m$. It is given by a character of the torus, $b \in \mathbb{Z}^d$. For the linearization given by matrix $A \in \mathbb{Z}^{d\times m}$ and vector $b \in \mathbb{Z}^d$, the torus element λ acts on the vector v in $\mathbb{C}^m$ via

$$v_j \mapsto \lambda_1^{a_{1j}-b_1} \cdots \lambda_d^{a_{dj}-b_d} v_j. \tag{7.7}$$

Example 7.3. We return to Example 7.1. The action of GT_6 given by (7.6) is as follows. The torus element

$$\begin{pmatrix} \nu_1 & \nu_2 & \nu_3 & \nu_4 & \nu_5 & \nu_6 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \mu_1 & \mu_2 & \mu_3 \end{pmatrix}$$

acts on a 3×3 matrix x by multiplying the entry x_{ij} by $\prod_{k=1}^6 \nu_k^a$, where a is the column of A with index (i, j). This is the left-right action of $\mathrm{GT}_3 \times \mathrm{GT}_3$ on the space of 3×3 matrices; it sends $x_{ij} \mapsto \lambda_i \mu_j x_{ij}$.

We will prove the following equivalence.

Theorem 7.4. *Consider a vector of counts $u \in \mathbb{Z}_{\geq 0}^m$ with sample size $u_+ = n$, matrix $A \in \mathbb{Z}^{d\times m}$ with $\mathbb{1} \in \mathbb{C}^m$ in the row span, and vector $b = Au \in \mathbb{Z}^d$. The stability under the action of the complex torus GT_d given by matrix nA with linearization b is related to ML estimation in $\mathcal{M}_A$ as follows:*

(a)	$\mathbb{1}$ *unstable*		*does not happen,*
(b)	$\mathbb{1}$ *semistable*	$\Leftrightarrow$	*extended MLE exists and is unique,*
(c)	$\mathbb{1}$ *polystable*	$\Leftrightarrow$	*MLE exists and is unique,*
(d)	$\mathbb{1}$ *stable*		*does not happen.*

7.3. The Hilbert–Mumford Criterion. For a torus action, there is a polyhedral translation of the different concepts of stability. Recall that $P(A)$ is the polytope

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

obtained by taking the convex hull of the columns $a_j \in \mathbb{Z}^d$ of the matrix A. We can also define subpolytopes for $v \in \mathbb{R}^m$ that depend on the support $\operatorname{supp}(v) := \{j : v_j \neq 0\} \subseteq [m]$. Namely, we define

$$P_v(A) := \operatorname{conv}\{a_j \mid j \in \operatorname{supp}(v)\}.$$

For a polytope $P \subseteq \mathbb{R}^d$, we denote its interior by $\operatorname{int}(P)$ and its relative interior by $\operatorname{relint}(P)$.

Theorem 7.5 (Hilbert–Mumford criterion for a torus). *Let $v \in \mathbb{C}^m$ and consider the action of the complex torus GT_d on $\mathbb{C}^m$ given by matrix $A \in \mathbb{Z}^{d\times m}$ with linearization $b \in \mathbb{Z}^d$. We have*

$$\begin{array}{llcl} \text{(a)} & v \textit{ unstable} & \Leftrightarrow & b \notin P_v(A), \\ \text{(b)} & v \textit{ semistable} & \Leftrightarrow & b \in P_v(A), \\ \text{(c)} & v \textit{ polystable} & \Leftrightarrow & b \in \operatorname{relint}(P_v(A)), \\ \text{(d)} & v \textit{ stable} & \Leftrightarrow & b \in \operatorname{int}(P_v(A)). \end{array}$$

For an elementary proof, see [3, Appendix A]. Armed with the Hilbert–Mumford criterion for a torus, we can now prove Theorem 7.4.

Proof of Theorem 7.4. We refer to the conditions for the different notions of stability from the Hilbert–Mumford criterion in Theorem 7.5. By Proposition 7.2, the MLE of u exists if and only if b lies in the relative interior of the polytope $P(nA)$, which is the condition for polystability in Theorem 7.5.

It remains to see that the cases of instability and stability do not occur. The all-ones vector $\mathbb{1}$ can never be unstable with respect to the action in Theorem 7.4, because $b = Au$ is in the polytope $P(nA)$. Finally, the stable case also cannot arise, due to the assumption that the vector $\mathbb{1}$ lies in the row span of A, as follows. Writing $\mathbb{1}$ as a linear combination of the rows, i.e., $r^{\mathsf{T}}A = \mathbb{1}$, we have that all vectors a_j lie on the hyperplane $r_1x_1 + \cdots + r_dx_d = 1$ and the polytope $P(A)$ has empty interior in $\mathbb{R}^d$. ∎

7.4. MLE via Norm Minimization.

We have encountered two optimization problems: finding the MLE in a log-linear model, and norm minimization in an orbit under a related torus action. One problem attains its optimum if and only if the other one does, by Theorem 7.4. We now relate these optima via the moment map. As in the Gaussian case, we see that the norm minimizer gives the MLE.

Theorem 7.6. *Let $u \in \mathbb{Z}^m_{\geq 0}$ be a vector of counts with $u_+ = n$. Consider a matrix $A \in \mathbb{Z}^{d\times m}$ with $\mathbb{1} \in \operatorname{rowspan}(A)$, and let $b = Au \in \mathbb{Z}^d$. Consider the orbit closure of $\mathbb{1}$ under the torus action of GT_d given by matrix nA with linearization b. Let $q \in \mathbb{C}^m$ be a point in the orbit closure where the moment map μ vanishes. Then the extended MLE given u for the model $\mathcal{M}_A$ has jth entry*

$$\frac{|q_j|^2}{\|q\|^2}. \tag{7.8}$$

If $\mathbb{1}$ is polystable, then this vector is the MLE.

Theorem 7.6 shows that the MLE can be obtained from norm minimization on an orbit. It suggests using algorithms from invariant theory to compute the MLE; see [3, section 5]. We provide a proof in [3, Theorem 4.7]. The main ingredient is the Kempf–Ness theorem for torus actions, which we now describe.

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

Theorem 7.7 (Kempf–Ness theorem for a torus). *Consider the torus action of* GT_d *given by matrix* $A \in \mathbb{Z}^{d\times m}$ *with linearization* $b \in \mathbb{Z}^d$*. Let* $v^{(2)}$ *be the vector with* j*th coordinate* $|v_j|^2$*. The moment map is*

$$\begin{array}{cccc} \mu: & \mathbb{C}^m & \longrightarrow & \mathbb{C}^d \\ & v & \longmapsto & 2(Av^{(2)} - \|v\|^2 b). \end{array}$$

Hence a vector is semistable (resp., polystable) if and only if there is a nonzero v *in its orbit closure (resp., orbit) with* $Av^{(2)} = \|v\|^2 b$*. This* v *is a vector of minimal norm in the orbit closure (resp., orbit).*

Two proofs are given in [3, Appendix B]. We end this section with an example.

Example 7.8. Consider the log-linear model $\mathcal{M}_A$ and vector of counts u, where

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix}, \qquad u = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \qquad b = Au = \begin{bmatrix} 5 \\ 3 \end{bmatrix}.$$

The existence of the MLE given u in $\mathcal{M}_A$ can be characterized by the torus action given by matrix nA with linearization b, by Theorem 7.4, where $n = u_+ = 4$. Since b is a positive combination of the columns of A, the vector $\mathbb{1}$ is polystable under this action and the MLE given u exists. The MLE relates to a point of minimal 2-norm in the orbit of $\mathbb{1}$ under the torus action given by matrix nA with linearization b, by Theorem 7.6. We show how to obtain the MLE from a point q of minimal norm in the orbit of $\mathbb{1}$.

Since q lies in the orbit of $\mathbb{1}$, its entries are $q_j = \lambda_1^{na_{1j}-b_1}\lambda_2^{na_{2j}-b_2}$, where λ_i are nonzero complex numbers; i.e., $q^\mathsf{T} = \begin{bmatrix} \lambda^3 & \lambda^{-1} & \lambda^{-5} \end{bmatrix}$, where $\lambda = \frac{\lambda_1}{\lambda_2}$. Moreover, the moment map vanishes at q, so we have $nAq^{(2)} = \|q\|^2 b$. Combining these gives the condition $3\nu^2 - \nu - 5 = 0$, where $\nu = |\lambda|^8$, and we obtain that the MLE is

$$\hat{p} = \frac{1}{\nu^2+\nu+1}\begin{bmatrix} \nu^2 \\ \nu \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{31+\sqrt{61}}{4\sqrt{61}+52} \\ \frac{3+3\sqrt{61}}{4\sqrt{61}+52} \\ \frac{9}{2\sqrt{61}+26} \end{bmatrix} \sim \begin{bmatrix} 0.4662 \\ 0.3175 \\ 0.2162 \end{bmatrix}.$$

8. Outlook.

In this paper, we have presented a bridge between invariant theory and maximum likelihood estimation. In this section, we discuss consequences of this connection and comment on future directions. The statistical models in this paper are multivariate Gaussians or log-linear models. There are intriguing parallels between the Gaussian and log-linear settings (compare Theorems 3.3 and 7.4, for instance). However, as of yet we have no unified description of the continuous and discrete settings. Our models are instances or submodels of exponential families. By considering group actions on these models, there is a natural connection to exponential transformation families; see [40, Remark 9.2.2]. It is natural to wonder whether our dictionary can be extended to the more general framework of exponential families.

We have presented two settings in which the full correspondence holds: complex Gaussian group models (Theorem 5.3) and DAG models (Theorem 6.5). It is an open problem to find all models for which the full correspondence holds and to find a proof technique that generalizes the approaches from these two cases.

Several groups whose Gaussian group models are of statistical interest are nonreductive, such as DAG models on transitive DAGs. While invariant theory traditionally focuses on reductive groups, there has been increasing interest in extensions to

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

nonreductive groups. It is an open problem to relate our correspondence, and the topological notions of stability from Definition 2.9, to the recent developments in nonreductive geometric invariant theory [8]. This connects to recent work using other algebro-geometric notions for ML estimation [7].

A consequence of our bridge between invariant theory and ML estimation is the resolution of the problem of obtaining ML thresholds for matrix and tensor normal models [13, 14]; see section 4.2. The study of ML thresholds is an ongoing area of study with several open problems, such as the determination of the thresholds for RDAG models from [35]. Perhaps connections to invariant theory can be used to find such thresholds.

Since the publication of [2], further invariant theoretic tools have been used to gain statistical insights into Gaussian group models. The paper [28] uses Vinberg theory to classify Gaussian group models that are convex in the classical sense. Moreover, it computes the uniqueness ML threshold of such models and gives an explicit rational function for the MLE.

If a group $G \subseteq \mathrm{GL}_m$ is Zariski closed and self-adjoint, then the Gaussian group model $\mathcal{M}_G$ is a geodesically convex submanifold of PD_m. Hence one can use geodesically convex methods, such as those of [10, 27], to compute the MLE. Geodesic convexity was used in the case of matrix and tensor normal models in [22]. Given a concentration matrix Ψ in such a model and a tuple of i.i.d. samples Y, the results [22, Theorems 2.4 and 2.7] bound the error between Ψ and the MLE $\hat{\Psi}$ given Y. These bounds are optimal up to logarithmic factors in the sample size n. Furthermore, [22, Theorems 2.9 and 2.10] show that the flip-flop algorithm [19, 32] for matrix and tensor normal models efficiently computes the MLE with high probability. It is natural to ask whether the results of [22] can be generalized from matrix and tensor normal models to all Gaussian group models with Zariski closed and self-adjoint groups; see [40, Problem 9.6.2]. Here, one may consider geodesically convex methods from [10, 27] as a replacement for the flip-flop algorithm.

We hope this paper encourages further study on the interplay between invariant theory and statistics, uncovering new connections between them.

Acknowledgments. We are grateful to Peter Bürgisser, Mathias Drton, Bernd Sturmfels, and Michael Walter for fruitful discussions. We also thank Jan Draisma, Visu Makam, Nikolay Nikolov, Panagiotis Papazoglou, and Piotr Zwiernik for useful hints and suggestions.

REFERENCES

[1] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson, *Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing*, in STOC'18–Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2018, pp. 172–181, https://doi.org/10.1145/3188745.3188942. (Cited on p. 724)

[2] C. Améndola, K. Kohn, P. Reichenbach, and A. Seigal, *Invariant theory and scaling algorithms for maximum likelihood estimation*, SIAM J. Appl. Algebra Geom., 5 (2021), pp. 304–337, https://doi.org/10.1137/20M1328932. (Cited on pp. 728, 729, 730, 731, 737, 738, 739, 740, 745)

[3] C. Améndola, K. Kohn, P. Reichenbach, and A. Seigal, *Toric invariant theory for maximum likelihood estimation in log-linear models*, Algebr. Stat., 12 (2021), pp. 187–211, https://doi.org/10.2140/astat.2021.12.187. (Cited on pp. 740, 743, 744)

[4] S. Andersson, *Invariant normal models*, Ann. Statist., 3 (1975), pp. 132–154. (Cited on p. 723)

[5] S. Andersson, D. Madigan, and M. Perlman, *On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs*, Scand. J. Stat., 24 (1997), pp. 81–102, https://doi.org/10.1111/1467-9469.t01-1-00050. (Cited on p. 734)

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

[6] O. Barndorff-Nielsen, P. Blaesild, J. L. Jensen, and B. Jørgensen, *Exponential transformation models*, Proc. Roy. Soc. London Ser. A, 379 (1982), pp. 41–65, https://doi.org/10.1098/rspa.1982.0004. (Cited on p. 723)

[7] G. Bérczi, E. Hamilton, P. Reichenbach, and A. Seigal, *Complete Collineations for Maximum Likelihood Estimation*, preprint, https://arxiv.org/abs/2311.03329, 2023. (Cited on p. 745)

[8] G. Bérczi and F. Kirwan, *Non-reductive geometric invariant theory and hyperbolicity*, Invent. Math., 235 (2024), pp. 81–127. (Cited on p. 745)

[9] D. I. Bernstein, S. Dewar, S. J. Gortler, A. Nixon, M. Sitharam, and L. Theran, *Maximum likelihood thresholds via graph rigidity*, Ann. Appl. Probab., 34 (2024), pp. 3288–3319. (Cited on p. 725)

[10] P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, *Towards a Theory of Non-commutative Optimization: Geodesic First and Second Order Methods for Moment Maps and Polytopes*, preprint, https://arxiv.org/abs/1910.12375, 2019. (Cited on pp. 723, 724, 728, 732, 733, 745)

[11] P. Bürgisser, A. Garg, R. Oliveira, M. Walter, and A. Wigderson, *Alternating Minimization, Scaling Algorithms, and the Null-Cone Problem from Invariant Theory*, preprint, https://arxiv.org/abs/1711.08039, 2017. (Cited on p. 724)

[12] H. Derksen and V. Makam, *Polynomial degree bounds for matrix semi-invariants*, Adv. Math., 310 (2017), pp. 44–63, https://doi.org/10.1016/j.aim.2017.01.018. (Cited on p. 724)

[13] H. Derksen and V. Makam, *Maximum likelihood estimation for matrix normal models via quiver representations*, SIAM J. Appl. Algebra Geom., 5 (2021), pp. 338–365, https://doi.org/10.1137/20M1369348. (Cited on pp. 723, 731, 745)

[14] H. Derksen, V. Makam, and M. Walter, *Maximum likelihood estimation for tensor normal models via castling transforms*, Forum Math. Sigma, 10 (2022), art. e50, https://doi.org/10.1017/fms.2022.37. (Cited on pp. 723, 731, 745)

[15] J. Draisma, S. Kuhnt, and P. Zwiernik, *Groups acting on Gaussian graphical models*, Ann. Statist., 41 (2013), pp. 1944–1969, https://doi.org/10.1214/13-AOS1130. (Cited on p. 723)

[16] M. Drton, C. Fox, A. Käufl, and G. Pouliot, *The maximum likelihood threshold of a path diagram*, Ann. Statist., 47 (2019), pp. 1536–1553, https://doi.org/10.1214/18-AOS1724. (Cited on pp. 725, 739)

[17] M. Drton, S. Kuriki, and P. Hoff, *Existence and uniqueness of the Kronecker covariance MLE*, Ann. Statist., 49 (2021), pp. 2721–2754, https://doi.org/10.1214/21-aos2052. (Cited on p. 731)

[18] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, Oberwolfach Semin. 39, Birkhäuser Basel, 2009, https://doi.org/10.1007/978-3-7643-8905-5. (Cited on p. 741)

[19] P. Dutilleul, *The MLE algorithm for the matrix normal distribution*, J. Stat. Comput. Simul., 64 (1999), pp. 105–123. (Cited on pp. 724, 745)

[20] M. L. Eaton, *Group Invariance in Applications in Statistics*, NSF-CBMS Regional Conf. Ser. Probab. Statist. 1, Institute of Mathematical Statistics and American Statistical Association, 1989. (Cited on p. 723)

[21] R. A. Fisher, *Two new properties of mathematical likelihood*, Proc. Roy. Soc. Lond. Ser. A, 144 (1934), pp. 285–307. (Cited on p. 722)

[22] C. Franks, R. Oliveira, A. Ramachandran, and M. Walter, *Near Optimal Sample Complexity for Matrix and Tensor Normal Models via Geodesic Convexity*, preprint, https://doi.org/10.48550/arXiv.2110.07583, 2021. (Cited on p. 745)

[23] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson, *A deterministic polynomial time algorithm for non-commutative rational identity testing*, in 57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016, IEEE Computer Soc., Los Alamitos, CA, 2016, pp. 109–117, https://doi.org/10.1109/FOCS.2016.95. (Cited on p. 724)

[24] A. Garg and R. Oliveira, *Recent progress on scaling algorithms and applications*, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS, (2018), pp. 14–49, http://eatcs.org/beatcs/index.php/beatcs/article/view/533; arXiv version, https://arxiv.org/abs/1808.09669. (Cited on p. 733)

[25] N. Goodman, *Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)*, Ann. Math. Statist., 34 (1963), pp. 152–177, https://doi.org/10.1214/aoms/1177704250. (Cited on p. 734)

[26] D. Hilbert, *Über die vollen Invariantensysteme*, Math. Ann., 42 (1893), pp. 313–373, https://doi.org/10.1007/BF01444162. (Cited on p. 723)

[27] H. Hirai, H. Nieuwboer, and M. Walter, *Interior-point methods on manifolds: Theory and applications*, in 2023 IEEE 64th Annual Symposium on Foundations of Computer

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Science (FOCS), IEEE, 2023, pp. 2021–2030, https://doi.org/10.1109/FOCS57990.2023.00123; detailed version available from https://arxiv.org/abs/2303.04771. (Cited on p. 745)

[28] H. Ishi, *On Gaussian group convex models*, in Geometric Science of Information, Lecture Notes in Comput. Sci. 12829, Springer, Cham, 2021, pp. 256–264, https://doi.org/10.1007/978-3-030-80209-7_29. (Cited on p. 745)

[29] G. Ivanyos, Y. Qiao, and K. Subrahmanyam, *Constructive non-commutative rank computation is in deterministic polynomial time*, Comput. Complexity, 27 (2018), pp. 561–593, https://doi.org/10.1007/s00037-018-0165-7. (Cited on p. 724)

[30] G. Kempf and L. Ness, *The length of vectors in representation spaces*, in Algebraic Geometry (Proc. Summer Meeting, Univ. Copenhagen, Copenhagen, 1978), Lecture Notes in Math. 732, Springer, Berlin, 1979, pp. 233–243, https://doi.org/10.1007/BFb0066647. (Cited on pp. 722, 732)

[31] S. Lauritzen, *Graphical Models*, Oxford Statist. Sci. Ser. 17, Clarendon Press, 1996. (Cited on pp. 739, 742)

[32] N. Lu and D. Zimmerman, *The likelihood ratio test for a separable covariance matrix*, Stat. Probab. Lett., 73 (2005), pp. 449–457, https://doi.org/10.1016/j.spl.2005.04.020. (Cited on p. 745)

[33] J. Madsen, *Invariant normal models with recursive graphical Markov structure*, Ann. Statist., 28 (2000), pp. 1150–1178. (Cited on p. 723)

[34] A. Mai, F. Bastin, and M. Toulouse, *On Optimization Algorithms for Maximum Likelihood Estimation*, Research Report, CIRRELT-2014-64, CIRRELT, 2014. (Cited on p. 723)

[35] V. Makam, P. Reichenbach, and A. Seigal, *Symmetries in directed Gaussian graphical models*, Electron. J. Stat., 17 (2023), pp. 3969–4010, https://doi.org/10.1214/23-EJS2192. (Cited on pp. 727, 737, 738, 739, 745)

[36] G. Mostow, *Self-adjoint groups*, Ann. of Math. (2), 62 (1955), pp. 44–55, https://doi.org/10.2307/2007099. (Cited on p. 729)

[37] D. Mumford, *Stability of projective varieties*, Enseign. Math. (2), 23 (1977), pp. 39–110. (Cited on p. 727)

[38] I. Myung, *Tutorial on maximum likelihood estimation*, J. Math. Psych., 47 (2003), pp. 90–100, https://doi.org/10.1016/S0022-2496(02)00028-7. (Cited on p. 723)

[39] L. Pachter and B. Sturmfels, eds., *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005, https://doi.org/10.1017/CBO9780511610684. (Cited on p. 741)

[40] P. Reichenbach, *Invariant Theory in Computational Complexity and Algebraic Statistics*, Ph.D. thesis, TU Berlin, 2023, https://doi.org/10.14279/depositonce-18306. (Cited on pp. 725, 727, 729, 733, 744, 745)

[41] R. Richardson and P. Slodowy, *Minimum vectors for real reductive algebraic groups*, J. London Math. Soc. (2), 42 (1990), pp. 409–429, https://doi.org/10.1112/jlms/s2-42.3.409. (Cited on p. 732)

[42] S. Sullivant, *Algebraic Statistics*, Grad. Stud. Math. 194, AMS, 2018, https://doi.org/10.1090/gsm/194. (Cited on pp. 723, 724, 725, 737, 741)

[43] N. Wallach, *Geometric Invariant Theory: Over the Real and Complex Numbers*, Universitext, Springer, 2017, https://doi.org/10.1007/978-3-319-65907-7. (Cited on p. 736)

[44] R. Wooding, *The multivariate distribution of complex normal variables*, Biometrika, 43 (1956), pp. 212–215, https://doi.org/10.1093/biomet/43.1-2.212. (Cited on p. 734)

Downloaded 11/07/24 to 130.44.166.84 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license