ALGEBRAIC AND TENSOR METHODS IN STATISTICS AND OPTIMIZATION

JULIA LINDBERG

Abstract. This is a preliminary report on topics discussed during the semester program *AGATES: Algebraic Geometry with Applications to Tensors and Secants.* The purpose of this report is to discuss the method of moments and identifiability of Gaussian mixture models. The proof of Theorem 3.2 was given by Fulvio Gesmundo and Nick Vannieuwenhoven.

1. Introduction

A fundamental problem in statistics is to estimate a density from samples. This problem is called *density estimation* and formally it asks, "Given n samples from an unknown distribution p, can we estimate p"? Often, one assumes the density p lies in a parameterized family of densities. One family, known as Gaussian mixture models, are a popular choice due to their broad expressive power.

Theorem 1.1. [5, Chapter 3] *A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model with enough components.*

Theorem 1.1 is a major motivation of the study of Gaussian mixture models. These statistical models appear in many applications from the modeling of geographic events [15], the spread of COVID-19 [13], the design of planar steel frame structures [4], speech recognition [11, 14, 12], image segmentation [3] and biometrics [7].

A *Gaussian random variable*, $X \sim \mathcal{N}(\mu, \sigma^2)$, has a probability density function

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma \in \mathbb{R}_{>0}$ is the standard deviation. A random variable X is the *mixture of k Gaussians* if its probability density function is the convex combination of k Gaussian densities. Here we write $X \sim \sum_{\ell=1}^{k} \lambda_\ell \mathcal{N}(\mu_\ell, \sigma_\ell^2)$ where $\mu_\ell \in \mathbb{R}$, $\sigma_\ell \in \mathbb{R}_{>0}$ for all $\ell \in [k] = \{1, \ldots, k\}$ and $(\lambda_1, \ldots, \lambda_k) \in \Delta_{k-1} = \{\lambda \in \mathbb{R}^k_{>0} : \sum_{i=1}^{k} \lambda_i = 1\}$. Each λ_ℓ, $\ell \in [k]$, is the mixture weight of the ℓth component.

The following problem is fundamental in statistics.

Problem 1.2. Given samples, $y_1, \ldots, y_N$, distributed as the mixture of k Gaussian densities, recover the parameters $\mu_i, \sigma_i^2, \lambda_i$ for $i \in [k]$.

Problem 1.2 is called *parameter recovery for Gaussian mixture models*. There are many techniques to address this problem, this note focuses on one technique known as the method of moments.

2. Method of moments

One method for parameter recovery is the *method of moments*. This is a classical approach for parameter estimation based on the law of large numbers. This approach uses the fact that the moments of parameterized families of densities are functions of these parameters.

Let $f : \mathbb{R} \to \mathbb{R}$ be the probability density function of a random variable X. For $r \geq 0$, the *$r-th$ moment of X* is

$$m_i = \mathbb{E}[X^r] = \int_{\mathbb{R}} x^r f(x) dx.$$

We consider a statistical model with n unknown parameters, $\theta = (\theta_1, \ldots, \theta_n)$, and consider the moments up to order d as functions of θ, $g_1(\theta), \ldots, g_d(\theta)$.

Assume $y_1, \ldots, y_N$ are independent samples from the same distribution. The rth *sample moment* is given by

$$\overline{m}_r = \frac{1}{N} \sum_{i=1}^{N} y_i^r.$$

The method of moments works by using samples from the statistical model to calculate sample moments $\overline{m}_1, \ldots, \overline{m}_d$, and then solve the corresponding system $\overline{m}_i = g_i(\theta)$, $i = 1, \ldots, d$, for the parameters $\theta_1, \ldots, \theta_n$.

The moments of Gaussian distributions are polynomials in the variables μ, σ^2 and can be calculated recursively as $M_0(\mu, \sigma^2) = 1, M_1(\mu, \sigma^2) = \mu$ and

$$M_i(\mu, \sigma^2) = \mu M_{i-1} + (i-1)\sigma^2 M_{i-2}, \quad i \geq 2. \tag{2.1}$$

We calculate the ith moment of a mixture of k Gaussian densities, m_i, as the convex combinations of $M_i(\mu_1, \sigma_1^2), \ldots, M_i(\mu_k, \sigma_k^2)$:

$$m_i = \lambda_1 M_i(\mu_1, \sigma_1^2) + \ldots + \lambda_k M_i(\mu_k, \sigma_k^2). \tag{2.2}$$

Example 2.1. When $k = 2$, the first six moments of a Gaussian mixture model are:

$$\begin{aligned}
1 &= \lambda_1 + \lambda_2 \\
m_1 &= \lambda_1\mu_1 + \lambda_2\mu_2 \\
m_2 &= \lambda_1(\mu_1^2 + \sigma_1^2) + \lambda_2(\mu_2^2 + \sigma_2^2) \\
m_3 &= \lambda_1(\mu_1^3 + 3\mu_1\sigma_1^2) + \lambda_2(\mu_2^3 + 3\mu_2\sigma_2^2) \\
m_4 &= \lambda_1(\mu_1^4 + 6\mu_1^2\sigma_1^2 + 3\sigma_1^4) + \lambda_2(\mu_2^4 + 6\mu_2^2\sigma_2^2 + 3\sigma_2^4) \\
m_5 &= \lambda_1(\mu_1^5 + 10\mu_1^3\sigma_1^2 + 15\mu_1\sigma_1^4) + \lambda_2(\mu_2^5 + 10\mu_2^3\sigma_2^2 + 15\mu_2\sigma_2^4)
\end{aligned}$$

Using (2.1), we also define the *Gaussian moment variety* $\mathcal{G}_{1,d} \subset \mathbb{P}^d$ to be the projective closure of the image of the map

$$\begin{aligned}
\phi : \mathbb{C}^2 &\to \mathbb{C}^d \\
(\mu, \sigma^2) &\mapsto (M_1, \ldots, M_d)
\end{aligned}$$

where M_i is as defined in (2.1). This variety was first defined in [1] and if we restrict in source of ϕ to $\mathbb{R} \times \mathbb{R}_{>0}$, it parameterizes the first d moments of all Gaussian mixture densities.

3. Identifiability

A fundamental question when applying the method of moments is how many moments does one need to uniquely recover the parameters of the density? This is the question of *identifiability* and is a fundamental question in statistics. It was shown in [9] that every Gaussian k-mixture model is identifiable using moments $m_1, \ldots, m_{4k-2}$. Moreover, there exist Gaussian mixture models where one needs moments up to order $4k-2$ to identify it. For example, consider a Gaussian k mixture model where all means $\mu_1 = \mu_2 = \ldots = \mu_k = 0$. In this case, all odd order moments are identically zero and one needs moments of order $4k-2$ to uniquely identify this Gaussian mixture model.

A looser question than asking how many moments one needs to uniquely every Gaussian mixture model, is to ask how many moments one needs to identify a *generic* Gaussian mixture model. We provide an answer to this question using the following recent result from algebraic geometry.

Theorem 3.1. [10, Theorem 1.5] *Let $X \subset \mathbb{P}^N$ be an irreducible and non-degenerate variety of dimension n, $h \geq 1$ an integer and assume that*

(1) $(h+1)n + h \leq N$
(2) X has non-degenerate Gauss map
(3) X is not $(h+1)-$defective,

then X is $h-$identifiable.

Recall that for an n dimensional projective variety $X \subset \mathbb{P}^N$, the *Gauss map* of X is a rational map from smooth points of X, $\tilde{X}$, to the Grassmanian $Gr(n, N)$ which maps a smooth point $x \in \tilde{X}$ to its tangent space T_xX:

$$\gamma : \tilde{X} - \to Gr(n, N)$$
$$x \mapsto T_xX$$

Theorem 3.2. *A generic Gaussian k mixture model is identifiable using moments $m_0, m_1, \ldots, m_{3k+2}$.*

Proof. We verify that the conditions of Theorem 3.1 hold when $X = \mathcal{G}_{1,3k+2}$. For item (1) observe that $X \subset \mathbb{P}^{3k+2}$, is two dimensional so $(h+1) \cdot 2 + h \leq 3k+2$ when $h = k$.

Next we show X has non-degenerate Gauss map. Degeneracy of projective surfaces was classified in [6, Section 2] and [8, Theorem 3.4.6], where it was shown that the Gauss map of a projective surface is degenerate if and only if it is one of the following:

(a) a linearly embedded $\mathbb{P}^2$;
(b) a cone over a curve;
(c) the tangential variety to a curve.

X is not a linearly embedded $\mathbb{P}^2$ since for $k \geq 2$, it is of degree $\binom{3k+2}{2} > 1$ [1, Corollary 2], meaning it is a non-linear surface. By [2, Theorem 1] X is not the cone over a curve. Finally, X is not the tangential variety to a curve since the singular locus of the tangential variety of a curve includes the curve. In this case, X has singular locus defined by the line $\{m_0 = 0, \ldots, m_{d-2} = 0\}$. Since the tangential variety of a line is the line itself, this means that if X were to be the tangential variety it would have to be a line, not a surface, which is a contradiction. This shows that X has non-degenerate Gauss map.

Finally, by [2, Theorem 1] X is not $(3k+3)-$defective proving the result. □

3.1. **Acknowledgements.** This work is partially supported by the Simons Foundation Award No. 663281 granted to the Institute of Mathematics of the Polish Academy of Sciences for the years 2021-2023.

References

[1] Carlos Améndola, Jean-Charles Faugère, and Bernd Sturmfels. Moment varieties of Gaussian mixtures. *J. Algebr. Stat.*, 7(1):14–28, 2016. 3, 4

[2] Carlos Améndola, Kristian Ranestad, and Bernd Sturmfels. Algebraic identifiability of Gaussian mixtures. *Int. Math. Res. Not. IMRN*, (21):6556–6580, 2018. 4

[3] Konstantinos Blekas, Aristidis Likas, Nikolaos Galatsanos, and Isaac Lagaris. A spatially constrained mixture model for image segmentation. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 16:494–8, 04 2005. 1

[4] Bach Do and Makoto Ohsaki. Gaussian mixture model for robust design optimization of planar steel frames. *Struct. Multidiscip. Optim.*, 63(1):137–160, 2021. 1

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 1

[6] Phillip Griffiths and Joseph Harris. Algebraic geometry and local differential geometry. *Ann. Sci. École Norm. Sup. (4)*, 12(3):355–452, 1979. 3

[7] Dan Hosseinzadeh and Sridhar Krishnan. Gaussian mixture modeling of keystroke patterns for biometric applications. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38:816 – 826, 12 2008. 1

[8] Thomas A. Ivey and J. M. Landsberg. *Cartan for beginners: differential geometry via moving frames and exterior differential systems*, volume 61 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. 3

[9] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of The ACM*, 55(2):113–120, 2012. 3

[10] Alex Massarenti and Massimiliano Mella. Bronowski's conjecture and the identifiability of projective varieties, 2022. 3

[11] Douglas Alan Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, 1992. 1

[12] Fei Sha and Lawrence K. Saul. Large margin Gaussian mixture modeling for phonetic classification and recognition. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages 265–268, 2006. 1

[13] Amit Singhal, Pushpendra Singh, Brejesh Lall, and Shiv Dutt Joshi. Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. *Chaos Solitons Fractals*, 138:110023, 8, 2020. 1

[14] Yaxin Zhang, Mike Alder, and Roberto Togneri. Using Gaussian mixture modeling in speech recognition. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–613. IEEE, 1994. 1

[15] Shuo Zhou, Howard Bondell, Antoinette Tordesillas, Benjamin I. P. Rubinstein, and James Bailey. Early identification of an impending rockslide location via a spatially-aided Gaussian mixture model. *Ann. Appl. Stat.*, 14(2):977–992, 2020. 1

Julia Lindberg, Max-Planck Institute for Mathematics in the Sciences, Germany
Email address: julia.lindberg@mis.mpg.de
URL: https://sites.google.com/view/julialindberg/home