Algebraic Statistics of Gaussian Mixtures

vorgelegt von
M.Sc. Mathematik
Carlos Enrique Améndola Cerón
geb. in Mexiko-Stadt

von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
- Dr.rer.nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Peter Friz
Gutachter:	Prof. Dr. Bernd Sturmfels
Gutachter:	Prof. Dr. Christian Haase
Gutachter:	Prof. Dr. Reinhold Schneider

Tag der wissenschaftlichen Aussprache: 15. November 2017

Berlin 2017

Acknowledgements

First of all, there are no words to appropriately thank my advisor, Bernd Sturmfels, for having given me this once-in-a-lifetime opportunity to be his student and having guided me with the best wisdom through this wonderful journey. It has been literally a dream come true to be able to learn from and to work with him. His passion for Math is unlimited, and an inexhaustible source of inspiration and admiration. I am also infinitely grateful to my co-advisor Christian Haase, for his continuous presence and consistent support, for welcoming me to and making me feel at home with Team Haase, and for being such an awesome and fun person.

I would like to thank my brilliant collaborators: Mathias Drton, Jean-Charles Faugère, Kristian Ranestad, Alexander Engström, Jose Rodriguez and the whole Mathematics Research Communities Likelihood group, especially Serkan Hoşten. They are just the tip of a whole iceberg of wonderful mathematicians I have had the chance to meet across the world in many exciting workshops and conferences: Caroline Uhler, Piotr Zwiernik, Shaowei Lin, Seth Sullivant, Timo de Wolff, Peter Bürgisser, Jonathan Hauenstein, Anton Leykin, Josephine Yu, Elizabeth Gross, Angélica Cueto, Diane Maclagan, Liam Solus, Jan Draisma, Emil Horobet, Rob Eggermont, Ralph Morrison, Bo Lin, Joe Kileel, Christiane Görgen, Eliana Duarte, Yue Ren, Jacinta Torres, Mario Kummer, Paul Breiding,...

Thanks to my Einstein family: to the highly talented and friendly postdocs Laura Escobar, Fatemeh Mohammadi and Marta Panizzut; and to my academic sister Kathlén Kohn who has been there from the beginning. It is always a pleasure to come to the office and find you there. I am grateful for the generous funding of the Einstein Foundation for making the whole project possible. In the same way, I am very grateful to Michael Joswig for accommodating us in the Discrete Mathematics/Geometry research group at TU Berlin. I thank each one of these colleagues and send 'comrade wishes' to Georg Loho, Benjamin Schröter and André Wagner who have also their corresponding dissertation defense in this period. Special thanks to Antje Schulz; without her help, patience and always friendly smile, it would have been impossible to sort through all the necessary paperwork/regulations and able to focus on research.

I would like to thank the Villa community at FU Berlin for the fun and interesting events (including *DiscreTea*), and specially my academic siblings Florian Kohl, Lena Walter and Jan Hofmann for uncountably many good times. I am grateful to the Berlin Mathematical School for being a very good promoter of Mathematics in Berlin and for the chance of being in such a friendly and international community.

Many special thanks to Krishnan Mody, Andra Constantinescu, Arya Tafvizi, Josué Tonelli Cueto, Anna Seigal, Elina Robeva, Kaie Kubjas, Corey and Debby Harris, and other amazing friends for all that uplifting encouragement in the end.

Very heartfelt thanks to my family: Mom, Dad, Andrea (+ Celine) and my *abuelitas*, without you I would not have even got to the first step of this voyage. Your love made missing you a little more bearable.

Last but not least, I am very grateful to Prof. Dr. Reinhold Schneider for agreeing to take the time to review this work and to Prof. Dr. Peter Friz for being the chair of the dissertation committee.

Thank you all, and may the Force be with you.

Abstract

In this work we study the statistical models known as Gaussian mixtures from an algebraic point of view.

First, we illustrate how algebraic techniques can be useful to address fundamental questions on the shape of Gaussian mixture densities, namely the problem of determining the maximum number of modes a mixture of Gaussians can have, depending on the number of components and the dimension.

We proceed to look at the statistical problem of estimation of the parameters of a Gaussian mixture. We present and compare the prominent methods of maximum likelihood estimation and moment matching, and from this study a fundamental difference in algebraic complexity is revealed between the two approaches.

With the above statistical motivations, we introduce the algebraic objects that will permit us to obtain statistical inference results, mainly on the identifiability problem. These objects are Gaussian moment varieties and their corresponding secant varieties. We study them by asking for their dimension, for their degree and for the equations that define them. We provide many answers, conjectures and open questions in this direction.

Finally, we explore further connections and analogues to algebraic geometry from commonly used submodels of statistical interest. We compare what we learn from the algebraic perspective to recent tensor decomposition methods in the machine learning community.

Throughout, we mention current research directions that continue the effort of including Gaussian densities and their mixtures in algebraic statistics.

Zusammenfassung

In dieser Arbeit studieren wir die statistischen Modelle, die als zusammengesetzte Normalverteilungen oder auch als Gaußsche Mischverteilungen bekannt sind, vom algebraischen Standpunkt aus.

Zunächst betrachten wir, wie algebraische Methoden dabei helfen können, grundlegende Fragen über die Form der Dichte einer solchen Verteilung zu klären: Zum Beispiel wollen wir die Anzahl lokaler Maxima bestimmen, die eine Mischung von Gaußverteilungen, abhängig von der Anzahl der Komponenten sowie der Dimension, höchstens haben kann.

Anschließend beschäftigen wir uns mit der statistischen Fragestellung, wie man die Modellparameter (Erwartungswerte und Varianzen der Komponenten sowie deren Gewichte) bestimmen kann. Wir vergleichen die gängigen Methoden: die Maximum-Likelihood-Methode und die Momentenmethode. Dieser Vergleich offenbart einen grundlegenden Unterschied in der algebraischen Komplexität zwischen den beiden Ansätzen.

Motiviert durch diese statistischen Erkenntnisse führen wir diejenigen algebraischen Objekte ein, die uns statistische Ergebnisse – hauptsächlich zum Identifizierbarkeitsproblem – liefern: die Gaußschen Momentenvarietäten und deren Sekantenvarietäten. Wir untersuchen ihre Dimension sowie ihren Grad und die algebraischen Gleichungen, die sie definieren. Wir beantworten viele dieser Fragen und formulieren weitere offene Fragen und Vermutungen.

Schließlich erforschen wir Verbindungen und Analogien zur algebraischen Geometrie aus gängigen Untermodellen, die von statistischer Relevanz sind. Wir vergleichen unsere Erkenntnisse aus der Perspektive der algebraischen Geometrie mit modernen Tensor-Zerlegungsmethoden aus dem Bereich des maschinellen Lernens.

Wir verweisen durchgehend auf vielversprechende Forschungsfragen, die Normalverteilungen und deren Mischverteilungen in die Algebraische Statistik miteinbeziehen.

Declaration of Authorship

I, CARLOS AMÉNDOLA, declare that this thesis titled *"Algebraic Statistics of Gaussian Mixtures"* and the work presented in it are my own or based on joint work of my own with co-authors. Parts of this thesis are prepublished.

- Chapter 2 is based on the joint paper *Maximum Number of Modes of Gaussian Mixtures* [AEH17] with Alexander Engström (Aalto University) and Christian Haase (FU Berlin). A preprint is available at `arXiv:1702.05066`.

- Chapter 3 is based on the joint paper *Maximum Likelihood Estimates for Gaussian Mixtures are Transcendental* [ADS15] with Mathias Drton (UW Seattle) and Bernd Sturmfels (MPI Leipzig and UC Berkeley). It was presented in Berlin at MACIS 2015 *International Conference on Mathematical Aspects of Computer and Information Sciences* and published as one of the revised selected papers. The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-32859-1

- Chapters 4 and 5 are based on two joint papers: the article *Moment Varieties of Gaussian Mixtures* [AFS16] with Jean-Charles Faugère (INRIA) and Bernd Sturmfels published in the *Journal of Algebraic Statistics* (available online at http://dx.doi.org/10.18409/jas.v7i1.42) and the article *Algebraic Identifiability of Gaussian Mixtures* [ARS17] with Kristian Ranestad (University of Oslo) and Bernd Sturmfels published in *International Mathematics Research Notices* (available online at: https://doi.org/10.1093/imrn/rnx090).

Contents

1. Introduction

In Algebraic Statistics, methods from Algebraic Geometry, Commutative Algebra and Combinatorics are used to address various problems in Statistics; these in turn provide these areas with new interesting questions [DS+98, PRW00, PS05, DSS08]. The study of these connections between algebra and statistics is made possible thanks to many common statistical models being described by polynomial equations and polynomial inequalities.

A particularly nice case is when the probability models are given by discrete exponential families, since these correspond geometrically to toric varieties [GMS+06, MSUZ16]. However, this discrete case does not cover the most prominent continuous distribution: the normal or Gaussian.

Definition 1.0.1. The n-dimensional multivariate *Gaussian* distribution has continuous probability density function

$$f(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{1.0.1}$$

where $x, \mu \in \mathbb{R}^n$ and Σ is a symmetric positive definite $n \times n$ matrix. We call μ the *mean vector* and Σ the *covariance matrix*.

If a random variable X follows a Gaussian distribution, we write $X \sim N(\mu, \Sigma)$. We recall one of the reasons that they are so essential in probability and statistics (see e.g. [VdV00, Section 2.3], and [BR10, Chapter 3] for generalizations):

Theorem 1.0.2 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random vectors with mean μ and covariance Σ. Then their centered scaled sum will converge in distribution as follows:*

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0, \Sigma).$$

That is, the scaled and centered sequence of sample means $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$ will follow a Gaussian distribution as $n \to \infty$: $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma)$.

While the Gaussian distribution has appeared in Algebraic Statistics, it has been mainly via the study of Gaussian graphical models with a focus on conditional independence relations, as in [U+12] and [Zwi15]. In this thesis we will add mixtures of Gaussians to the statistical models studied in Algebraic Statistics.

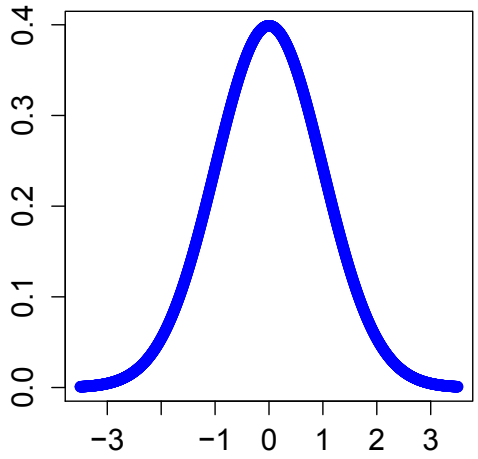

Figure 1.0.1.: Standard univariate Gaussian distribution

We now introduce the concept of a *mixture* of Gaussians by presenting its actual origin at the end of the 19th century [Pea94].

1.1. Pearson's Crabs: Algebraic Statistics in 1894

In 1894, Karl Pearson wanted to explain the asymmetry observed in data measured from a population of Naples' crabs, believing it was possible that two subpopulations of crabs were present in the sample. The corresponding statistical model is known as a Gaussian mixture; in this case a mixture of two univariate Gaussian distributions, each with its own mean and variance. In his seminal paper [Pea94, p.72], Pearson writes:

"It may happen that we have a mixture of 2,3,...,n [we will use k in place of n in what follows] *homogeneous groups, each of which deviates about its own mean symmetrically and in a manner represented with sufficient accuracy by the normal curve"*

The histogram corresponding to the observed data is shown in Figure 1.1.1.

Let us say that indeed we have $k = 2$ subpopulations of crabs distributed $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with proportions α and $1 - \alpha$ respectively, so that $0 < \alpha < 1$ (here and throughout we use the standard notation σ^2 for the variance σ_{11} when $n = 1$). Sampling from this mixture distribution is the result of tossing a biased coin with probability α of heads and then drawing from the first population

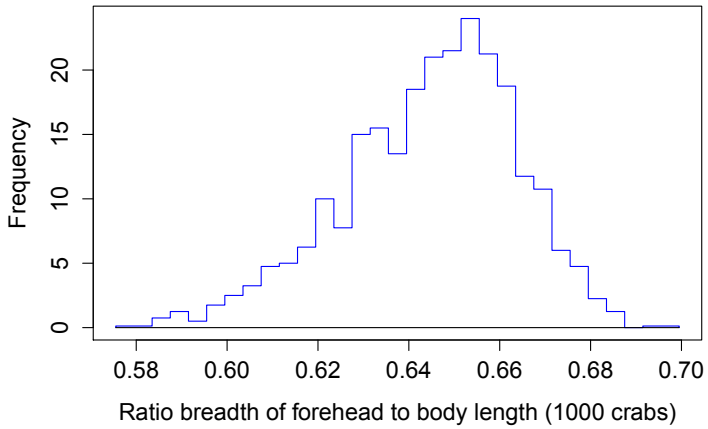

Figure 1.1.1.: Histogram for Crabs Data

if heads comes up, else drawing from the second. The probability density function of the mixture is then a *convex combination* of the individual ones. That is,

$$f_{mixt}(x) = \frac{1}{\sqrt{2\pi}}\left[\frac{\alpha}{\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-\alpha}{\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right]. \quad (1.1.1)$$

The main guiding question that Pearson faced was:

Problem 1. How can we fit a Gaussian mixture density f_{mixt} to the given data above? That is, how can we find appropriate parameters $(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2)$ so that f_{mixt} approximates the given frequency curve?

In order to solve this problem, Pearson introduced the *method of moments*, which consists of matching the model density moments to the sample moments. For example, the mean μ of the mixture density f_{mixt} can be computed as

$$\mu = \alpha\mu_1 + (1-\alpha)\mu_2.$$

This equation restricts the possible values of α, μ_1, μ_2 given μ. We can estimate the value of μ by taking the mean of the observed sample. Explicitly, if $x_1, x_2, \ldots, x_N$ are the observed values (in the crabs data, $N = 1000$) then

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1.1.2)$$

is the sample mean or first sample moment $\hat{m}_1$.

In similar fashion, Pearson proposes the following system of equations for recovering the five parameters from the first five moments (details will be given in Chapter 4):

$$\begin{aligned}
\alpha_1\mu_1 + (1-\alpha)\mu_2 &= m_1 \\
\alpha_1(\mu_1^2+\sigma_1^2) + (1-\alpha)(\mu_2^2+\sigma_2^2) &= m_2 \\
\alpha_1(\mu_1^3+3\mu_1\sigma_1^2) + (1-\alpha)(\mu_2^3+3\mu_2\sigma_2^2) &= m_3 \quad (1.1.3) \\
\alpha_1(\mu_1^4+6\mu_1^2\sigma_1^2+3\sigma_1^4) + (1-\alpha)(\mu_2^4+6\mu_2^2\sigma_2^2+3\sigma_2^4) &= m_4 \\
\alpha_1(\mu_1^5+10\mu_1^3\sigma_1^2+15\mu_1\sigma_1^4) + (1-\alpha)(\mu_2^5+10\mu_2^3\sigma_2^2+15\mu_2\sigma_2^4) &= m_5.
\end{aligned}$$

What is essential to note here is that the above system of equations is *polynomial*. This beautiful fact will allow us to use algebraic geometry to study this method for solving for the parameters.

After considerable effort and cleverness, Pearson managed to do elimination and obtain a *ninth* degree polynomial in the single unknown $x = \mu_1\mu_2$,

$$24x^9 - 28\lambda_4 x^7 + 36m_3^2x^6 - (24m_3\lambda_5 - 10\lambda_4^2)x^5 - (148m_3^2\lambda_4 + 2\lambda_5^2)x^4$$

$$+(288m_3^4 - 12\lambda_4\lambda_5 m_3 - \lambda_4^3)x^3 + (24m_3^3\lambda_5 - 7m_3^2\lambda_4^2)x^2 + 32m_3^4\lambda_4 x - 24m_3^6 = 0, \quad (1.1.4)$$

where $\lambda_4 = 9m_2^2 - 3m_4$ and $\lambda_5 = 30m_2m_3 - 3m_5$.

This is the first step in a back-substitution procedure that will yield possible solutions for the unknowns $\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2$. Indeed, Pearson can now substitute the empirical moments from his numerical crabs data into (1.1.4), find the real roots of this nonic and determine if they can correspond to a solution for the mixture model. His approach can be seen as the first instance of Algebraic Statistics.

1.2. Fisher's Approach and Hill-Climbing

Finding the roots of Pearson's nonic (1.1.4), which we will call *Pearson's polynomial*, was certainly intimidating at the time. Several potential users of this method were not enthusiastic. For instance, we have the following quote from Charlier in 1906 (see [MP04, p.3]):

"The solution of an equation of the ninth degree, where almost all powers, to the ninth, of the unknown quantity are existing, is, however, a very laborious task. Mr. Pearson has indeed possessed the energy to perform this heroic task in some instances in his first memoir on these topics from the year 1894. But I fear that he will have few successors, if the dissection of the frequency curve into two components is not very urgent."

A century later, we see how Gaussian mixture models have proven to be very

useful in a wide range of applications, from image segmentation [GS98] to speech recognition [RR95]. While only a few followed Pearson's method of moments ideas for this problem, many followed the idea of mixture modeling once an alternative method gained wide popularity: the so-called *EM algorithm* (introduced in 1977 by Dempster, Laird and Rubin [DLR77]).

To understand the EM algorithm, which stands for *Expectation-Maximization algorithm*, we go back to 1922 when Ronald Fisher had managed to convince most of the world that *maximum likelihood estimation* is the method to be preferred [A$^+$97]. This attractive approach looks for parameter values among the ones that maximize the likelihood of having observed a particular sample.

Definition 1.2.1. Let $x_1, x_2, \ldots, x_N$ be sample data points and consider the parametrized model density function $f(x;\theta)$ where $\theta \in \Theta$ is a vector of parameters. Then the *likelihood function* for the sample is:

$$L(x_1, x_2, \ldots, x_N; \theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_N;\theta) \tag{1.2.1}$$

and the *log-likelihood function* is $\ell(x_1, \ldots, x_N; \theta) = \log L(x_1, \ldots, x_N; \theta)$.

If the points $x_1, x_2, \ldots, x_N$ are drawn independently from the probability density $f(x;\theta)$ for a given parameter θ, then $L(x_1, x_2, \ldots, x_N; \theta)$ is indeed the likelihood assigned to them in the joint density $f(x_1, \ldots, x_n; \theta)$. However, we do not know θ, so the proposed approach is that we find such a $\hat{\theta}$ that will maximize the likelihood of having observed $x_1, x_2, \ldots, x_N$.

In this way, the maximum likelihood approach translates the parameter estimation problem to an optimization problem: maximizing the likelihood function, or equivalently, the log-likelihood function. This does not mean that this becomes an easy problem to solve (one could argue the opposite), but gives way to application of other techniques. In addition to desirable statistical properties such as consistency and asymptotic normality, the philosophy of maximum likelihood is equivalent to minimizing the so-called Kullback-Leibler divergence [Her05].

However, it is often the case that there are no closed form solutions for the parameters in the critical equations coming from the gradient of the likelihood function. The EM algorithm provides a way to overcome this situation. From a starting parameter value, it iteratively updates them with the very nice property that the likelihood never decreases with each step. Such local search methods are sometimes labeled 'hill-climbing'. Convergence will often lead to a local maximum [Wu83]. We will give a more precise description in Chapter 3.

For Gaussian mixtures, the EM algorithm is particularly simple to implement (see Section 3.3), and thus widely used. For instance, one could easily call the `R` packages *mclust* [FR03], *mixtools* [BCHY09] or *EMcluster* [CMM12]. An implementation for the univariate case can be found in the Appendix.

As one could expect, there are drawbacks to these methods. They usually depend heavily on the starting point of the algorithm and there is no general guarantee of converging to a global maximum. Even worse, for general Gaussian mixtures the optimization problem is ill-posed: the likelihood can be an unbounded function of the parameters! (see Remark 3.2.2)

One may ask, where is the algebra present in the maximum likelihood approach? How can tools from algebraic geometry be useful in this case? In many frequently used statistical models, including the parametric discrete exponential family and log-linear models, computing the maximum likelihood estimate $\hat{\theta}$ is equivalent to solving a system of polynomial (or rational) equations (similar to system (1.1.3)). These are usually the critical equations that one obtains from equating the gradient of the likelihood to zero $\nabla \ell(x_1, \ldots, x_N; \theta) = 0$ and trying to solve for the parameters θ. For generic data, independent of the size N, the number of solutions to these systems over the *complex* numbers is constant, and this is known as the *ML degree* of the model [DSS08, Chapter 2]. The ML degree is an intrinsic invariant of a statistical model, with interesting geometric and topological properties [HS14]. This number gives a measure of the algebraic complexity of solving the maximum likelihood estimation problem. When the ML degree is moderate, exact tools are guaranteed to find the optimal solution to the ML problem [BHSPR07, GDP+12].

However, the ML degree of a statistical model is only defined when the MLE is an algebraic function of the data, so we will need to investigate if this is the case for Gaussian mixtures, and we do so in Chapter 3. We refer to [CHKS06, ABB+17] for more on ML degree.

1.3. Motivating Questions

"The equations for the dissection of a frequency-curve into k normal curves can be written down in the same manner as for the special case of $k = 2$ treated in this paper; they require us only to calculate higher moments. But the analytical difficulties, even for the case of $k = 2$, are so considerable, that it may be questioned whether the general theory could ever be applied in practice to any numerical case."

-Karl Pearson [Pea94]

In this work we will address the following natural (and perhaps ambitious, if we judge by the above quote) questions:

Problem 2. How does Pearson's method generalize for a mixture of $k > 2$ Gaussians? What about Gaussians in higher dimension $n > 1$?

Problem 3. How many moments are needed to recover the parameters for general n, k? Is there an analogous polynomial to Pearson's polynomial (1.1.4)? If so, what is its degree?

Problem 4. Can we analyze algebraically Fisher's method (maximum likelihood) for mixtures of Gaussians? Specifically, is there an ML degree that indicates the algebraic complexity of maximum likelihood estimation for Gaussian mixtures?

1.3.1. The Chapters

In Chapter 2 we aim to better understand the shape of Gaussian mixture densities, statistical inference apart. We know that a single Gaussian has a unique global maximum: its *mode*. Certainly, Gaussian mixtures with a common mean vector will continue to have only one global maximum. But in general the number of modes (local maxima) can grow depending on the Gaussian components of the mixture. Can there be more modes than the number of components k? Is there a maximum number of modes, maybe as a function that depends on n, k? We describe what we know in this chapter.

In Chapter 3 we focus on Problem 4. We will 'spoil' early the main result: the maximum likelihood estimates for Gaussian mixtures are transcendental functions of the data. This means that the parameter estimation problem transcends in a way the algebraic geometric techniques. As a consequence there will be no ML degree but we will ask if there is an analogous number that could bound the number of critical points in the likelihood function.

Starting from the results obtained in the previous chapter, in Chapter 4 we revisit the idea of Pearson's method of moments and study his approach from a modern perspective. To lay the groundwork for partial solutions to Problems 2 and 3, we introduce *moment varieties* and study their properties. They encode the moments of a single Gaussian distribution.

In Chapter 5 we go to the natural next step and study the secant varieties of moment varieties, since they correspond to moments of Gaussian mixture distributions. We study the first instances of both the one-dimensional case $n = 1$ with $k > 1$, and the higher dimensional case when $n > 1$. Our results contrast both situations in terms of identifiability of Gaussian mixtures from their moments up to a certain order.

In Chapter 6 we describe the situation for some special submodels of Gaussian mixtures with restricted structure; in particular, homoscedastic Gaussian mixtures in which the components share a fixed covariance matrix. We explain how this connects algebraically in a strong way to the celebrated Alexander-Hirschowitz theorem, and we compare to recent techniques in the Machine Learning literature.

Finally, a conclusion briefly summarizes the main results and contributions of this work. This includes the current status of the three problems presented at the beginning of this section and some open questions that remain.

2. The Peaks of Mixture Densities

We begin by recalling the probability density function of an n-dimensional Gaussian mixture with k components.

Given k mean vectors $\mu_i \in \mathbb{R}^n$, $n \times n$ positive definite covariance matrices Σ_i, and mixture weights $\alpha_i > 0$, $i = 1, \ldots, k$ we have

$$f(x) = \sum_{i=1}^{k} \alpha_i f_i(x), \tag{2.0.1}$$

where $\alpha_1 + \cdots + \alpha_k = 1$ and

$$f_i(x) = \frac{1}{\sqrt{\det(2\pi\Sigma_i)}} \mathrm{e}^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}. \tag{2.0.2}$$

A fundamental property of a probability density function is the number of modes, i.e. local maxima, that it possesses. For Gaussian mixtures, this is especially relevant in applications such as clustering [HMMR15]. For example, the *mean shift algorithm* converges if there are only finitely many critical points [Wal13].

We will be interested in the maximal number $m(n, k)$ of local maxima for n-dimensional Gaussian mixtures with k components. Shockingly, it is not known whether this maximal number is always finite for general Gaussian mixtures. The main results of this chapter will be giving a lower bound (Theorem 2.3.1) and an upper bound (Theorem 2.4.4) on this number.

Remark 2.0.1. Since a single Gaussian has a unique global maximum at its mean μ, we have that $m(n, 1) = 1$ for all $n \geq 1$.

2.1. Background

The simplest case when there is actually a mixture has $n = 1$ and $k = 2$: a mixture of two univariate Gaussians $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$, with mixture parameter $\alpha \in (0, 1)$. It was observed historically that in this scenario the number of modes was either 1 or 2, with the following heuristics:

- If the distance between the component means is small, then the mixture is unimodal (independently of α).

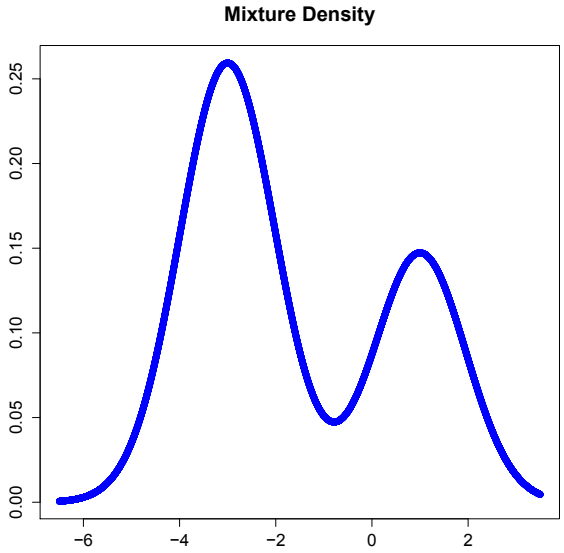

Figure 2.1.1.: A mixture of two univariate Gaussians with 2 modes

- If the distance between the component means is large enough, then there is bimodality unless α is close to 0 or 1.

A.C. Cohen (1953) and Eisenberger (1964) obtained some first explicit conditions in these directions [Eis64]. Notably, if $\alpha = \frac{1}{2}$ and $\sigma_1 = \sigma_2 = \sigma$, then the mixture is unimodal (with mode at $\frac{\mu_1+\mu_2}{2}$) if and only if $|\mu_2 - \mu_1| \leq 2\sigma$. A few years later, J. Behboodian [Beh70] gives a proof that indeed $m(1,2) = 2$ by showing the number of critical points is at most three, and finds that

$$|\mu_2 - \mu_1| \leq 2\min(\sigma_1, \sigma_2)$$

is a sufficient condition for unimodality. Furthermore, if $\sigma_1 = \sigma_2 = \sigma$, then

$$|\mu_2 - \mu_1| \leq 2\sigma\sqrt{1 + \frac{|\log(\alpha) - \log(1-\alpha)|}{2}}$$

is again a sufficient condition for having only one mode in the mixture. Starting the 21st century, it was Carreira-Perpiñán and Williams who had particular interest in the problem [CPW03b]. Using scale-space theory, they prove that $m(1,k) = k$; any univariate Gaussian mixture with k components has at most k modes.

A natural conjecture is that $m(n,k) = k$ for all n, k. However, this fails already when $n = k = 2$, since a mixture of two bivariate Gaussians can have 3 distinct modes (actually, we'll see shortly that $m(2,2) = 3$).

Remark 2.1.1. In June 2016, a discussion thread on the `ANZstat` mailing list (e-

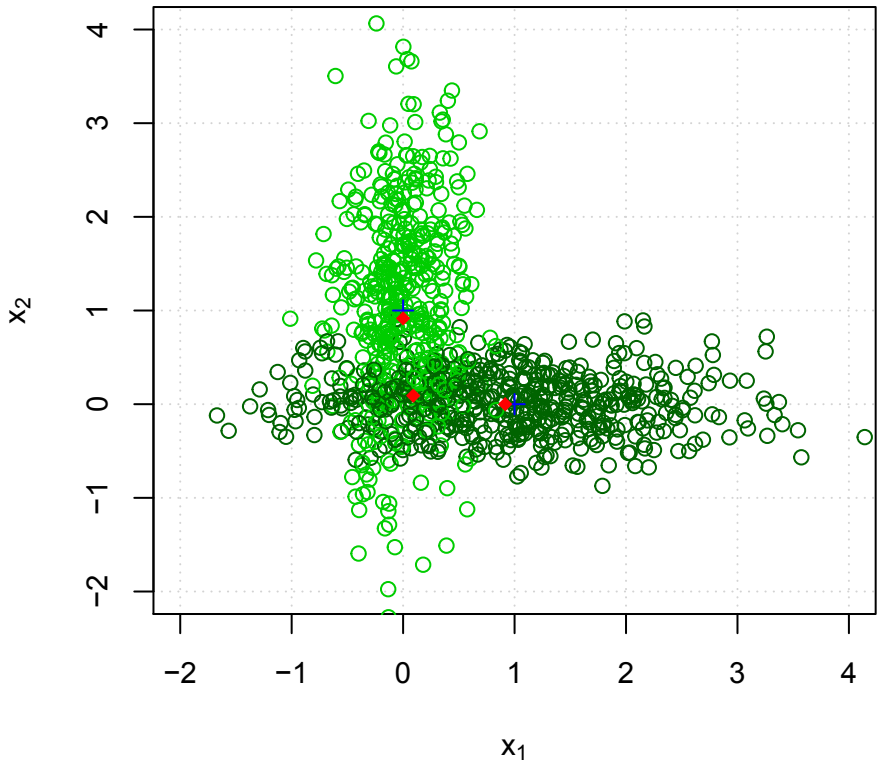

Figure 2.1.2.: Sample from a mixture of two bivariate Gaussians

mail bulletin board for statistics in Australia and New Zealand) with the title "*an interesting counter-intuitive fact*" referred to the fact that a Gaussian mixture can have more modes than components.

Example 2.1.2. Consider $X_1 \sim N\left(\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\0 & 0.1\end{smallmatrix}\right)\right)$ and $X_2 \sim N\left(\left(\begin{smallmatrix}0\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}0.1 & 0\\0 & 1\end{smallmatrix}\right)\right)$, with $\alpha = \frac{1}{2}$. There are two modes close to the original means at $(1, 0)$ and $(0, 1)$ but there is also a third mode near the origin. This situation is illustrated in Figure 2.1.2, with the two means marked in blue and the three modes marked in red.

Special attention can be paid to assumptions on the variances. A mixture is said to be *homoscedastic* if all the variances in the components are equal: $\Sigma_i = \Sigma$ for $i = 1, \ldots, k$. On the other hand, a mixture is said to be *isotropic* if $\Sigma_i = \sigma_i I$ for $i = 1, \ldots, k$, so that covariances are scalar matrices and the densities have a 'spherical' shape. Note that, up to coordinate change, homoscedastic mixtures are (homoscedastic) isotropic. Carreira and Williams conjectured in [CPW03b]

that if one is restricted to homoscedastic Gaussian mixtures, then the maximum number of modes is actually k, and verified this numerically for many examples in a brute force search. Denoting by $h(n,k)$ this maximum number, they asserted that $h(n,k) = k$ for any $n, k \geq 1$.

Remark 2.1.3. It holds that $h(n,k) \leq m(n,k)$ for all n, k, and $h(n,k)$ is also the maximum number of possible modes of a Gaussian mixture with all unit covariances (by the note above and the fact that the number of modes remains invariant under affine transformations)

However, later J.J. Duistermaat emailed the authors of [CPW03a] with a counterexample in $n = 2$ with $k = 3$ isotropic components, each on the vertex of an equilateral triangle. This configuration gives 4 modes for a small window of parameters, disproving the conjecture.

Example 2.1.4. Consider an isotropic mixture with components

$$X_1 \sim N((1,0), \sigma I_2) \quad X_2 \sim N\left((-\frac{1}{2}, \frac{\sqrt{3}}{2}), \sigma I_2\right) \quad X_3 \sim N\left((-\frac{1}{2}, -\frac{\sqrt{3}}{2}), \sigma I_2\right)$$

with $\sigma = 0.72$ and $\alpha = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. There are three modes close to the original means and there is also a fourth mode at the origin. This situation is illustrated in Figure 2.1.3, with the three means marked in blue and the four modes marked in red.

In terms of contribution to the study of the topography of Gaussian mixture densities, S. Ray and B. Lindsay initiate a systematic study and ask interesting questions in [RL05]. They consider the *ridgeline function* $x^* : \Delta_k \to \mathbb{R}^n$ given by

$$x^*(\alpha) = [\alpha_1\Sigma_1^{-1} + \alpha_2\Sigma_2^{-1} + \ldots + \alpha_k\Sigma_k^{-1}]^{-1}[\alpha_1\Sigma_1^{-1}\mu_1 + \alpha_2\Sigma_2^{-1}\mu_2 + \ldots + \alpha_k\Sigma_k^{-1}\mu_k]$$

and obtaining as its image the $(k-1)$-dimensional *ridgeline manifold* $\mathcal{M} = \mathrm{Im}(x^*)$ that contains all critical points of f_X. This fact is useful, for example, in the case of homoscedastic mixtures, whose critical points (and in particular all modes) lie in the convex hull of the component means (a result that appeared first in [CPW03b]). In the conclusion of [RL05], the following line appears: "*one might ask if there exists an upper bound for the number of modes, one that can be described as a function of k and n*". Assuming this bound is finite, we answer this question in the affirmative in Section 2.4 below.

2.2. Examples and Conjecture

The appearance of a possible extra mode in dimension $n = 2$ when having $k = 2$ components carries over to higher dimensions. In [RR12], it was proven that

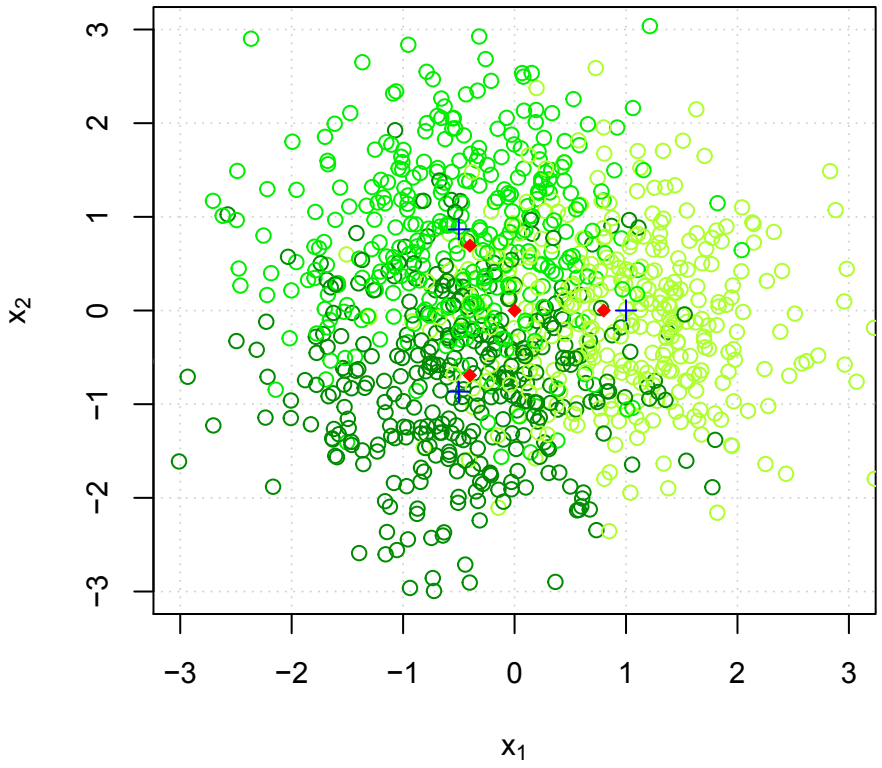

Figure 2.1.3.: Duistermaat's counterexample: 4 modes for mixture of 3 bivariate Gaussians

$m(n,2) = n+1$. That is, one can get as many as $n+1$ modes from just a two component Gaussian mixture in dimension n. Looking for further progress, Ray proposed the maximum number of modes problem for the 2011 AIM Workshop on Singular Learning Theory, organized by Steele, Sturmfels and Watanabe [Ste11]. The problem was discussed, and it led to the following conjecture:

Conjecture 2.2.1. *(Sturmfels, AIM 2011) For all $n, k \geq 1$,*

$$m(n,k) = \binom{n+k-1}{n}. \tag{2.2.1}$$

This conjecture matches correctly all the known values for $m(n,k)$ that we have presented so far. In the next section we will show that for $n = 2$ there exist Gaussian mixtures that achieve as many as $\binom{k+1}{2}$ modes, showing that (2.2.1) is a

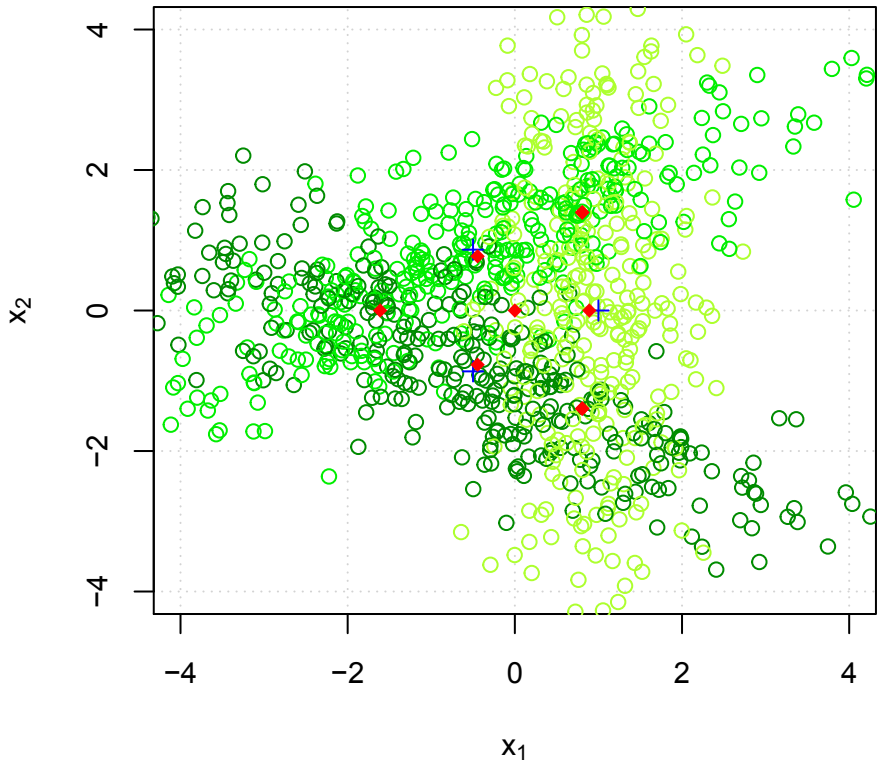

Figure 2.2.1.: Possible 7 modes for mixture of 3 bivariate Gaussians?

lower bound on $m(2, k)$.

One could ask if there exists a counterexample to Conjecture 2.2.1 similar to Duistermaat's for an isotropic mixture. Specifically, let $n = 2$ and $k = 3$, where we locate the $\binom{2+3-1}{2} = 6$ modes coming from the deformation of 3 lines arranged in a equilateral triangle and means as the middle points on the 3 sides. Each of the other 3 means lie near the corresponding triangle vertices. Could an extra mode be formed at the origin for some values of the parameters? This would give a total of 7 modes (see Figure 2.2.1). Note that by rotational symmetry, the origin is a critical point. If the Gaussians are very concentrated on the lines (see Figure 2.2.2), then everything points at the origin being a local minimum, but if they diffuse enough, then it becomes a mode (the problem is that we lose the other modes at the vertices). We argue that an intermediate scenario of a total of 7 modes is actually impossible. Indeed, consider any height of the equilateral

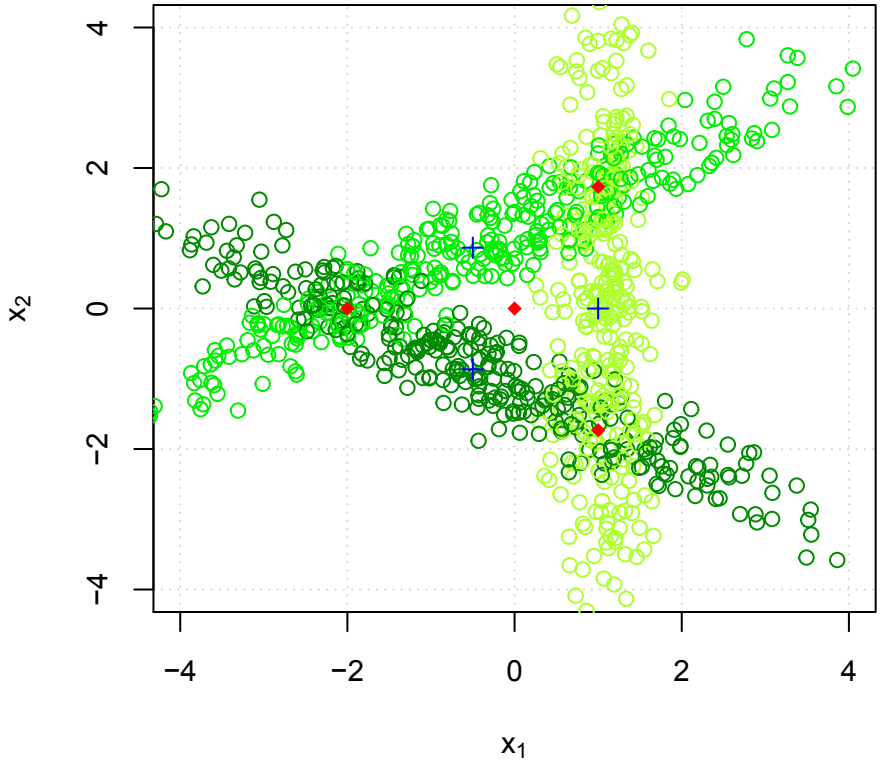

Figure 2.2.2.: 4 modes for mixture of 3 bivariate Gaussians

triangle. Again by symmetry, the corresponding modes near the vertex and middle point of the triangle lie on this height. Restricting the Gaussian mixture density to that line, the components corresponding to the opposite sides project to the same kernel; thus obtaining a combination of two Gaussian kernels. Since we know that the number of modes is at most two in this case, not all three of the critical points lying on the line can be modes.

In [EFR13], a construction of a finite configuration of isotropic Gaussians with many modes is presented. One considers products of triangles, using the counterexample with 4 modes of Section 2.1 as the basic building block to obtain 4^n modes in dimension $2n$ with $k = 3^n$ components. This gives an example where the number of modes is superlinear $k^{1.261}$ in the number of components (however, note that the dimension $2n = 2\log_3 k$ also grows with k).

In the following section, we provide configurations for any choice of n and k

having $\binom{k}{n} + k$ modes. If we let n grow logarithmically with k as in [EFR13], we obtain superpolynomially (but subexponentially) many modes:

$$\binom{k}{\log k} = k^{\log k + O(\log\log k)}. \tag{2.2.2}$$

2.3. Many Modes

In this section, we prove that Gaussian mixtures can have many modes. For $n = 2$, this agrees with (2.2.1).

Theorem 2.3.1. *Given $k, n \in \mathbb{N}$, there is a mixture of k Gaussians in $\mathbb{R}^n$ with at least $\binom{k}{n} + k$ modes.*

Proof. Starting from a generic arrangement of k hyperplanes in $\mathbb{R}^n$, we are going to define a family $\Phi = \Phi^\delta$ of Gaussian mixtures depending on a parameter $\delta > 0$. Around each of the $\binom{k}{n}$ intersection vertices p of the arrangement, we construct neighborhoods $Q = Q(p)$, also depending on δ, so that for δ small enough, we have $\Phi^\delta|_{\partial Q} < \Phi^\delta(p)$. This certifies the existence of a mode in Q. In addition, there will be a mode near each of the k means. Consider a generic arrangement $H_1, \dots, H_k$ of k hyperplanes in $\mathbb{R}^n$. For each $i = 1, \dots, k$, denote by $\pi_i \colon \mathbb{R}^n \to H_i$ the orthogonal projection, and pick an affine map $\eta_i \colon \mathbb{R}^n \to \mathbb{R}$ such that $|\eta_i(x)|$ is the distance from x to H_i. Further, choose means $\mu_i \in H_i$ outside the other H_j. Then, our ith component will be a standard Gaussian with mean μ_i along H_i with variance δ^3 in the direction normal to H_i.

$$\Phi_i(x) := \frac{1}{\sqrt{(2\pi)^d\ \delta^3}}\ \exp\left(-\frac{1}{2\,\delta^3}\,|\eta_i(x)|^2\ -\frac{1}{2}\,\|\pi_i(x) - \mu_i\|^2\right) \tag{2.3.1}$$

For the mixture, we take all coefficients to be equal: $\Phi = \frac{1}{k}\sum_i \Phi_i$. Let p be one of the intersection vertices; without loss of generality, $\{p\} = H_1 \cap \cdots \cap H_n$. For $\delta > 0$, we define the neighborhood Q of p to be

$$Q(p) = \{x \in \mathbb{R}^n \,|\, |\eta_i(x)| \le \delta,\ \forall 1 \le i \le n\}$$

(note that $\eta_i(p) = 0$). This is an affine cube with center p. Now we consider each of its $2d$ facets $F_i^\pm$, $1 \le i \le d$, where

$$F_i^\pm = \{x \in \mathbb{R}^n \,|\, \eta_i(x) = \pm\delta,\ |\eta_j(x)| \le \delta\ \ \forall j \ne i\}\,.$$

Along $F_i^\pm$, we have that

$$\Phi_i(x) \le \frac{1}{\sqrt{(2\pi)^d\ \delta^3}}\ \exp\left(-\frac{1}{2\,\delta^3}\,\delta^2\ \right) \quad \xrightarrow{\delta\to 0} \quad 0.$$

At p, we have $\eta_i(p) = 0$. We have seen that, around the point p, as $\delta \to 0$,

$$\sqrt{\delta^3} \max_{x \in F_j^{\pm}} \Phi_i(x) \quad \to \quad \phi_i \quad \text{for } i, j = 1, \ldots, d,\ j \neq i \tag{2.3.2}$$

$$\sqrt{\delta^3} \max_{x \in F_i^{\pm}} \Phi_i(x) \quad \to \quad 0 \quad \text{for } i = 1, \ldots, d \tag{2.3.3}$$

$$\sqrt{\delta^3} \max_{x \in Q} \Phi_i(x) \quad \to \quad 0 \quad \text{for } i = d+1, \ldots, k \tag{2.3.4}$$

where ϕ_i is the positive number

$$\phi_i := \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\left\|p - \mu_i\right\|^2\right).$$

Adding up we get for $\Phi = \frac{1}{k}\sum_{i=1}^{k} \Phi_i$ that

$$\sqrt{\delta^3} \max_{x \in \partial Q} \Phi(x) \quad \to \quad \frac{1}{k} \sum_{\substack{i=1 \\ i \neq j}}^{k} \phi_i \tag{2.3.5}$$

$$\sqrt{\delta^3}\ \Phi(p) \quad \to \quad \frac{1}{k} \sum_{i=1}^{k} \phi_i \tag{2.3.6}$$

where $j = \text{argmin}\phi_i$. As the limit (2.3.5) is smaller (by $\phi_j > 0$) than the limit (2.3.6), there must be some $\delta^\star(p) > 0$ so that for $0 < \delta < \delta^\star(p)$ we have $\max_{x \in \partial Q} \Phi(x) < \Phi(p)$. Then the point p' where the continuous function Φ takes it maximum over the compact set Q will be in the interior of Q, and hence is a local maximum. Choosing $\delta^\star$ to be the minimum over the $\delta^\star(p)$ over all intersection vertices p, we obtain a mixture with at least $\binom{k}{n}$ modes.

The argument for the existence of a mode near μ_i is similar, but much simpler. Fix a compact neighborhood Q of μ_i which avoids the hyperplanes H_j for $j \neq i$. For small δ, the Φ_j for $j \neq i$ are negligible along Q. So the value of the mixture Φ at μ_i will be bigger than the values along ∂Q, because this is true for Φ_i. □

2.4. Not Too Many Modes

The main result of this section is to present an upper bound on the number of modes of a Gaussian mixture: Theorem 2.4.4. We start by looking at the set of critical points and we will use Khovanskii's theory on Fewnomials, see [Kho91].

Theorem 2.4.1. *For all $n, k \geq 1$, the number of non-degenerate critical points*

for the density of a mixture of k Gaussians in $\mathbb{R}^n$ is bounded by

$$2^{n+\binom{k}{2}}(5+3n)^k. \tag{2.4.1}$$

This will follow from a Khovanskii-type theorem that bounds the number of nondegenerate solutions to a system of polynomial equations that includes transcendental functions. Such a version where the transcendental functions are exponentials of linear forms was first presented by Khovanskii to illustrate his theory of fewnomials [Kho91, p. 12]. In our case, however, we will be interested in exponentials of quadratic forms.

Theorem 2.4.2. *For $1 \leq i \leq n$, let $F_i \in \mathbb{R}[x_1, \ldots, x_n, y_1, \ldots, y_k]$ be polynomials of degree d_i and for $1 \leq j \leq k$ consider the exponential quadratic forms $y_j(x) = \mathrm{e}^{(x-\mu_j)^T Q_j (x-\mu_j)}$, with $\mu_j \in \mathbb{R}^n$ and $Q_j \in \mathbb{R}^{n\times n}$. If $g_i : \mathbb{R}^n \to \mathbb{R}$ are given by $g_i(x) = F_i(x_1, \ldots, x_n, y_1(x), \ldots, y_k(x))$ then the number of non-degenerate solutions to the system $g_1 = g_2 = \cdots = g_n = 0$ is finite and bounded by*

$$d_1 \cdots d_n (5 + n + d_1 + \ldots + d_n)^k \cdot 2^{\frac{k(k-1)}{2}}. \tag{2.4.2}$$

In order to prove the theorem, Khovanskii gives first a sketch making simplifying assumptions (and skipping technical details) and fills the theory in his next two chapters. This sketch is also presented in [Sot11] and [BCS13], and we will present the proof of our theorem in the same way. We will need the following lemma (for a proof see e.g. Theorem 4.3 in [Sot11]).

Lemma 2.4.3 (Khovanskii-Rolle)**.** *Let $C \subset \mathbb{R}^{n+1}$ be a smooth curve that intersects $H := x_{n+1} = 0$ transversally, and $v = (v_1, \ldots, v_{n+1}) : C \to \mathbb{R}^{n+1}$ a smooth nonvanishing tangential vector field to C. Then $|(C \cap H)| < N + q$, where N is the number of points of C where $v_{n+1} = 0$ and q is the number of unbounded components of C.*

Proof. (of Theorem 2.4.2) By Induction on k. If $k = 0$, there are no exponentials and the bound (2.4.2) reduces to the product of the degrees $d_1 \cdots d_n$. This is the well known Bézout bound for a multivariate system of polynomial equations.
Now we will give the sketch of the proof and mention how our estimates change if the assumptions in the induction step do not hold. In any case, the final inequalities needed to prove the bound (2.4.2) will hold.
For $k \geq 1$, to reduce the number of exponentials to $k-1$, we introduce a new variable t such that the system with equations

$$\widehat{g}_i(x,t) = F_i(x_1, \ldots, x_n, y_1(x), \ldots, y_{k-1}(x), t y_k(x)) \tag{2.4.3}$$

will have as solutions the ones from the original system when intersecting with the hyperplane $t = 1$. We assume the system (2.4.3) defines a smooth curve C in $\mathbb{R}^{n+1}$

(this is the critical step that needs to be modified later), so that we can apply the Khovanskii-Rolle Lemma. Indeed, a tangential vector field to C is

$$v_r = (-1)^{n+1-r} \det \frac{\partial(\widehat{g_1}, \dots \widehat{g_n})}{\partial(x_1, \dots, \hat{x_i}, \dots, x_n, t)}. \tag{2.4.4}$$

Thus, the bound for N is the number of solutions to the system in the $n+1$ variables $x_1, x_2, \dots, x_n, u$ with $u = ty_k(x)$ and $k-1$ exponentials

$$\widehat{g_1} = 0, \dots, \widehat{g_n} = 0, v_{n+1} = 0. \tag{2.4.5}$$

Now, each $\frac{\partial y_j}{\partial x_i}(x) = y_j \cdot l_{ij}(x)$ where l_{ij} is a linear function. Hence, $\frac{\partial \widehat{g_i}}{\partial x_j} = h_{ij}(x, y_1(x), \dots, y_{k-1}(x), u)$ where h_{ij} is a polynomial of degree at most $d_i + 1$. Thus $v_{n+1}(x, u)$ is a polynomial of degree at most $(d_1+1)+\ldots+(d_n+1) = n+D$, where $D = d_1 + \ldots + d_n$. By induction hypothesis,

$$N \leq d_1 \cdots d_n (n+D)(5 + (n+1) + (D+n+D))^{k-1} \cdot 2^{\frac{(k-1)(k-2)}{2}} \tag{2.4.6}$$

In order to bound q, the number of unbounded components of C, one observes that, since each component has two infinite branches, there are (counted with multiplicity) $2q$ accumulation points on the sphere S^n, so a hyperplane sufficiently far from the origin will meet C in at least q points. In other words, a hyperplane can be chosen so that q can be bounded by the number of solutions of a system

$$\widehat{g_1} = 0, \dots, \widehat{g_n} = 0, \lambda_1 x_1 + \ldots + \lambda_n x_n + \lambda_{n+1} u + \mu = 0. \tag{2.4.7}$$

for some $\lambda_i, \mu \in \mathbb{R}$. Under non-degeneracy of the solutions, we get by induction hypothesis,

$$q \leq d_1 \cdots d_n \cdot 1 \cdot (5 + (n+1) + (D+1))^{k-1} \cdot 2^{\frac{(k-1)(k-2)}{2}} \tag{2.4.8}$$

So, in total,

$$\begin{aligned} N + q &\leq d_1 \cdots d_n \left[(n+D)(6+2n+2D)^{k-1} + (7+n+D)^{k-1}\right] 2^{\frac{(k-1)(k-2)}{2}} \\ &< d_1 \cdots d_n \left[(n+D)(5+n+D)^{k-1} 2^{k-1} + 5(5+n+D)^{k-1} 2^{k-1}\right] 2^{\frac{(k-1)(k-2)}{2}} \\ &= d_1 \cdots d_n (5+n+D)^k \cdot 2^{\frac{k(k-1)}{2}}. \end{aligned}$$

as we wanted. This is the end of the sketch.

If the smoothness assumptions for the system after introducing t are not satisfied, the argument is modified via the Morse-Sard Theorem. The details of such modifications can be found along [Kho91], although we find that for our theorem these are better summarized in [BCS13, p. 293-295]. Essentially, one slightly per-

turbs the system from $\hat{g}_i = 0$ to $\hat{g}_i = \epsilon_i$ (ϵ in a neighborhood of 0) to guarantee obtaining a smooth curve C. The asserted bound (2.4.2) remains unchanged, and the number of non-degenerate solutions of the perturbed system cannot be less than the number for the original system.

Another change is that since the polynomial system might not define a proper map, one adds an extra variable x_0 with an extra equation

$$g_0(x_0, x_1, \ldots, x_n) = x_0^2 + x_1^2 + \ldots + x_n^2 - R^2 = 0 \tag{2.4.9}$$

with $R > 0$ so that every preimage is now bounded. Morse-Sard now applies to conclude the set of regular values of $g = (g_0, \ldots, g_n) : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is open and dense. The number of non-degenerate solutions of the new system has twice the number of non-degenerate solutions of the original system that lie in the open ball of radius R centered at the origin . In terms of the bounding of the corresponding N', q', we now have

$$N' \leq 2 \cdot d_1 \cdots d_n \cdot (n+1+D)(5+(n+2)+(2+D+n+1+D))^{k-1} \cdot 2^{\frac{(k-1)(k-2)}{2}} \tag{2.4.10}$$

(the extra 2 comes from the degree of 2.4.9, and we now have $n+2$ variables). For q', the bound becomes

$$q' \leq 2 \cdot d_1 \cdots d_n \cdot 1 \cdot (5 + (n+2) + 2 + D + 1)^{k-1} \cdot 2^{\frac{(k-1)(k-2)}{2}} \tag{2.4.11}$$

(the extra 1 from the hyperplane equation). Since the $R > 0$ does not affect the bound computation, it can be taken large enough to include all the solutions to the original system. Thus $N' + q'$ is a bound for twice as many non-degenerate solutions of said original system. Finally, this way the induction step inequality can again be completed

$$\begin{aligned} \frac{N' + q'}{2} &\leq d_1 \cdots d_n \left[(1+n+D)(10+2n+2D)^{k-1} + (10+n+D)^{k-1}\right] 2^{\binom{k-1}{2}} \\ &< d_1 \cdots d_n \left[(1+n+D)(5+n+D)^{k-1} + 4(5+n+D)^{k-1}\right] 2^{k-1} 2^{\binom{k-1}{2}} \\ &= d_1 \cdots d_n (5+n+D)^k \cdot 2^{\frac{k(k-1)}{2}}, \end{aligned}$$

as needed. □

Now we can obtain Theorem 2.4.1 as a corollary of the above.

Proof. (of Theorem 2.4.1) Let $f_X(x) = \sum_{i=1}^k \alpha_i f_{X_i}(x)$ be a Gaussian mixture and consider the system $g_i(x_1, \ldots, x_n)$ given by the partial derivatives $g_i = \frac{\partial f_X}{\partial x_i}$. These can be interpreted as polynomials $F_i(x_1, \ldots, x_n, y_1, \ldots, y_k)$ by taking y_i as the exponential kernel of f_{X_i}. The system now has the form as in Theorem 2.4.2, and note that the degree of each F_i is 2. The number of non-degenerate critical

points of f_X is thus the number of non-degenerate solutions to the system of g_i, and according to (2.4.2), it is bounded by

$$2 \cdots 2 \, (5 + n + 2 + \ldots + 2)^k \cdot 2^{\frac{k(k-1)}{2}} = 2^n (5 + n + 2n)^k \, 2^{\binom{k}{2}} = 2^{n+\binom{k}{2}} (5 + 3n)^k.$$

□

Finally, the promised upper bound on the number of modes of a Gaussian mixture, provided it is finite.

Theorem 2.4.4. *Consider a Gaussian mixture with $n, k > 1$ that has finitely many modes. Then its number of modes is bounded by*

$$2^{n+\binom{k}{2}} (5 + 3n)^k. \tag{2.4.12}$$

Proof. Let $f(x) = \sum_{i=1}^{k} \alpha_i f_i(x)$ be a Gaussian mixture. If all of its modes are non-degenerate, then Theorem 2.4.2 applies and we're done. The difficulty stems from considering possible degenerate modes. Note that by the finiteness hypothesis, all of them are isolated so we may fix disjoint neighborhoods $Q_i \subset \mathbb{R}^n$ over which each mode is a global maximum.

For any linear function $\ell(x_1, \ldots, x_n) = c \cdot x$, the function $f + \ell$ has a gradient that differs by the constant vector c from ∇f. In particular, the system given by the partial derivatives of $f + \ell$ are still polynomials $F_i(x_1, \ldots, x_n, y_1, \ldots, y_k)$ of degree 2. By Theorem 2.4.1 we have the bound (2.4.1) on the non-degenerate modes of $f + \ell$. Since $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, one of Morse's Lemmas [MSS65, Lemma A, p.11] states that for almost all $c \in \mathbb{R}^n$ (all except for a set of measure zero), $f + \ell$ has only non-degenerate critical points.

Now, let $c \in \mathbb{R}^n$ in the complement of such measure zero set, with norm small enough so that $f + \ell$ still has modes inside each of the neighborhoods Q_i. Then, $f + \ell$ may have fewer critical points than f but we know $f + \ell$ has at least as many modes as f does. Since the modes of $f + \ell$ are bounded by (2.4.1), the result follows. □

Corollary 2.4.5. *If every mixture of k Gaussians in $\mathbb{R}^n$ has finitely many modes, then*

$$m(n, k) \leq 2^{n+\binom{k}{2}} (5 + 3n)^k. \tag{2.4.13}$$

2.5. Future Work

Are there finitely many critical points? This was the main goal sought in [Wal13] but could only be proven for $k = 2$. As our bound (2.4.2) only bounds the number of non-degenerate critical points of Gaussian mixtures, a final answer is still open.

However, since non-degenerate critical points are isolated, we see that the set of critical points of a Gaussian mixture is finite if and only if it consists of isolated points, since this set is closed and bounded (compare [Gha15]).

Quantitatively, we do not expect our upper bound to be tight. Rather, proving the lower bound $\binom{n+k-1}{n}$ for all n, k will be the main focus of a forthcoming paper, extending the technique used to prove Theorem 2.3.1 .

We observe that our lower bound can be extended to elliptical distributions, not only Gaussians. A study of the number of modes for mixtures of elliptical distributions, including t-distributions, is done in [AHR13]. This is done by extending the concept of the ridgeline manifold to their densities (upper bounds are 2 or 3 modes for a mixture of two t-distributions with equal dispersion matrices or the same degrees of freedom).

The main takeaway of this chapter is that Gaussian mixture densities can have quite complex behavior. We should not expect to easily get results about them, especially when we move to statistical aspects in the next chapters. However, it is precisely the versatility of the mixture density that makes it attractive for modeling. Therefore, we will make an effort and delve deeper into these models.

3. Maximum Likelihood Estimation and Transcendence

Suppose we have a random sample $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ that we know or believe was generated from a mixture of k Gaussians in $\mathbb{R}^n$. We would like to estimate all the parameters: the means $\mu_i \in \mathbb{R}^n$, the positive definite $n \times n$ symmetric matrices Σ_i and the weights α_i for $i = 1, \ldots, k$ that generated the observed data.

From the Introduction, we know one central approach is via maximum likelihood inference.

3.1. In Search of the ML Degree

Recall that we want to maximize the log-likelihood function

$$\ell(x_1, \ldots, x_n; \theta) = \log f(x_1; \theta) + \log f(x_2; \theta) + \cdots + \log f(x_N; \theta)$$

where for a Gaussian mixture, θ is the parameter vector consisting of the means μ_i, the variances Σ_i and the weights α_i, $i = 1, \ldots, k$.

To illustrate the method, let's compute the ML estimate for a single Gaussian (that is, $k = 1$).

Example 3.1.1. Sample $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ from $X \sim N(\mu, \Sigma)$. The likelihood is

$$l(x_1, \ldots, x_N; \mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{\det(2\pi\Sigma)}} \mathrm{e}^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$

and the log-likelihood is thus

$$\ell(x_1, \ldots, x_N; \mu, \Sigma) = -\frac{nN}{2}\log 2\pi - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu).$$

To find the maximizer $(\hat{\mu}, \hat{\Sigma})$ of ℓ we solve the critical equations $\dfrac{\partial\ell}{\partial\mu} = 0$ and $\dfrac{\partial\ell}{\partial\Sigma} = O$ (that is, $\dfrac{\partial\ell}{\partial\sigma_{ij}} = 0$ for all $1 \le i, j \le n$).

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x}$$

so the ML estimator for the mean is the sample mean. For the second equation, we use the 'trace trick' to rewrite the log-likelihood:

$$\begin{aligned} \ell(x_1, \ldots, x_N; \mu, \Sigma) &= -\frac{nN}{2} \log 2\pi - \frac{N}{2} \log \det \Sigma - \frac{1}{2} \mathrm{tr} \left(\sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right) \\ &= -\frac{nN}{2} \log 2\pi - \frac{N}{2} \log \det \Sigma - \frac{1}{2} \sum_{i=1}^{N} \mathrm{tr} \left((x_i - \mu)(x_i - \mu)^T \Sigma^{-1} \right) \end{aligned}$$

Using the matrix differentiation identities (see e.g. [PP+08]):

$$\frac{\partial}{\partial A} \log \det A = A^{-T} \qquad \frac{\partial}{\partial A} \mathrm{tr}(BA) = B^T$$

we get that

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} \left(\sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \right) \Sigma^{-1} = O$$

So the ML estimator for the covariance matrix is

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T,$$

assuming it is full rank (for instance, if $N > n$ then it is positive definite with probability 1). To argue that this unique interior critical point is indeed the global maximum, one can check that the log-likelihood approaches $-\infty$ as Σ approaches the boundary of the cone of positive definite matrices. □

We see then that maximum likelihood estimates for $k = 1$:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x} \qquad \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

are rational functions of the data $x_1, \ldots, x_n \in \mathbb{R}^n$, even when the density function for the Gaussian involves an exponential. Since there is a unique critical point, this means the ML degree (cf. section 1.2) for the Gaussian model is *1*.

What would be the ML degree, that is, the general number of solutions for the critical equations, for $k > 1$? We asked this in Problem 4 and now we give an

answer that will be the main result of this chapter.

Theorem 3.1.2. *The maximum likelihood estimators of Gaussian mixture models are transcendental functions. More precisely, there exist $x_1, x_2, \ldots, x_N$ rational samples in $\mathbb{Q}^n$ whose maximum likelihood parameters for the mixture of two n-dimensional Gaussians are not algebraic numbers over $\mathbb{Q}$.*

Theorem 3.1.2 means that there is *no* ML degree for Gaussian mixtures.

3.2. Reaching Transcendence

Transcendental number theory [Bak90, Gel15] is a field that furnishes tools for deciding whether a given real number τ is a root of a nonzero polynomial in $\mathbb{Q}[t]$. If this holds then τ is algebraic; otherwise τ is transcendental. For instance, $\sqrt{2}+\sqrt{7} = 4.059964873...$ is algebraic, and so are the parameter estimates computed by Pearson in his 1894 study of crab data [Pea94]. By contrast, the famous constants $\pi = 3.141592653...$ and $e = 2.718281828...$ are transcendental. Our proof will be based on the following classical result. A textbook reference is [Bak90, Theorem 1.4]:

Theorem 3.2.1 (Lindemann-Weierstrass)**.** *If $u_1, \ldots, u_r$ are distinct algebraic numbers then $e^{u_1}, \ldots, e^{u_r}$ are linearly independent over the algebraic numbers.*

For now, consider the case of $n = 1$, that is, mixtures of two univariate Gaussians. We allow mixtures with arbitrary means and variances. Our model then consists of all probability distributions on the real line $\mathbb{R}$ with density

$$f_{\alpha,\mu,\sigma}(x) \;=\; \frac{1}{\sqrt{2\pi}} \cdot \left[\frac{\alpha}{\sigma_1} \exp\Big(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\Big) \;+\; \frac{1-\alpha}{\sigma_2} \exp\Big(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\Big)\right]. \quad (3.2.1)$$

It has five unknown parameters, namely, the means $\mu_1, \mu_2 \in \mathbb{R}$, the standard deviations $\sigma_1, \sigma_2 > 0$, and the mixture weight $\alpha \in [0, 1]$. The aim is to estimate the five model parameters from a collection of data points $x_1, x_2, \ldots, x_N \in \mathbb{R}$.

The *log-likelihood function* of the model (3.2.1) is

$$\ell(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) \;\;=\;\; \sum_{i=1}^{N} \log f_{\alpha,\mu,\sigma}(x_i). \quad (3.2.2)$$

This is a function of the five parameters, while $x_1, \ldots, x_N$ are fixed constants.

The principle of maximum likelihood suggests to find estimates by maximizing the function ℓ over the five-dimensional parameter space $\Theta = [0, 1] \times \mathbb{R}^2 \times \mathbb{R}^2_{>0}$.

Remark 3.2.2. The log-likelihood function ℓ in (3.2.2) is never bounded above. To see this, we argue as in [Bis06, Section 9.2.1]. Set $N = 2$, fix arbitrary values $\alpha_0 \in [0,1]$, $\mu_{20} \in \mathbb{R}$ and $\sigma_{20} > 0$, and match the first mean to the first data point $\mu_1 = x_1$. The remaining function of one unknown σ_1 equals

$$\ell(\alpha_0, x_1, \mu_{20}, \sigma_1, \sigma_{20}) \quad \geq \quad \log\left[\frac{\alpha_0}{\sigma_1} + \frac{1-\alpha_0}{\sigma_{20}} \exp\Big(-\frac{(x_1-\mu_{20})^2}{2\sigma_{20}^2}\Big)\right] + \text{const.}$$

The lower bound tends to ∞ as $\sigma_1 \to 0$.

Remark 3.2.2 means that there is no global solution to the MLE problem. This is usually remedied by restricting to a subset of the parameter space Θ. In practice, maximum likelihood for Gaussian mixtures means computing local maxima of the function ℓ. These are found numerically by a hill climbing method, such as the EM algorithm, with particular choices of starting values. See Section 3.3. This method is implemented, for instance, in the `R` packages `mclust` [FR03], `mixtools` [BCHY09] and `EMCluster` [CMM12]. In order for Theorem 3.1.2 to cover such local maxima, we prove the following statement:

> *There exist samples $x_1, \ldots, x_N \in \mathbb{Q}$ such that every non-trivial critical point $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ of the log-likelihood function ℓ in the domain Θ has at least one transcendental coordinate.*

Here, a critical point is *non-trivial* if it yields an honest mixture, i.e. a distribution that is not Gaussian. By the identifiability results of [Tei63], this happens if and only if the estimate $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ satisfies $0 < \hat{\alpha} < 1$ and $(\hat{\mu}_1, \hat{\sigma}_1) \neq (\hat{\mu}_2, \hat{\sigma}_2)$.

Remark 3.2.3. The log-likelihood function always has some algebraic critical points, for any $x_1, \ldots, x_N \in \mathbb{Q}$. Indeed, if we define the empirical mean and variance as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad\qquad s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

then any point $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ with $\hat{\mu}_1 = \hat{\mu}_2 = \bar{x}$ and $\hat{\sigma}_1 = \hat{\sigma}_2 = s$ is critical. This is the case of a single Gaussian distribution treated in Example 3.1.1, with mean $\bar{x}$ and variance s^2, so it is trivial.

Proof. (of Theorem 3.1.2) First, we treat the univariate case $n = 1$. Consider the partial derivative of (3.2.2) with respect to the mixture weight α:

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{N} \frac{1}{f_{\alpha,\mu,\sigma}(x_i)} \cdot \frac{1}{\sqrt{2\pi}} \left[\frac{1}{\sigma_1}\exp\Big(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\Big) - \frac{1}{\sigma_2}\exp\Big(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\Big)\right]. \quad (3.2.3)$$

Clearing the common denominator

$$\sqrt{2\pi} \cdot \prod_{i=1}^{N} f_{\alpha,\mu,\sigma}(x_i),$$

we see that $\partial \ell / \partial \alpha = 0$ if and only if

$$\sum_{i=1}^{N} \left[\frac{1}{\sigma_1} \exp\Big(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\Big) \; - \; \frac{1}{\sigma_2} \exp\Big(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\Big) \right] \times \prod_{j \neq i} \left[\frac{\alpha}{\sigma_1} \exp\Big(-\frac{(x_j-\mu_1)^2}{2\sigma_1^2}\Big) + \frac{1-\alpha}{\sigma_2} \exp\Big(-\frac{(x_j-\mu_2)^2}{2\sigma_2^2}\Big) \right] = 0. \quad (3.2.4)$$

Letting $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$, we may rewrite the left-hand side of (3.2.4) as

$$\sum_{i=1}^{N} \left[\sum_{k_i=1}^{2} \frac{(-1)^{k_i - 1}}{\sigma_{k_i}} \exp\Big(-\frac{(x_i-\mu_{k_i})^2}{2\sigma_{k_i}^2}\Big) \right] \prod_{j \neq i} \left[\sum_{k_j=1}^{2} \frac{\alpha_{k_j}}{\sigma_{k_j}} \exp\Big(-\frac{(x_j-\mu_{k_j})^2}{2\sigma_{k_j}^2}\Big) \right]. \quad (3.2.5)$$

We expand the products, collect terms, and set $N_i(k) = |\{j : k_j = i\}|$. With this, the partial derivative $\partial \ell / \partial \alpha$ is zero if and only if the following vanishes:

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{k \in \{1,2\}^N} \exp\left(-\sum_{j=1}^{N} \frac{(x_j - \mu_{k_j})^2}{2\sigma_{k_j}^2} \right) (-1)^{k_i - 1} \alpha^{|\{j \neq i : k_j = 1\}|} (1-\alpha)^{|\{j \neq i : k_j = 2\}|} \left(\prod_{j=1}^{N} \frac{1}{\sigma_{k_j}} \right) \\
&= \sum_{k \in \{1,2\}^N} \exp\left(-\sum_{j=1}^{N} \frac{(x_j - \mu_{k_j})^2}{2\sigma_{k_j}^2} \right) \left(\prod_{j=1}^{N} \frac{1}{\sigma_{k_j}} \right) \sum_{i=1}^{N} (-1)^{k_i - 1} \alpha^{|\{j \neq i : k_j = 1\}|} (1-\alpha)^{|\{j \neq i : k_j = 2\}|} \\
&= \sum_{k \in \{1,2\}^N} \exp\left(-\sum_{j=1}^{N} \frac{(x_j - \mu_{k_j})^2}{2\sigma_{k_j}^2} \right) \left(\prod_{j=1}^{N} \frac{1}{\sigma_{k_j}} \right) \alpha^{N_1(k)-1} (1-\alpha)^{N_2(k)-1} \begin{bmatrix} N_1(k)(1-\alpha) \\ +N_2(k)(-\alpha) \end{bmatrix} \\
&= \sum_{k \in \{1,2\}^N} \exp\left(-\sum_{j=1}^{N} \frac{(x_j - \mu_{k_j})^2}{2\sigma_{k_j}^2} \right) \left(\prod_{j=1}^{N} \frac{1}{\sigma_{k_j}} \right) \alpha^{N_1(k)-1} (1-\alpha)^{N - N_1(k) - 1} (N_1(k) - N\alpha).
\end{aligned}$$

Let $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ be a non-trivial isolated critical point of the likelihood function. This means that $0 < \hat{\alpha} < 1$ and $(\hat{\mu}_1, \hat{\sigma}_1) \neq (\hat{\mu}_2, \hat{\sigma}_2)$. This point depends continuously on the choice of the data $x_1, x_2, \ldots, x_N$. By moving the vector with these coordinates along a general line in $\mathbb{R}^N$, the mixture parameter $\hat{\alpha}$ moves continuously in the critical equation $\partial \ell / \partial \alpha = 0$ above. By the Implicit Function Theorem, it takes on all values in some open interval of $\mathbb{R}$, and we can thus choose our data points x_i general enough so that $\hat{\alpha}$ is not an integer multiple of $1/N$.

We can further ensure that the last sum above is a $\mathbb{Q}(\alpha)$-linear combination of exponentials with nonzero coefficients.

Suppose that $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ is algebraic. The Lindemann-Weierstrass Theorem implies that the arguments of exp are all the same. Then the 2^N numbers

$$\sum_{j=1}^{N} \frac{(x_j - \hat{\mu}_{k_j})^2}{2\hat{\sigma}_{k_j}^2}, \qquad k \in \{1,2\}^N,$$

are all identical. However, for $N \geq 3$, and for general choice of data $x_1, \ldots, x_N$ as above, this can only happen if $(\hat{\mu}_1, \hat{\sigma}_1) = (\hat{\mu}_2, \hat{\sigma}_2)$. This contradicts our hypothesis that the critical point is non-trivial. We conclude that all non-trivial critical points of the log-likelihood function (3.2.2) are transcendental.

In the multivariate case, the model parameters comprise the mixture weight $\alpha \in [0,1]$, mean vectors $\mu_1, \mu_2 \in \mathbb{R}^n$ and positive definite covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^n$. Arguing as above, if a non-trivial critical $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2)$ is algebraic, then the Lindemann-Weierstrass Theorem implies that the numbers

$$\sum_{j=1}^{N} (x_j - \hat{\mu}_{k_j})^T \hat{\Sigma}_{k_j}^{-1} (x_j - \hat{\mu}_{k_j}), \qquad k \in \{1,2\}^N,$$

are all identical. For N sufficiently large and a general choice of $x_1, \ldots, x_N$ in $\mathbb{R}^n$, the 2^N numbers are identical only if $(\hat{\mu}_1, \hat{\Sigma}_1) = (\hat{\mu}_2, \hat{\Sigma}_2)$. Again, this constitutes a contradiction to the hypothesis that $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2)$ is non-trivial. □

Many variations and specializations of the Gaussian mixture model are used in applications. In the case $n = 1$, the variances are sometimes assumed equal, so $\sigma_1 = \sigma_2$ for the above two-mixture. This avoids the issue of an unbounded likelihood function (as long as $N \geq 3$). Our proof of Theorem 3.1.2 applies to this setting. In higher dimensions ($n \geq 2$), the covariance matrices are sometimes assumed arbitrary and distinct, sometimes arbitrary and equal, but often also have special structure such as being diagonal. Various default choices are discussed in the paper [FR03] that introduces the `R` package `mclust`. See also Chapter 6. Our results imply that maximum likelihood estimation is transcendental also for all of these models.

Example 3.2.4. We illustrate Theorem 3.1.2 for a specialization of (3.2.1) obtained by fixing three parameters: $\mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1/\sqrt{2}$. The remaining two free parameters are α and $\mu = \mu_1$. We take only $N = 2$ data points, namely $x_1 = 0$ and $x_2 = x > 0$. Omitting an additive constant, the log-likelihood equals

$$\ell(\alpha, \mu) \;=\; \log\bigl(\alpha \cdot e^{-\mu^2} + (1-\alpha)\bigr) \;+\; \log\bigl(\alpha \cdot e^{-(\mu - x)^2} + (1-\alpha) \cdot e^{-x^2}\bigr). \quad (3.2.6)$$

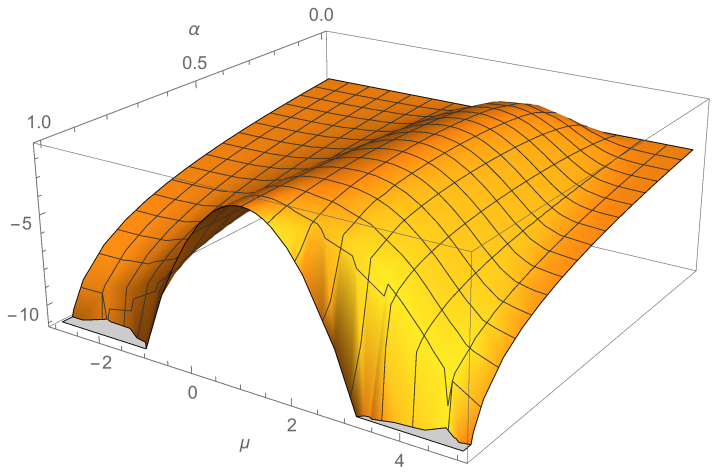

Figure 3.2.1.: Graph of the log-likelihood function for two data points $x_1 = 0$ and $x_2 = 2$.

For a concrete example take $x = 2$. The graph of (3.2.6) for this choice is shown in Figure 3.2.1. By maximizing $\ell(\alpha, \mu)$ numerically, we find the parameter estimates

$$\hat{\alpha} = 0.50173262959803874... \quad \text{and} \quad \hat{\mu} = 1.95742494230308167... \tag{3.2.7}$$

Our technique can be applied to prove that $\hat{\alpha}$ and $\hat{\mu}$ are transcendental over $\mathbb{Q}$. We illustrate it for $\hat{\mu}$.

For any $x \in \mathbb{R}$, the function $\ell(\alpha, \mu)$ is bounded from above and achieves its maximum on $[0, 1] \times \mathbb{R}$. If $x > 0$ is large, then any global maximum $(\hat{\alpha}, \hat{\mu})$ of ℓ is in the interior of $[0, 1] \times \mathbb{R}$ and satisfies $0 < \hat{\mu} \leq x$. According to a `Mathematica` computation, the choice $x \geq 1.56125...$ suffices for this. Assume that this holds. Setting the two partial derivatives equal to zero and eliminating the unknown α in a further `Mathematica` computation, the critical equation for μ is found to be

$$(x - \mu)e^{\mu^2} - x + \mu e^{-\mu(2x-\mu)} = 0. \tag{3.2.8}$$

Suppose for contradiction that both x and $\hat{\mu}$ are algebraic numbers over $\mathbb{Q}$. Since $0 < \hat{\mu} \leq x$, we have $-\hat{\mu}(2x - \hat{\mu}) < 0 < \hat{\mu}^2$. Hence $u_1 = \hat{\mu}^2$, $u_2 = 0$ and $u_3 = -\hat{\mu}(2x - \hat{\mu})$ are distinct algebraic numbers. The Lindemann-Weierstrass Theorem implies that e^{u_1}, e^{u_2} and e^{u_3} are linearly independent over the field of algebraic numbers. However, from (3.2.8) we know that

$$(x - \hat{\mu}) \cdot e^{u_1} - x \cdot e^{u_2} + \hat{\mu} \cdot e^{u_3} = 0.$$

This is a contradiction. We conclude that the number $\hat{\mu}$ is transcendental over $\mathbb{Q}$.

Table 3.3.1.: Seven critical points of the log-likelihood function in Theorem 3.3.1 with $K = 7$.

k	α	μ_1	μ_2	σ_1	σ_2	log-likelihood
1	0.1311958	1.098998	4.553174	0.09999497	1.746049	-27.29187821475
2	0.1032031	2.097836	4.330408	0.09997658	1.988948	-28.63974638055
3	0.07883084	3.097929	4.185754	0.09997856	2.06374	-29.15502775347
4	0.06897294	4.1	4.1	0.1	2.07517	-29.28589815510
5	0.07883084	5.102071	4.014246	0.09997856	2.06374	-29.15502775347
6	0.1032031	6.102164	3.869592	0.09997658	1.988948	-28.63974638055
7	0.1311958	7.101002	3.646826	0.09999497	1.746049	-27.29187821475

3.3. Many Critical Points

Theorem 3.1.2 shows that Gaussian mixtures do not admit an ML degree. This raises the question of how to find any bound for the number of critical points.

Problem 5. Does there exist a universal bound on the number of non-trivial critical points for the log-likelihood function of the mixture of two univariate Gaussians? Or, can we find a sequence of samples on the real line such that the number of non-trivial critical points increases beyond any bound?

We shall resolve this problem by answering the second question affirmatively. The idea behind our solution is to choose a sample consisting of many well-separated clusters of size 2. Then each cluster gives rise to a distinct non-trivial critical point $(\hat{\alpha}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ of the log-likelihood function ℓ from (3.2.2). We propose one particular choice of data, but many others would work too.

Theorem 3.3.1. *Fix sample size $N = 2K$ for $K \geq 2$, and take the ordered sample $(x_1, \ldots, x_{2K}) = (1, 1.2,\ 2, 2.2,\ \ldots, K, K{+}0.2)$. Then, for each $k \in \{1, \ldots, K\}$, the log-likelihood function ℓ from* (3.2.2) *has a non-trivial critical point with $k < \hat{\mu}_1 < k + 0.2$. Hence, there are at least K non-trivial critical points.*

Before turning to the proof, we offer a numerical illustration.

Example 3.3.2. For $K = 7$, we have $N = 14$ data points in the interval $[1, 7.2]$. Running the EM algorithm (as explained in the proof of Theorem 3.3.1 below) yields the non-trivial critical points reported in Table 3.3.1. Their μ_1 coordinates are seen to be close to the cluster midpoints $k + 0.1$ for all k. The observed symmetry under reversing the order of the rows also holds for all larger K.

Our proof of Theorem 3.3.1 will be based on the *EM algorithm*. We first recall this algorithm. Let $f_{\alpha,\mu,\sigma}$ be the mixture density from (3.2.1), and let

$$f_j(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_j}\,\exp\Big(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\Big), \quad j=1,2,$$

be the two Gaussian component densities. Define

$$\gamma_i = \frac{\alpha\cdot f_1(x_i)}{f_{\alpha,\mu,\sigma}(x_i)}, \tag{3.3.1}$$

which can be interpreted as the conditional probability that data point x_i belongs to the first mixture component. Further, define $N_1 = \sum_{i=1}^N \gamma_i$ and $N_2 = N - N_1$, which are expected cluster sizes. Following [Bis06, Section 9.2.2], the likelihood equations for our model can be written in the following fixed-point form:

$$\alpha = \frac{N_1}{N}, \tag{3.3.2}$$

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^N \gamma_i x_i, \qquad \mu_2 = \frac{1}{N_2}\sum_{i=1}^N (1-\gamma_i) x_i, \tag{3.3.3}$$

$$\sigma_1 = \frac{1}{N_1}\sum_{i=1}^N \gamma_i (x_i-\mu_1)^2, \qquad \sigma_2 = \frac{1}{N_2}\sum_{i=1}^N (1-\gamma_i)(x_i-\mu_2)^2. \tag{3.3.4}$$

In the present context, the EM algorithm amounts to solving these equations iteratively. More precisely, consider any starting point $(\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2)$. Then the E-step ("expectation") computes the estimated frequencies γ_i via (3.3.1). In the subsequent M-step ("maximization"), one obtains a new parameter vector $(\alpha,\mu_1,\mu_2,\sigma_1,\sigma_2)$ by evaluating the right-hand sides of the equations (3.3.2)-(3.3.4). The two steps are repeated until a fixed point is reached, up to the desired numerical accuracy. The updates never decrease the log-likelihood. For our problem it can be shown that the algorithm will converge to a critical point; see e.g. [RW84].

Proof. (of Theorem 3.3.1) Fix $k \in \{1,\dots,K\}$. We choose starting parameter values to suggest that the pair $(x_{2k-1}, x_{2k}) = (k, k+0.2)$ belongs to the first mixture component, while the rest of the sample belongs to the second. Explicitly,

we set

$$\alpha = \frac{2}{N} = \frac{1}{K},$$
$$\mu_1 = k + 0.1, \qquad \mu_2 = \frac{K^2 + 1.2K - 2k - 0.2}{2(K-1)},$$
$$\sigma_1 = 0.1, \qquad \sigma_2 = \frac{\sqrt{\frac{1}{12}K^4 - \frac{1}{3}K^3 + (k - \frac{43}{75})K^2 - (k^2 - k + \frac{14}{75})K + 0.01}}{K-1}.$$

We shall argue that, when running the EM algorithm, the parameters will always stay close to these starting values. Specifically, we claim that, throughout all EM iterations, the parameter values satisfy the inequalities

$$\frac{1}{4K} \leq \alpha \leq \frac{1}{K}, \qquad 0.09 \leq \mu_1 - k \leq 0.11, \qquad 0.099 \leq \sigma_1 \leq 0.105, \tag{3.3.5}$$

$$\frac{K}{2} + 0.1 \ \leq \mu_2 \leq \frac{K}{2} + 1.1, \tag{3.3.6}$$

$$\sqrt{\frac{K^2}{12} - \frac{K}{6} + 0.01} \ \leq \sigma_2 \leq \sqrt{\frac{K^2}{12} + \frac{K}{12} + 0.01}. \tag{3.3.7}$$

The starting values proposed above obviously satisfy the inequalities in (3.3.5), and it is not difficult to check that (3.3.6) and (3.3.7) are satisfied as well. To prove the theorem, it remains to show that (3.3.5)-(3.3.7) continue to hold after an EM update.

In the remainder, we assume that $K > 22$. For smaller values of K the claim of the theorem can be checked by running the EM algorithm. In particular, for $K > 3$, the second standard deviation satisfies the simpler bounds

$$\frac{K}{\sqrt{12}} - \frac{\sqrt{3}}{5} \leq \sigma_2 \leq \frac{K}{\sqrt{12}} + \frac{\sqrt{3}}{12}. \tag{3.3.8}$$

A key property is that the quantity γ_i, computed in the E-step, is always very close to zero for $i \neq 2k-1, 2k$. To see why, rewrite (3.3.1) as

$$\gamma_i = \frac{1}{1 + \frac{1-\alpha}{\alpha}\frac{f_2(x_i)}{f_1(x_i)}} = \frac{1}{1 + \frac{1-\alpha}{\alpha}\frac{\sigma_1}{\sigma_2}\exp\left\{\frac{1}{2}\left((\frac{x_i-\mu_1}{\sigma_1})^2 - (\frac{x_i-\mu_2}{\sigma_2})^2\right)\right\}}.$$

Since $\alpha \leq 1/K$, we have $\frac{1-\alpha}{\alpha} \geq K - 1$. On the other hand, $\frac{\sigma_1}{\sigma_2} \geq \frac{0.099}{K/\sqrt{12}+\sqrt{3}/12}$. Using that $K > 22$, their product is thus bounded below by 0.3209. Turning to the exponential term, the second inequality in (3.3.5) implies that $|x_i - \mu_1| \geq 0.89$ for $i = 2k-2$ or $i = 2k+1$, which index the data points closest to the kth pair.

Using (3.3.8), we obtain

$$\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2-\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2 \;\geq\; \left(\frac{0.89}{0.105}\right)^2-\left(\frac{K/2+0.1}{K/\sqrt{12}-\sqrt{3}/5}\right)^2 \;\geq\; 67.86.$$

From $e^{33.93} > 5.4\cdot 10^{14}$, we deduce that $\gamma_i < 10^{-14}$. The exponential term becomes only smaller as the considered data point x_i move away from the kth pair. As $|i-(2k-1/2)|$ increases, γ_i decreases and can be bounded above by a geometric progression starting at 10^{-14} and with ratio 10^{-54}. This makes γ_i with $i \neq 2k, 2k-1$ negligible. Indeed, from the limit of geometric series, we have

$$s_1 = \sum_{i\neq 2k-1,2k} \gamma_i \;<\; 10^{-13}, \tag{3.3.9}$$

and similarly, $s_2 = \sum_{i\neq 2k-1,2k} \gamma_i(x_i-k)$ satisfies

$$|s_2| \;=\; |\gamma_{2k-2}(-0.8)+\gamma_{2k+1}(1)+\gamma_{2k-3}(-1)+\gamma_{2k+2}(1.2)+\ldots| \;<\; 10^{-13}. \tag{3.3.10}$$

The two sums s_1 and s_2 are relevant for the M-step.

The probabilities γ_{2k-1} and γ_{2k} give the main contribution to the averages that are evaluated in the M-step. They satisfy $0.2621 \leq \gamma_{2k-1}, \gamma_{2k} \leq 0.9219$. Moreover, we may show that the values of γ_{2k-1} and γ_{2k} are similar, namely:

$$0.8298 \leq \frac{\gamma_{2k-1}}{\gamma_{2k}} \leq 1.2213, \tag{3.3.11}$$

which we prove by writing

$$\frac{\gamma_{2k-1}}{\gamma_{2k}} = \frac{1+y\exp(z/2)}{1+y},$$

and using $K > 22$ to bound

$$\begin{aligned} y &= \frac{1-\alpha}{\alpha}\frac{\sigma_1}{\sigma_2}\exp\left\{\frac{1}{2}\left(\left(\frac{k-\mu_1}{\sigma_1}\right)^2-\left(\frac{k-\mu_2}{\sigma_2}\right)^2\right)\right\}, \\ z &= \frac{0.4(k-\mu_1)+0.04}{\sigma_1^2}-\frac{0.4(k-\mu_2)+0.04}{\sigma_2^2}. \end{aligned}$$

Bringing it all together, we have

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N}\gamma_i x_i = k+\frac{0.2\gamma_{2k}+s_2}{\gamma_{2k-1}+\gamma_{2k}+s_1}.$$

Using $\gamma_{2k} + \gamma_{2k-1} > 0.5$ and (3.3.10), as well as the lower bound in (3.3.11), we find

$$\mu_1 - k \leq \frac{0.2\gamma_{2k}}{\gamma_{2k-1} + \gamma_{2k}} + \frac{s_2}{\gamma_{2k-1} + \gamma_{2k}} \leq \frac{0.2\gamma_{2k}}{0.8298\gamma_{2k} + \gamma_{2k}} + 10^{-12} \leq 0.11.$$

Using the upper bound in (3.3.11), we also have $0.09 \leq \mu_1 - k$. Hence, the second inequality in (3.3.5) holds.

The inequalities for the other parameters are verified similarly. For instance,

$$\frac{1}{4K} < \frac{0.2621 + 0.2621}{2K} \leq \frac{\gamma_{2k-1} + \gamma_{2k} + s_1}{2K} \leq \frac{0.9219 + 0.9219 + 10^{-13}}{2K} < \frac{1}{K}$$

holds for $\alpha = \frac{N_1}{N}$. Therefore, the first inequality in (3.3.5) continues to be true.

We conclude that running the EM algorithm from the chosen starting values yields a sequence of parameter vectors that satisfy the inequalities (3.3.5)-(3.3.8). The sequence has at least one limit point, which must be a non-trivial critical point of the log-likelihood function. Therefore, for every $k = 1, \ldots, K$, the log-likelihood function has a non-trivial critical point with $\mu_1 \in (k, k + 0.2)$. □

3.4. Further Discussion

We showed that the maximum likelihood estimator (MLE) in Gaussian mixture models is not an algebraic function of the data, and that the log-likelihood function may have arbitrarily many critical points. Hence, in contrast to the models studied so far in algebraic statistics [BHSPR07, DSS08, GDP+12, SU10], there is no notion of an ML degree for Gaussian mixtures. However, certified likelihood inference may still be possible, via transcendental root separation bounds, as in [CCK+06, CPPY06].

Remark 3.4.1. The *Cauchy-location model*, treated in [Ree85], is an example where the ML estimation is algebraic but the ML degree, and also the maximum number of local maxima, depends on the sample size and increases beyond any bound.

Remark 3.4.2. The ML estimation problem admits a population/infinite-sample version. Here the maximization of the likelihood function is replaced by *minimization of the Kullback-Leibler divergence* between a given data-generating distribution and the distributions in the model [Her05]. The question of whether this population problem is subject to local but not global maxima was raised in [Sre07]—in the context of Gaussian mixtures with known and equal variances. Examples where this can actually happen have been given in [JZB+16], meaning estimation is not that easy even in simple submodels. Further, it is known that the Kullback-Leibler divergence for such Gaussian mixtures is not an analytic function [Wat09, §7.8]. Readers of Japanese can find details in [Wat04].

However, in the introduction we saw that Pearson's classical method of moments only involved the solution of polynomial equation systems. Thus, this approach will be in the scope of algebraic geometry. This will be the main focus of the next chapter. In Section 4.3 we illustrate the behavior of Pearson's method for the sample used in Theorem 3.3.1.

4. Method of Moments and Moment Varieties

In the previous chapter we saw that while the ML estimates for a Gaussian are algebraic, this breaks down with an honest Gaussian mixture ($k > 1$). The method of moments seems to be more appropriate algebraically, for these and for any model with moments that are polynomial in the parameters [BS15]. In this chapter we take a closer look at the Gaussian moments and revisit Pearson's method from a computational algebra point of view. We also introduce the algebraic varieties that will encode the moments of Gaussians.

4.1. Moments and Cumulants

Let us first recall what moments are.

Definition 4.1.1. Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be the probability density function for a random vector $X = (X_1, X_2, \ldots, X_n)$. The *moment* with indices $i_1, i_2, \ldots, i_n \geq 0$ is

$$m_{i_1 i_2 \cdots i_n} = \mathbb{E}(X_1^{i_1} X_2^{i_2} \cdots X_n^{i_n}) = \int_{\mathbb{R}^n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

With compact notation $i = (i_1, i_2, \ldots, i_n)$:

$$m_i = \int_{\mathbb{R}^n} x^i f(x) dx.$$

The integer $i_1 + i_2 + \cdots + i_n$ is the *order* of the moment. Note that $m_{00\cdots0} = 1$.

Example 4.1.2. Let $n = 1$. The first moments of a Gaussian $X \sim N(\mu, \sigma^2)$ are:

$$\begin{aligned}
m_1 &= \mu \\
m_2 &= \mu^2 + \sigma^2 \\
m_3 &= \mu^3 + 3\mu\sigma^2 \\
m_4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
m_5 &= \mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4 \\
m_6 &= \mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6.
\end{aligned}$$

We note that each of them is a polynomial in the parameters μ, σ^2. We will see this is true in general.

It is also relatively easy to generate the higher order moments from lower order ones:

Lemma 4.1.3. *Let $n = 1$ and $X \sim N(\mu, \sigma^2)$, then its moments satisfy the recurrence relation:*

$$m_{i+1} = \mu m_i + i\sigma^2 m_{i-1}$$

for $i \geq 1$. In particular, every m_i is a polynomial in $\mathbb{Z}[\mu, \sigma^2]$.

Proof. This follows from an adequate application of integration by parts:

$$\begin{aligned} m_{i-1} &= \int_{-\infty}^{\infty} x^{i-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \left[\frac{x^i}{i}\right]' e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \left[\frac{x^i}{i} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right]_{x=-\infty}^{x=\infty} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{x^i}{i} \frac{(x-\mu)}{\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= 0 \quad + \quad \frac{1}{i\sigma^2} m_{i+1} - \frac{\mu}{i\sigma^2} m_i \end{aligned}$$

Multiplying by $i\sigma^2$ gives the desired result. □

Remark 4.1.4. The classical *Hermite* polynomials $H_n(x)$ are recovered as $H_n(\mu) = m_n$ when substituting the negative 'variance' $\sigma^2 = -1$.

Example 4.1.5. Let $n = 2$:

$$X \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}\right).$$

Then one can compute:

- moments of order 1: $m_{10} = \mu_1$, $m_{01} = \mu_2$
- moments of order 2: $m_{20} = \mu_1^2 + \sigma_{11}$, $m_{11} = \mu_1\mu_2 + \sigma_{12}$, $m_{02} = \mu_2^2 + \sigma_{22}$
- moments of order 3: $m_{30} = \mu_1^3 + 3\mu_1\sigma_{11}$, $m_{21} = \mu_1^2\mu_2 + 2\mu_1\sigma_{12} + \mu_2\sigma_{11}$, $m_{12} = \mu_1\mu_2^2 + 2\mu_2\sigma_{12} + \mu_1\sigma_{22}$, $m_{03} = \mu_2^3 + 3\mu_2\sigma_{22}$

One useful way to obtain and represent the moments is as the coefficients of a generating function.

Definition 4.1.6. Let $X \sim N(\mu, \Sigma)$, then its *moment generating function* is $M_X(t) = e^{t^T\mu+\frac{1}{2}t^T\Sigma t}$. Writing out the entries we have that $M_X(t_1, t_2, \ldots, t_n)$ is given by

$$\sum_{i_1,i_2,\ldots,i_n\geq 0} \frac{m_{i_1 i_2 \cdots i_n}}{i_1! i_2! \cdots i_n!} t_1^{i_1} t_2^{i_2} \cdots t_n^{i_n} = \exp\left(\sum_{r=1}^{n} \mu_r t_r\right) \cdot \exp\left(\frac{1}{2}\sum_{i,j=1}^{n} \sigma_{ij} t_i t_j\right) \quad (4.1.1)$$

In this way, a Taylor series expansion of $M_X(t)$ reveals the moments $m_{i_1 i_2 \cdots i_n}$.

There is another coordinate system that is also used to measure properties of a distribution: *cumulants*. It can simplify many expressions. For this reason we will frequently present expressions in them. Here is the definition.

Definition 4.1.7. Let $M_X(t_1, t_2, \ldots, t_n)$ be a moment generating function. Then the *cumulant generating function* is given by $K_X(t) = \log(M_X(t))$. Its expansion

$$K_X(t_1, \ldots, t_n) = \sum_{i_1,i_2,\ldots,i_n\geq 0} \frac{k_{i_1 i_2 \cdots i_n}}{i_1! i_2! \cdots i_n!} t_1^{i_1} t_2^{i_2} \cdots t_n^{i_n} \quad (4.1.2)$$

gives the *cumulant* $k_{i_1 i_2 \cdots i_n}$ of index $i_1 i_2 \ldots i_n$.

Note that $m_{00..0} = 1$ implies $k_{00..0} = 0$.

Example 4.1.8. Let $n = 1$. The first six cumulants in terms of the first six moments are:

$$\begin{aligned}
k_1 &= m_1 \\
k_2 &= m_2 - m_1^2 \\
k_3 &= m_3 - 3m_1m_2 + 2m_1^3 \\
k_4 &= m_4 - 4m_1m_3 - 3m_2^2 + 12m_1^2m_2 - 6m_1^4 \\
k_5 &= m_5 - 5m_1m_4 - 10m_2m_3 + 20m_1^2m_3 + 30m_1m_2^2 - 60m_1^3m_2 + 24m_1^5 \\
k_6 &= m_6 - 6m_1m_5 - 15m_2m_4 + 30m_1^2m_4 - 10m_3^2 + 120m_1m_2m_3 \\
&\quad - 120m_1^3m_3 + 30m_2^3 - 270m_1^2m_2^2 + 360m_1^4m_2 - 120m_1^6
\end{aligned} \quad (4.1.3)$$

Remark 4.1.9. If the assumption $m_1 = 0$ is made, then the first three cumulants and the first three moments coincide: $k_1 = m_1, k_2 = m_2, k_3 = m_3$.

Remark 4.1.10. In Pearson's polynomial (1.1.4) (where $m_1 = 0$), the simplifying expressions $\lambda_4 = 9m_2^2 - 3m_4, \lambda_5 = 30m_2m_3 - 3m_5$ are precisely the cumulant multiples $-3k_4$ and $-3k_5$. In the next section we will derive Pearson's polynomial entirely in cumulants, simplifying expression (1.1.4) further.

One can always express moments of order $\leq d$ as polynomials in cumulants of order $\leq d$, and vice versa, via the generating function identities:

$$M_X(t) = \exp(K_X(t)) \quad \text{and} \quad K_X(t) = \log(M_X(t)). \tag{4.1.4}$$

Finally, we mention a very important property of the Gaussian distribution with respect to its cumulants.

Example 4.1.11. For the Gaussian $X \sim N(\mu, \Sigma)$ we have the moment generating function $M_X(t) = e^{t^T\mu + \frac{1}{2}t^T\Sigma t}$, so the cumulant generating function $K = \log(M)$ is simply:

$$K_X(t) = t^T\mu + \frac{1}{2}t^T\Sigma t.$$

Thus, all Gaussian cumulants $k_{i_1 i_2 \ldots, i_n}$ of order greater than 2 vanish!

4.2. The Pearson Polynomial

Recall that the method of moments in statistics was introduced by Pearson in his 1894 paper [Pea94]. In our view, this can be regarded as the beginning of Algebraic Statistics. In this section we revisit Pearson's computation and related work of Lazard [Laz04], and we extend them further.

Pearson's method of moments identifies the parameters in a mixture of two univariate Gaussians. In order to simplify the use of subindices in the derivations that follow, in this section we will denote the two Gaussians as: $N(\mu, \sigma^2)$ and $N(\nu, \tau^2)$ (instead of μ_1, μ_2 and σ_1, σ_2).

Suppose the first five moments m_1, m_2, m_3, m_4, m_5 are given numerically from data. Pearson [Pea94] solves the corresponding five equations

$$\begin{array}{rcc} m_0 &=& 1 \\ m_1 &=& \alpha\mu + (1-\alpha)\nu \\ m_2 &=& \alpha(\mu^2+\sigma^2) + (1-\alpha)(\nu^2+\tau^2) \\ m_3 &=& \alpha(\mu^3+3\mu\sigma^2) + (1-\alpha)(\nu^3+3\nu\tau^2) \\ m_4 &=& \alpha(\mu^4+6\mu^2\sigma^2+3\sigma^4) + (1-\alpha)(\nu^4+6\nu^2\tau^2+3\tau^4) \\ m_5 &=& \alpha(\mu^5+10\mu^3\sigma^2+15\mu\sigma^4) + (1-\alpha)(\nu^5+10\nu^3\tau^2+15\nu\tau^4) \end{array} \tag{4.2.1}$$

for the five unknowns $\alpha, \mu, \nu, \sigma, \tau$.

As mentioned in the previous section, we will prefer to work in cumulants. We can obtain numerical values for k_1, k_2, k_3, k_4, k_5 from the formulas in (4.1.3). We will derive a simplified version of Pearson's polynomial (1.1.4) in cumulants.

Indeed, to solve the system (4.2.1) the crucial first step is to find the roots of the following univariate polynomial of degree 9 in p.

Proposition 4.2.1. *The product of normalized means $p = (\mu - m_1)(\nu - m_1)$ is a root of the polynomial*

$$\begin{array}{c} 8p^9 + 28k_4p^7 + 12k_3^2p^6 + (24k_3k_5 + 30k_4^2)p^5 + (148k_3^2k_4 - 6k_5^2)p^4 \\ +(96k_3^4 + 9k_4^3 - 36k_3k_4k_5)p^3 + (-21k_3^2k_4^2 - 24k_3^3k_5)p^2 - 32k_3^4k_4p - 8k_3^6. \end{array} \tag{4.2.2}$$

Proof. We first prove the identity (4.2.2) under the assumption that the empirical mean is zero:

$$m_1 \;=\; \alpha\mu + (1-\alpha)\nu \;=\; 0. \tag{4.2.3}$$

In order to work modulo the symmetry that switches the two Gaussian components, we replace the unknown means μ and ν by their first two elementary symmetric polynomials:

$$p = \mu\nu \quad \text{and} \quad s = \mu + \nu. \tag{4.2.4}$$

In [Pea94], Pearson applies considerable effort and cleverness to eliminating the unknowns $\mu, \nu, \sigma, \tau, \alpha$ from the constraints (4.2.1),(4.2.3),(4.2.4). We here offer a derivation that can be checked easily in a computer algebra system. We start by solving (4.2.3) for α. Substituting

$$\alpha \;=\; \frac{-\nu}{\mu - \nu}. \tag{4.2.5}$$

into $k_2 = \alpha(\mu^2+\sigma^2)+(1-\alpha)(\nu^2+\tau^2)$, we obtain the relation $k_2 = -R_1 - p$, where

$$R_1 \;=\; \frac{\sigma^2\nu - \tau^2\mu}{\mu - \nu}. \tag{4.2.6}$$

This the first of a series of semi-invariants R_i that appear naturally when trying to write the cumulant expressions in terms of p and s. In the next instance, by letting

$$R_2 \;=\; \frac{\sigma^2 - \tau^2}{\mu - \nu} \tag{4.2.7}$$

we can write $k_3 = -(3R_2 + s)p$. In a similar way, we obtain

$$\begin{array}{rcl} k_4 &=& 3R_3 + p(p - s^2) - 3k_2^2 \\ k_5 &=& 5R_4p - sp(s^2 - 2p) - 10k_2k_3 \\ k_6 &=& 15R_5 - p(s^4 - 3s^2p + p^2) - 15k_2^3 - 15k_2k_4 - 10k_3^2 \end{array} \tag{4.2.8}$$

where

$$\begin{array}{rcl} R_3 & = & (\mu\sigma^4 - \nu\tau^4 + 2\mu\nu^2\tau^2 - 2\mu^2\nu\sigma^2)/(\mu-\nu) \\ R_4 & = & (3\tau^4 - 3\sigma^4 + 2\nu^2\tau^2 - 2\mu^2\sigma^2)/(\mu-\nu) \\ R_5 & = & (\mu^4\nu\sigma^2 - \mu\nu^4\tau^2 + 3\mu^2\nu\sigma^4 - 3\mu\nu^2\tau^4 + \nu\sigma^6 - \mu\tau^6)/(\mu-\nu). \end{array} \tag{4.2.9}$$

It turns out that R_3, R_4, R_5 are not independent of R_1, R_2. Namely, we find

$$\begin{array}{rcl} R_3 & = & R_1^2 + 2pR_1 - 2spR_2 - pR_2^2 \\ R_4 & = & 2sR_1 + 6R_1R_2 + 2(p-s^2)R_2 - 3sR_2^2 \\ R_5 & = & -R_1^3 - 3pR_1^2 + (s^2p - p^2)R_1 + 6spR_1R_2 + 3pR_1R_2^2 \\ & & +(2sp^2 - s^3p)R_2 + (3p^2 - 3s^2p)R_2^2 - spR_2^2. \end{array} \tag{4.2.10}$$

We now express the three right hand sides in terms of p, s, k_2, k_3 using the relations

$$R_1 \,=\, -k_2 - p \qquad \text{and} \qquad R_2 \,=\, -\frac{s}{3} - \frac{k_3}{3p}. \tag{4.2.11}$$

Plugging the resulting expressions for R_3 and R_4 into the first two equations of (4.2.8), we get

$$\begin{array}{rcl} -2p^2s^2 - 4spk_3 + 6p^3 + 3k_4p + k_3^2 & = & 0, \\ -2p^2s^3 + 4p^3s + 5sk_3^2 - 20p^2k_3 + 3k_5p & = & 0. \end{array} \tag{4.2.12}$$

Pearson's polynomial (4.2.2) is the resultant of these two polynomials with respect to s.

The proof is completed by noting that the entire derivation is invariant under replacing the parameters for the means μ and ν by the normalized means $\mu - m_1$ and $\nu - m_1$. □

Remark 4.2.2. Gröbner bases reveal the following consequence of the two equations in (4.2.12):

$$(4p^3k_3 - 4k_3^3 - 6pk_3k_4 - 2p^2k_5)s + 4p^5 + 14p^2k_3^2 + 8p^3k_4 + k_3^2k_4 + 3pk_4^2 - 2pk_3k_5 \;=\; 0.$$

This furnishes an expression for s as rational function in the quantities k_3, k_4, p. Note that this expression and (4.2.2) do not depend on k_2 at all. The second moment m_2 is only involved via k_4.

Once the Pearson polynomial has been obtained, the method of moments for $k = 2, n = 1$ works as follows. From the data, compute the empirical moments $m_1, \ldots, m_5$, and derive the cumulants k_3, k_4, k_5 via (4.1.3). Next compute the nine complex zeros of the Pearson polynomial (4.2.2). We are only interested in zeros p that are real and non-positive, because $(\mu - m_1)(\nu - m_1) \leq 0$. All other zeros get discarded. For each non-positive zero p of (4.2.2), compute the corresponding

s from the equation in Remark 4.2.2. By (4.2.4), we obtain μ and ν as the two zeros of the equation $x^2 - sx + p = 0$. The mixture parameter α is given by (4.2.5). Finally, since R_1 and R_2 are now known by (4.2.11), we obtain σ^2 and τ^2 by solving an inhomogeneous system of two linear equations, (4.2.6) and (4.2.7). Note that what we described above computes $\mu - m_1, \nu - m_1$, so we should add m_1 to recover μ and ν.

The algorithm in the previous paragraph works well when m_1, m_2, m_3, m_4, m_5 are general enough. For special values of the empirical moments, however, one might encounter zero denominators and other degeneracies. Extra care is needed in those cases. We implemented a complete method of moments (for $n = 1, k = 2$) in the statistics software `R`. See Appendix C for the corresponding code and scripts.

Pearson [Pea94] applied his method to measurements taken from crabs in the Bay of Naples. His data set is the histogram presented in Figure 1.1.1.

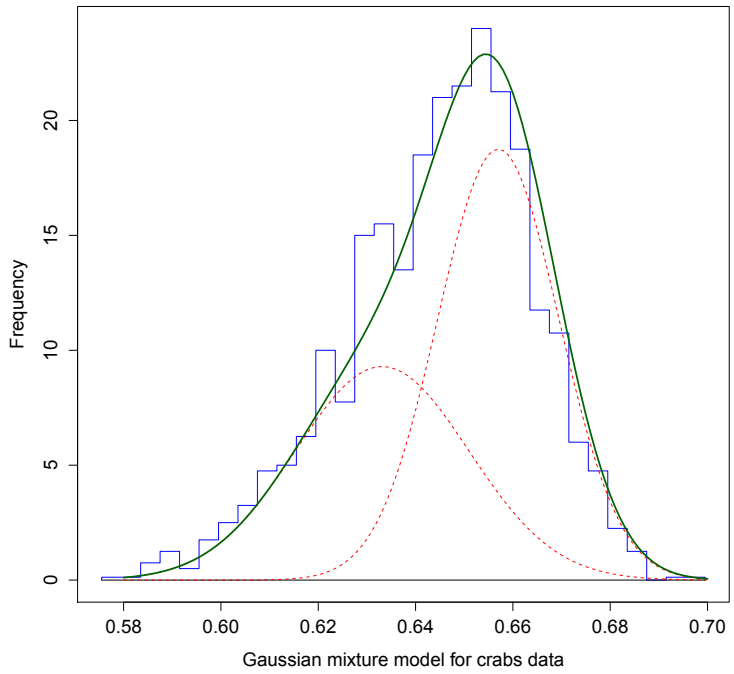

Figure 4.2.1.: Approximation of crabs data by a mixture of two Gaussians.

Pearson computes the empirical moments from the crab data, and he takes these as the numerical values for m_1, m_2, m_3, m_4, m_5. The resulting nonic polynomial (4.2.2) has three real roots, two of which are non-positive. One computes the model parameters as above. At this point, Pearson has two statistically meaningful solutions. To choose between them, he computes m_6 in each case, and selects the

model that is closest to the empirical m_6. Pearson's method gives the parameters $\mu = 0.633, \sigma = 0.018, \nu = 0.657, \tau = 0.012, \alpha = 0.414$. The resulting probability density function scaled to match the histogram and its mixture components are shown in Figure 4.2.1.

Suppose the moments $m_1, \ldots, m_6$ are measured exactly from a mixture of two univariate Gaussians. Pearson's experiment suggests that if we add the equation for the sixth moment m_6:

$$m_6 = \alpha(\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6) + (1-\alpha)(\nu^6 + 15\nu^4\tau^2 + 45\nu^2\tau^4 + 15\tau^6),$$

to the polynomial system (4.2.1), the parameters that created the 6 moments will be the unique solution of the new system. In fact, we can obtain a single relation among the first six moments that identifies mixtures of two univariate Gaussians.

Theorem 4.2.3. *There exists an irreducible polynomial relation of degree 39 in $m_1, m_2, m_3, m_4, m_5, m_6$ with 31154 terms that vanishes on the six moments of a mixture of two univariate Gaussians. This polynomial has degrees $33, 32, 23, 17, 12, 9$ in $m_1, m_2, m_3, m_4, m_5, m_6$ respectively.*

Proof. Using (4.2.10) and (4.2.11), the last equation in (4.2.8) translates into

$$\begin{array}{c} 5sk_3^3 - 144p^5 + (72s^2 - 270k_2)p^4 + (90s^2k_2 + 180sk_3 - 4s^4)p^3 + \\ (-135k_2k_4 + 180sk_2k_3 - 30s^3k_3 - 90k_3^2 - 9k_6)p^2 - 30k_3^2(s^2 + \frac{3}{2}k_2)p = 0. \end{array} \quad (4.2.13)$$

We now eliminate the unknowns p and s from the three equations in (4.2.12) and (4.2.13). After removing an extraneous factor k_3^3, we obtain an irreducible polynomial in k_3, k_4, k_5, k_6 of degree 23 with 195 terms. This polynomial is also mentioned in [Laz04, Proposition 12].

We finally substitute the expressions in (4.1.3) to get an inhomogeneous polynomial in $m_1, m_2, \ldots, m_6$ of degree 39 with 31154 terms. At this point, we check that this polynomial vanishes at the parametrization (4.2.1). The degree in each moment m_i is read off by inspection. □

We will come back to this result and make it more precise in terms of secants of moment varieties in Chapter 5.

4.3. Comparison to Maximum Likelihood

To make a first comparison between the two methods, we now note what happens when we already apply the Method of Moments (MOM) when $k = 1$:

Example 4.3.1. Sample $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ from $X \sim N(\mu, \Sigma)$. The vector of first order moments is given by $M_1 = \mu$ and the matrix of second order moments

is $M_2 = \mu\mu^T + \Sigma$. Thus, the method of moments equations are:

$$\hat{M}_1 = \hat{\mu} \qquad \hat{M}_2 = \hat{\mu}\hat{\mu}^T + \hat{\Sigma}$$

which give immediately the solutions:

$$\hat{\mu} = \hat{M}_1 \qquad \hat{\Sigma} = \hat{M}_2 - \hat{M}_1 \hat{M}_1^T,$$

where the sample moments are:

$$\hat{M}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \hat{M}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T.$$

This means that the MOM estimates are given by

$$\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \\ \hat{\Sigma} &= \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T - \bar{x}\bar{x}^T = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \end{aligned}$$

We see then the nice fact that the method of moments estimates for $k = 1$ coincide with the ones from maximum likelihood (Example 3.1.1)! □

Remark 4.3.2. Here we arranged the first order moments in a vector M_1 and the second order moments into a matrix M_2. In general, in dimension n, all moments of order d can be arranged naturally in a symmetric *tensor* of order d with format $n \times n \times \cdots \times n$. For a friendly introduction to tensors and their eigenvectors see [Stu16]. Of particular interest are tensors that admit an orthogonal decomposition, as is the case of real symmetric matrices for $d = 2$ [Rob16]. We will come back to this point of view in Section 6.1.

In Section 3.3, the sample consisting of the following $N = 2K$ data points was examined:

$$1,\ 1.2,\ 2,\ 2.2,\ 3,\ 3.2,\ 4, \ldots, K,\ K + 0.2 \qquad (\text{for } K > 1). \tag{4.3.1}$$

Its main purpose was to illustrate that, unlike most models studied in Algebraic Statistics, there is no notion of *ML degree* for a mixture of two Gaussians. In fact, the particular sample in (4.3.1) has the property that, as K increases, the number of critical points of the log-likelihood function grows without any bound. We proved in Theorem 3.3.1 that for each 'cluster' or pair $(k, k + 0.2)$, one can find a non-trivial critical point $(\hat{\alpha}, \hat{\mu}, \hat{\nu}, \hat{\sigma}, \hat{\tau})$ of the likelihood equations such that the mean estimate $\hat{\mu}$ lies between them.

Now let's apply Pearson's method of moments to this sample. The special nature of the data raises some interesting considerations. As we shall see, the even spacing of the points in the list (4.3.1) implies that all empirical cumulants of odd order ≥ 3 vanish:

$$k_3 \,=\, k_5 \,=\, k_7 \,=\, k_9 \,=\, \cdots \,=\, 0. \tag{4.3.2}$$

Let us analyze what happens when applying the method of moments to *any* sample that satisfies (4.3.2). Under this hypothesis Pearson's polynomial (4.2.2) factors as follows:

$$8p^9 + 28p^7k_4 + 30p^5k_4^2 + 9p^3k_4^3 \;=\; 8p^3\left(p^2 + \frac{3}{2}k_4\right)^2\left(p^2 + \frac{1}{2}k_4\right) \;=\; 0. \tag{4.3.3}$$

Recall that p represents $p = (\mu - m_1)(\nu - m_1)$. The first root of the Pearson polynomial is $p = 0$. This implies $m_1 = \mu$ or $m_1 = \nu$. Since m_1 is the weighted average of μ and ν, we conclude that the means are equal: $m_1 = \mu = \nu$. However, the equal-means model cannot be recovered from the first five moments. To see this, note that the equations for cumulants $k_1 = 0$, $k_3 = 0$ and $k_5 = 0$ become $0 = 0$, yielding no information on the remaining three parameters.

If we assume that also the sixth moment m_6 is known from the data, then the parameters can be identified. The original system (4.2.1) under the equal-means model $\mu = \nu = 0$ equals

$$\begin{aligned} m_2 &= \alpha\sigma^2 + (1-\alpha)\tau^2 \\ m_4 &= 3\alpha\sigma^4 + 3(1-\alpha)\tau^4 \\ m_6 &= 15\alpha\sigma^6 + 15(1-\alpha)\tau^6. \end{aligned} \tag{4.3.4}$$

After some rewriting and elimination:

$$\begin{aligned} \alpha(\sigma^2 - \tau^2) &= k_2 - \tau^2 \\ 5k_4(\sigma^2 + \tau^2) &= 10k_2k_4 + k_6 \\ 15k_4(\sigma^2\tau^2) &= 3k_2k_6 + 15k_2^2k_4 - 5k_4^2. \end{aligned} \tag{4.3.5}$$

Assuming $k_4 \neq 0$, this system can be solved easily in radicals for α, σ, τ.

If $k_4 \geq 0$ then $p = 0$ is the only real zero of (4.3.3). If $k_4 < 0$ then two other solutions are:

$$p \,=\, -\sqrt{\frac{-3}{2}k_4} \quad \text{and} \quad p \,=\, -\sqrt{\frac{-1}{2}k_4}. \tag{4.3.6}$$

Note that p must be negative because it is the product of the two normalized means.

The mean of the sample in (4.3.1) is $m_1 = K/2 + 3/5$. The central moments

are

$$m_r \quad = \quad \frac{1}{2K} \cdot \Big(\sum_{i=1}^{K} \big(i - m_1\big)^r + \sum_{i=1}^{K} \big(i - m_1 + \frac{1}{5}\big)^r \Big) \qquad \text{for } r = 2, 3, 4, \ldots \quad (4.3.7)$$

This expression is a polynomial of degree r in K. That polynomial is zero when r is odd. Using (4.1.3), this implies the vanishing of the odd sample cumulants (4.3.2). For even r, we get

$$m_2 = \frac{1}{12}K^2 - \frac{11}{150}, \quad m_4 = \frac{1}{80}K^4 - \frac{11}{300}K^2 + \frac{91}{3750}$$

$$m_6 = \frac{1}{448}K^6 - \frac{11}{800}K^4 + \frac{91}{3000}K^2 - \frac{12347}{656250}.$$

These polynomials nicely simplify to binomials when we substitute the moments into (4.1.3):

$$k_1 = \frac{K}{2} + 0.6, \;\; k_2 = \frac{K^2}{12} - \frac{11}{150}, \;\; k_4 = -\frac{K^4}{120} + \frac{61}{7500}, \;\; k_6 = \frac{K^6}{252} - \frac{7781}{1968750}. \quad (4.3.8)$$

These are the sample cumulants. If our purpose were to estimate the cumulants, these are biased estimators, and *k-statistics* may be preferable. However, we are estimating the moments so we shall use (4.3.8) in our derivation.

Since $K \geq 1$, we have $k_4 < 0$ in (4.3.8). Hence the Pearson polynomial has three distinct real roots. For $p = 0$, substituting (4.3.2) and (4.3.8) into (4.3.5) shows that, for every value of K, there are no positive real solutions for both σ and τ. Thus the method of moments concludes that the sample does *not* come from a mixture of two Gaussians with the same mean.

Next we consider the two other roots in (4.3.6). To recover the corresponding s-values, we use the system (4.2.12) with all odd cumulants replaced by zero:

$$\begin{aligned} p(6p^2 - 2s^2 p + 3k_4) &= 0 \\ 2sp^2(2p - s^2) &= 0 \end{aligned} \qquad (4.3.9)$$

For $p = -\sqrt{\frac{-3}{2}k_4}$, the first equation gives $s \neq 0$, and the second yields a non-real value for s, so this is not viable. For $p = -\sqrt{\frac{-1}{2}k_4}$, we obtain $s = 0$, and this is now a valid solution.

In conclusion, Pearson's method of moments infers a non-equal-means model for the data (4.3.1). Using central moments, i.e. after subtracting $m_1 = K/2 + 3/5$ from each data point, we find $\mu = -\nu = \sqrt[4]{\frac{-k_4}{2}}$. These values lead to $\alpha = \frac{1}{2}$ and

$\sigma = \tau$. The final estimate is $(\alpha, \mu, \sigma^2, \nu, \tau^2) =$

$$\left(\frac{1}{2}, m_1 - \sqrt[4]{\frac{-k_4}{2}}, k_2 - \sqrt{\frac{-k_4}{2}}, m_1 + \sqrt[4]{\frac{-k_4}{2}}, k_2 - \sqrt{\frac{-k_4}{2}}\right). \tag{4.3.10}$$

We are now in a position to compare this estimate to those found by maximum likelihood.

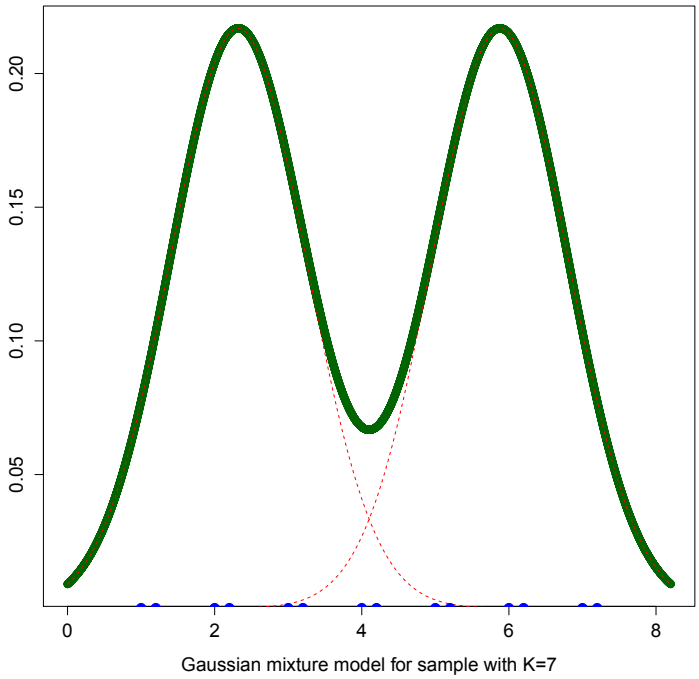

Figure 4.3.1.: The sample data for $K = 7$ (in blue) is approximated by a mixture of two Gaussians via the method of moments. The parameter values are derived in Example 4.3.3.

Example 4.3.3. (sample from Theorem 3.3.1 with $K = 7$) The sample consists of the 14 data points 1,1.2,2,2.2,3,3.2,4,4.2,5,5.2,6,6.2,7,7.2. The method of moments estimator (4.3.10) $(\alpha, \mu, \sigma, \nu, \tau)$ evaluates to

$$\left(\frac{1}{2}, \frac{41 - \sqrt[4]{100001}}{10}, \frac{\sqrt{401 - \sqrt{100001}}}{10}, \frac{41 + \sqrt[4]{100001}}{10}, \frac{\sqrt{401 - \sqrt{100001}}}{10}\right).$$

For general k_3, k_4, k_5, Pearson's equation (4.2.2) of degree 9 cannot be solved in radicals, as its roots are algebraic numbers with Galois group S_9 over $\mathbb{Q}$. However,

for our special data, the algebraic degree of the solution drops, and we could write the estimate in radicals.

The situation is dramatically different for likelihood inference. It was shown in Chapter 3 that the critical points for the likelihood function of the mixture of two Gaussians with data (4.3.1) have transcendental coordinates, and that the number of these critical points grows with K.

It is thus interesting to assess the quality of our solution (4.3.10) from the likelihood perspective. The probability density function for the Gaussian mixture with these parameters is shown in Figure 4.3.1. The corresponding value of the log-likelihood function is -28.79618895.

If the estimate (4.3.10) is used as starting point in the EM algorithm, then it converges to the stationary point

$$(\alpha, \mu, \sigma, \nu, \tau) = (0.500000, 2.420362, 5.77968, 1.090329, 1.090329).$$

That point has a log-likelihood value of approximately -28.43415. Comparing to Table 3.3.1, this value is only beaten by the critical points associated to the endpoints $k = 1$ and $k = 7$. □

We make the following observation: of all the critical points listed in Table 3.3.1, the middle clusters get the lowest log-likelihood. Hence an equal-means model is not very likely for this sample. This is further confirmed by the method of moments (MOM) since, as mentioned above, the equal-means model is inconsistent with our polynomial equations.

Behavior similar to Example 4.3.3 is observed for all $K \geq 2$. The MOM estimate separates the sample into two halves, and assigns the same variance to both Gaussian components. The exact parameter estimates are obtained by substituting m_1, k_2, k_4 from (4.3.8) into (4.3.10). For $K = 20$, the estimate computed by the EM algorithm with the MOM estimate as starting point beats in likelihood value all K critical points listed in Theorem 3.3.1. For $K > 20$, the likelihood value of the MOM estimate itself appears to be already better than the critical points listed in Theorem 3.3.1. In other words, the MOM produces good starting points for maximum likelihood.

4.4. Varieties of Moments

We have got this far without really getting into the algebraic theory that we will need to study the method of moments. We will do so in this section.

Recall from Section 4.1 that the n-dimensional Gaussian distribution is defined

by the moment generating function

$$\sum_{i_1,i_2,\ldots,i_n\geq 0} \frac{m_{i_1 i_2 \cdots i_n}}{i_1! i_2! \cdots i_n!} t_1^{i_1} t_2^{i_2} \cdots t_n^{i_n} \;=\; \exp\Big(\sum_{r=1}^{n} \mu_r t_r\Big) \cdot \exp\Big(\frac{1}{2} \sum_{i,j=1}^{n} \sigma_{ij} t_i t_j\Big). \quad (4.4.1)$$

The model parameters are the entries of the mean $\mu = (\mu_1, \ldots, \mu_n)$ and of the covariance matrix $\Sigma = (\sigma_{ij})$. The unknowns μ_i have degree 1, and the unknowns σ_{ij} have degree 2. Then $m_{i_1 i_2 \cdots i_n}$ is a homogeneous polynomial of degree $i_1 + i_2 + \cdots + i_n$ in the $n + \binom{n+1}{2}$ unknowns.

Definition 4.4.1. Let $\mathbb{P}^N$ be the projective space of dimension $N = \binom{n+d}{d} - 1$ whose coordinates are all $N+1$ moments $m_{i_1 i_2 \cdots i_n}$ with $i_1 + i_2 + \cdots + i_n \leq d$. The closure of the image of parametrization (4.4.1) is a subvariety $\mathcal{G}_{n,d}$ of $\mathbb{P}^N$. We call this the *Gaussian moment variety of order d*.

Note that the dimension of $\mathcal{G}_{n,d}$ equals $n + \binom{n+1}{2}$. In this section we will discuss this variety and its defining polynomials. Let us examine the Gaussian moment varieties $\mathcal{G}_{n,d}$, starting with the case $n = 1$. The moment variety $\mathcal{G}_{1,d}$ is a surface in $\mathbb{P}^d$. Its defining polynomial equations are as follows:

Proposition 4.4.2. *Let $d \geq 3$. The homogeneous prime ideal of the Gaussian moment surface $\mathcal{G}_{1,d}$ is minimally generated by $\binom{d}{3}$ cubics. These are the 3×3-minors of the $3 \times d$-matrix*

$$H_d \;=\; \begin{pmatrix} 0 & m_0 & 2m_1 & 3m_2 & 4m_3 & \cdots & (d-1)m_{d-2} \\ m_0 & m_1 & m_2 & m_3 & m_4 & \cdots & m_{d-1} \\ m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_d \end{pmatrix}.$$

Proof. Let $I_d = \mathcal{I}(\mathcal{G}_{1,d})$ be the vanishing ideal of the moment surface, and let J_d be the ideal generated by the 3×3-minors of H_d. Now, by Lemma 4.1.3, we know that the moments of the univariate Gaussian distribution satisfy the recurrence relation

$$m_i \;=\; \mu\, m_{i-1} + (i-1)\sigma^2 m_{i-2} \;\text{ for }\; i \geq 1. \quad (4.4.2)$$

Hence the row vector $(\sigma^2, \mu, -1)$ is in the left kernel of H_d. Thus $\operatorname{rank}(H_d) = 2$, and this means that all 3×3-minors of H_d indeed vanish on the surface $\mathcal{G}_{1,d}$. This proves $J_d \subseteq I_d$.

From the previous inclusion we have $\dim(V(J_d)) \geq 2$. Fix a monomial order such that the antidiagonal product is the leading term in each of the 3×3-minors of H_d. These leading terms are the distinct cubic monomials in $m_1, m_2, \ldots, m_{d-2}$. Hence the initial ideal satisfies

$$\langle m_1, m_2, \ldots, m_{d-2} \rangle^3 \;\subseteq\; \operatorname{in}(J_d). \quad (4.4.3)$$

This shows that $\dim(V(J_d)) = \dim(V(\text{in}(J_d))) \leq 2$, and hence $V(J_d)$ has dimension 2 in $\mathbb{P}^d$.

We next argue that $V(J_d)$ is an irreducible surface. On the affine space $\mathbb{A}^d = \{m_0 = 1\}$, this holds, even ideal-theoretically, because the minor indexed by 1, 2 and i expresses m_i as a polynomial in m_1 and m_2. Consider the intersection of $V(J_d)$ with $\mathbb{P}^{d-1} = \{m_0 = 0\}$. The matrix H_d shows that $m_1 = m_2 = \cdots = m_{d-2} = 0$ holds on that hyperplane at infinity, so $V(J_d) \cap \{m_0 = 0\}$ is a curve. Every point on that curve is the limit of points in $V(J_d) \cap \{m_0 = 1\} = V(I_d) \cap \{m_0 = 1\}$, obtained by making (μ, σ) larger in an appropriate direction. This shows that $V(J_d)$ is irreducible, and we conclude that $V(J_d) = V(I_d)$.

At this point we only need to exclude the possibility that J_d has lower dimensional embedded components. However, there are no such components because the ideal of maximal minors of a $3 \times d$-matrix of unknowns is Cohen-Macaulay (see Theorem 18.18 in [Eis13]), and $V(J_d)$ has the expected dimension for an intersection with $\mathbb{P}^d$. This shows that J_d is a Cohen-Macaulay ideal. Hence J_d has no embedded associated primes, and we conclude that $J_d = I_d$ as desired. □

Corollary 4.4.3. *The 3×3-minors of the matrix H_d form a Gröbner basis for the prime ideal of the Gaussian moment surface $\mathcal{G}_{1,d} \subset \mathbb{P}^d$ with respect to the reverse lexicographic term order.*

Proof. The ideal J_d of $\mathcal{G}_{1,d}$ is generated by the 3×3-minors of H_d. Our claim states that equality holds in (4.4.3). This can be seen by examining the Hilbert series of both ideals. It is known that the ideal of $r \times r$-minors of a generic $r \times d$-matrix has the same numerator of the Hilbert series as the r-th power of the monomial ideal $\langle m_1, m_2, \ldots, m_{d-r+1} \rangle$ (see e.g. [BV06, Remark 9.20]). Since that ideal is Cohen-Macaulay, this Hilbert series numerator remains unchanged under transverse linear sections. Hence our ideal J_d has the same Hilbert series numerator as $\langle m_1, m_2, \ldots, m_{d-2} \rangle^3$. This implies that the two ideals in (4.4.3) have the same Hilbert series, so they are equal. □

The argument above tells us that our surface has the same degree as the ideal $\langle m_1, m_2, \ldots, m_{d-2} \rangle^3$:

Corollary 4.4.4. *The Gaussian moment surface $\mathcal{G}_{1,d}$ has degree $\binom{d}{2}$ in $\mathbb{P}^d$.*

We go one step further and determine the singular locus on the Gaussian moment surface.

Lemma 4.4.5. *The singular locus of the surface $\mathcal{G}_{1,d}$ is a single line. It is defined by $\langle m_0, m_1, \ldots, m_{d-2} \rangle$.*

Proof. Let $\mathcal{L}$ be the line defined by $\langle m_0, m_1, \ldots, m_{d-2} \rangle$ and $\mathcal{S} = \text{Sing}(\mathcal{G}_{1,d})$. We claim $\mathcal{L} = \mathcal{S}$.

We first show that $\mathcal{S} \subseteq \mathcal{L}$. Consider the affine open chart $\{m_0 = 1\}$ of $\mathcal{G}_{1,d}$. On that chart, the coordinates m_i are polynomial functions in the unknowns $m_0, \ldots, m_{i-1}$, for $i \geq 3$. Indeed, the 3×3-minor of H_d with column indices $1, 2$ and i has the form $m_i - h(m_0, \ldots, m_{i-1})$. Hence $\mathcal{G}_{1,d} \cap \{m_0 = 1\} \simeq \mathbb{A}^2$, and therefore $\mathcal{S} \subset \{m_0 = 0\}$. Next suppose $m_0 = 0$. The leftmost 3×3-minor of H_d implies $m_1 = 0$. Now, the minor with columns $2, 3, 4$ implies that $m_2 = 0$, the minor with columns $3, 4, 5$ implies that $m_3 = 0$, etc. From the rightmost minor we conclude $m_{d-2} = 0$. This shows that $\mathcal{G}_{1,d} \cap \{m_0 = 0\} = \mathcal{L}$, and we conclude $\mathcal{S} \subseteq \mathcal{L}$.

For the reverse inclusion $\mathcal{L} \subseteq \mathcal{S}$, we consider the Jacobian matrix of the cubics that define $\mathcal{G}_{1,d}$. That matrix has $d+1$ rows and $\binom{d}{3}$ columns. We claim that it has rank $\leq d-3$ on $\mathcal{L}$. To see this, note that the term $m_i m_{d-1}^2$ appears in the minor of H_d with columns $i, d-1, d$ for $i = 2, \ldots, d-2$, and that all other occurrences of m_{d-1} or m_d in any of the 3×3-minors of H_d is linear. Therefore the Jacobian matrix restricted to $\mathcal{L}$ has only $d-3$ non-zero entries, and so its rank is at most $d-3$. This is less than $d-2 = \operatorname{codim}(\mathcal{G}_{1,d})$. We conclude that all points on the line $\mathcal{L}$ are singular points in the Gaussian moment surface $\mathcal{G}_{1,d}$. □

It is natural to ask whether the nice determinantal representation extends to the varieties $\mathcal{G}_{n,d}$ when $n \geq 2$. The answer is yes and no. Let us explain. For $n \geq 2$ it is difficult to compute the prime ideal of the Gaussian moment variety $\mathcal{G}_{n,d}$ in $\mathbb{P}^N$, so one approach is to work on the affine open set $\mathbb{A}^N = \{m_{00\cdots0} = 1\}$. We define the *affine Gaussian moment variety* to be the intersection of $\mathcal{G}_{n,d}$ with the the affine chart $\mathbb{A}^N = \{m_{00\cdots0} = 1\}$ in $\mathbb{P}^N$.

On that affine space, $\mathcal{G}_{n,d}$ is a complete intersection defined by the vanishing of all cumulants $k_{i_1 i_2 \cdots i_n}$ whose order $i_1 + i_2 + \cdots + i_n$ is between 3 and d. Why?

The transformation (4.1.4) between moments and cumulants is an isomorphism. Under this isomorphism, the affine Gaussian moment variety is the linear space defined by the vanishing of all cumulants of orders $3, 4, \ldots, d$. Indeed, recall Example 4.1.11. This implies:

Remark 4.4.6. The affine moment variety $\mathcal{G}_{n,d} \cap \mathbb{A}^N$ is an affine space of dimension $n + \binom{n+1}{2}$.

For instance, the 5-dimensional affine variety $\mathcal{G}_{2,3} \cap \mathbb{A}^9$ is isomorphic to the 5-dimensional linear space defined by $k_{30} = k_{21} = k_{12} = k_{03} = 0$. Each such cumulant is a polynomial in the moments, as explained in section 4.1. The ideal of $\mathcal{G}_{n,d}$ is then obtained from the ideal of cumulants by saturating with respect to $m_{00\cdots0}$. In that way, the affine Gaussian moment variety is still determinantal when $n > 1$ but this does not hold necessarily for $\mathcal{G}_{n,d}$.

Actually, the determinantal representation for $\mathcal{G}_{n,d}$ does not hold even in the first nontrivial case, when $n = 2$ and $d = 3$:

Proposition 4.4.7. *The 5-dimensional variety $\mathcal{G}_{2,3}$ has degree 16 in $\mathbb{P}^9$. Its homogeneous prime ideal is minimally generated by 14 cubics and 4 quartics, and the Hilbert series equals*

$$\frac{1+4t+10t^2+6t^3-4t^4-t^5}{(1-t)^6}.$$

Starting from four of the cubics, the ideal can be computed by a saturation:

$$\langle\, 2m_{10}^3-3m_{00}m_{10}m_{20}+m_{00}^2m_{30}\,,\ 2m_{01}m_{10}^2-2m_{00}m_{10}m_{11}-m_{00}m_{01}m_{20}+m_{00}^2m_{21},\\ 2m_{01}^2m_{10}-m_{00}m_{02}m_{10}-2m_{00}m_{01}m_{11}+m_{00}^2m_{12}, 2m_{01}^3-3m_{00}m_{01}m_{02}+m_{00}^2m_{03}\,\rangle : \langle m_{00}\rangle^\infty$$

The four special cubics above are the cumulants $k_{30}, k_{21}, k_{12}, k_{03}$ when expressed in terms of moments.

Remark 4.4.8. We want to stress the moment-cumulant relation (4.1.4). Either moments or cumulants can serve as an affine coordinate system on the $\mathbb{P}^N$ whose points are inhomogeneous polynomials of degree $\leq d$ in n variables. To be precise, the affine space $\mathbb{A}^N = \{m_{00\cdots 0} = 1\}$ consists of those polynomials whose constant term is nonzero. Hence the formulas (4.1.4) represent a non-linear change of variables on $\mathbb{A}^N$. This was called *Cremona linearization* in [CCM$^+$16]. We agree with the authors of [CCM$^+$16] that passing from m-coordinates to k-coordinates usually simplifies the description of interesting varieties in $\mathbb{P}^N$.

We next exhibit an alternative representation of $\mathcal{G}_{n,d} \cap \mathbb{A}^N$ as a determinantal variety. This is derived from Willink's recursion in [Wil05]. Recall from Remark 4.1.4 that the classical Hermite polynomials are very closely related to the moment polynomials for a univariate Gaussian. What Willink does is to look at the multivariate Hermite polynomials, known to satisfy certain recurrence relations. Translating back to multivariate Gaussian moments, he can obtain analogous recurrences for them. These are [Wil05, eqn (13)]:

$$m_{i_1\cdots i_r+1\cdots i_n} = \mu_r \cdot m_{i_1\cdots i_r\cdots i_n} + \sum_{j=1}^{n} \sigma_{rj} \cdot i_j \cdot m_{i_1\cdots i_r-1\cdots i_n}, \tag{4.4.4}$$

for each index $r = 1, \ldots, n$. When $n = 1$ we recover Lemma 4.1.3.

In this way, we can generalize the matrix B_d in Proposition 4.4.2. We define the *Willink matrix* $W_{n,d}$ as follows. Its rows are indexed by vectors $u \in \mathbb{N}^n$ with $|u| \leq d-1$. The matrix $W_{d,n}$ has $2n+1$ columns. The first entry in the row u is the corresponding moment m_u. The next n entries in the row u are $m_{u+e_1}, m_{u+e_2}, \ldots, m_{u+e_n}$. Finally, the last n entries in the row u are $u_1 m_{u-e_1}, u_2 m_{u-e_2}, \ldots, u_n m_{u-e_n}$. Thus the Willink matrix $W_{n,d}$ has format $\binom{n+d-1}{d-1} \times (2n+1)$ and each entry is a scalar multiple of one of the moments. For $n = 1$, the $d \times 3$-matrix $W_{1,d}$ equals the transpose of the matrix B_d after permuting rows.

Proposition 4.4.9. *The affine Gaussian moment variety $\mathcal{G}_{n,d} \cap \mathbb{A}^N$ is defined by the vanishing of the $(n+2) \times (n+2)$-minors of the Willink matrix $W_{n,d}$.*

Proof. Suppose that the matrix $W_{n,d}$ is filled with the moments of a Gaussian distribution on $\mathbb{R}^n$, and consider the n linearly independent vectors

$$\big(\,\mu_i,\, 0, \ldots, 0, -1, 0, \ldots, 0,\, \sigma_{1i}, \sigma_{2i}, \ldots, \sigma_{ni}\,\big)^T \qquad \text{for } i = 1, 2, \ldots, n. \tag{4.4.5}$$

Here the entry -1 appears in the $(i+1)$st coordinate. By (4.4.4), these n vectors are in the kernel of $W_{n,d}$. Hence the rank of $W_{n,d}$ is $\leq n+1$, and the $(n+2)$-minors are zero.

Conversely, let m be an arbitrary point in $\mathbb{A}^N$ for which the matrix $W_{n,d}$ has rank $\leq n+1$. The square submatrix indexed by the rows $1, 2, \ldots, n+1$ and the columns $1, n+2, \ldots, 2n+1$ has determinant equal to $m_{00\cdots0}^{n+1} = 1$. Hence the rank of $W_{n,d}$ is exactly $n+1$. The kernel of the submatrix given by the first $n+1$ rows is an n-dimensional space for which we can pick a basis of the form (4.4.5). The entries can be interpreted as the mean and the covariance matrix of a Gaussian distribution. The rank hypothesis on $W_{n,d}$ now ensures that the n vectors in (4.4.5) are in the kernel of the full matrix $W_{n,d}$. This means that the higher moments satisfy the recurrences in (4.4.4), and hence the chosen point m lies in $\mathcal{G}_{n,d}$. □

Example 4.4.10. Consider the moments of order at most four for a bivariate Gaussian ($n = 2, d = 4$). The variety $\mathcal{G}_{2,4}$ has dimension 5 in $\mathbb{P}^{14}$. Its Willink matrix has format 10×5:

$$W_{2,4} \;=\; \begin{pmatrix} m_{00} & m_{01} & m_{10} & 0 & 0 \\ m_{01} & m_{02} & m_{11} & 0 & m_{00} \\ m_{10} & m_{11} & m_{20} & m_{00} & 0 \\ m_{02} & m_{03} & m_{12} & 0 & 2m_{01} \\ m_{11} & m_{12} & m_{21} & m_{01} & m_{10} \\ m_{20} & m_{21} & m_{30} & 2m_{10} & 0 \\ m_{03} & m_{04} & m_{13} & 0 & 3m_{02} \\ m_{12} & m_{13} & m_{22} & m_{02} & 2m_{11} \\ m_{21} & m_{22} & m_{31} & 2m_{11} & m_{20} \\ m_{30} & m_{31} & m_{40} & 3m_{20} & 0 \end{pmatrix}$$

The ideal of 4×4-minors of $W_{2,4}$ is minimally generated by 657 quartics. Saturation with respect to the coordinate m_{00} yields the prime ideal of $\mathcal{G}_{2,4}$, and we describe it next.

The affine moment variety $\mathcal{G}_{2,4} \cap \mathbb{A}^{14}$ is defined by the vanishing of the nine

cumulants of order 3:

$$
\begin{aligned}
k_{03} &= 2m_{01}^3 - 3m_{01}m_{02} + m_{03} \\
k_{12} &= 2m_{01}^2 m_{10} - 2m_{01}m_{11} - m_{02}m_{10} + m_{12} \\
k_{21} &= 2m_{01}m_{10}^2 - m_{01}m_{20} - 2m_{10}m_{11} + m_{21} \\
k_{30} &= 2m_{10}^3 - 3m_{10}m_{20} + m_{30} \\
k_{04} &= -6m_{01}^4 + 12m_{01}^2 m_{02} - 4m_{01}m_{03} - 3m_{02}^2 + m_{04} \\
k_{13} &= -6m_{01}^3 m_{10} + 6m_{01}^2 m_{11} + 6m_{01}m_{02}m_{10} - 3m_{01}m_{12} - 3m_{02}m_{11}m_{03}m_{10} + m_{13} \\
k_{22} &= -6m_{01}^2 m_{10}^2 + 2m_{01}^2 m_{20} + 8m_{01}m_{10}m_{11} + 2m_{02}m_{10}^2 - 2m_{01}m_{21} - m_{02}m_{20} - 2m_{10}m_{12} - 2m_{11}^2 + m_{22} \\
k_{31} &= -6m_{01}m_{10}^3 + 6m_{01}m_{10}m_{20} + 6m_{10}^2 m_{11} - m_{01}m_{30} - 3m_{10}m_{21} - 3m_{11}m_{20} + m_{31} \\
k_{40} &= -6m_{10}^4 + 12m_{10}^2 m_{20} - 4m_{10}m_{30} - 3m_{20}^2 + m_{40}
\end{aligned}
$$

The ideal of the projective variety $\mathcal{G}_{2,4}$ is obtained from these nine polynomials by homogenizing and saturating with a new unknown m_{00}. The result of this computation is as follows.

Proposition 4.4.11. *The* 5*-dimensional variety* $\mathcal{G}_{2,4}$ *has degree* 102 *in* $\mathbb{P}^{14}$*. Its prime ideal is minimally generated by* 99 *cubics,* 41 *quartics, and one quintic. The Hilbert series equals*

$$\frac{1 + 9t + 45t^2 + 66t^3 - 27t^4 + 13t^5 - 8t^6 + 4t^7 - t^8}{(1-t)^6}.$$

Remark 4.4.12. The Gaussian moment variety $\mathcal{G}_{2,5}$ has dimension 5 in $\mathbb{P}^{19}$, and we found its degree to be 332. However, at present, we do not know a generating set for its prime ideal.

We close this section by reporting the computation of the first interesting case for $n = 3$.

Proposition 4.4.13. *The Gaussian moment variety* $\mathcal{G}_{3,3}$ *has dimension* 9 *and degree* 130 *in* $\mathbb{P}^{19}$*. Its prime ideal is minimally generated by* 84 *cubics,* 192 *quartics,* 21 *quintics,* 15 *sextics,* 36 *septics, and* 35 *octics. The Hilbert series equals*

$$\frac{1+10t+55t^2+136t^3-26t^4-150t^5+139t^6-127t^7+310t^8-449t^9+360t^{10}-160t^{11}+32t^{12}-t^{13}}{(1-t)^{10}}$$

Remark 4.4.14. As seen from negative coefficients in the Hilbert series, the corresponding ideals when $n > 1$ are in general no longer Cohen-Macaulay.

In this chapter we have seen the algebraic advantage of using the method of moments for parameter inference of Gaussian mixtures. With our moment varieties $\mathcal{G}_{n,d}$ well defined by moments of a single Gaussian, we are ready to move in the next chapter to moments of mixtures.

5. Secants and Algebraic Identifiability

In the previous chapter we have introduced the Gaussian moment variety $\mathcal{G}_{n,d}$ as a subvariety of $\mathbb{P}^N$, where $N = \binom{n+d}{d} - 1$. Its points are the vectors of all moments of order $\leq d$ of an n-dimensional Gaussian distribution, parametrized birationally by the entries of the mean vector $\mu = (\mu_1, \ldots, \mu_n)$ and the covariance matrix $\Sigma = (\sigma_{ij})$. The variety $\mathcal{G}_{n,d}$ is rational of dimension $n(n+3)/2$ for $d \geq 2$.

However, we want to deal with moments of *mixtures* of Gaussians, not only Gaussians. Thankfully (as we have seen already with Pearson's method in $n = 1, k = 2$), the moments of a mixture are easy to compute from the moments of the individual components. Since the expectation is linear, if we have a mixture $f = \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_k f_k$, then the moment m_i of f is just the convex combination of the ith moments of $f_1, \ldots, f_k$, with weights $\alpha_1, \ldots, \alpha_k$ respectively.

The algebraic concept that we need is the *secant variety* [Ådl87]:

Definition 5.0.1. Let X be algebraic variety and $k \geq 1$, the *kth secant variety* of X, $Sec_k(X)$ is the Zariski closure of the union of all linear spaces spanned by collections of k points on X.

- If $k = 1$, we recover the original variety X.
- If $k = 2$, we have the union of the secant lines to the variety X.

With the general definition in mind, we now focus on our Gaussian moment variety case:

Definition 5.0.2. The kth *secant moment variety* $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ is the Zariski closure in $\mathbb{P}^N$ of the set of vectors of moments of order $\leq d$ of any probability distribution on $\mathbb{R}^n$ that is the mixture of k Gaussians, for $k \geq 2$.

In short, $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ is the projective variety that represents mixtures of k Gaussians. These varieties are the main object of study in this chapter.

The parametrization of $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ is given by replacing the right hand side of (4.4.1) with a convex combination of k such expressions. The number of model parameters is $k \cdot \left[n + \binom{n+1}{2} \right] + k - 1$.

Example 5.0.3. Consider the familiar case $n = 1$ and $d = 6$. We know from Proposition 4.4.2 that the Gaussian moment variety $\mathcal{G}_{1,6}$ is a surface of degree 15 in $\mathbb{P}^6$ that is cut out by 20 cubics. For $k = 2$ we obtain the variety of secant lines, here denoted $\mathrm{Sec}_2(\mathcal{G}_{1,6})$. This represents mixtures of two univariate Gaussians. It has the parametric representation from (4.2.1) (with the extra equation for m_6):

$$\begin{aligned}
m_0 &= 1 \\
m_1 &= \alpha\mu + (1-\alpha)\nu \\
m_2 &= \alpha(\mu^2+\sigma^2) + (1-\alpha)(\nu^2+\tau^2) \\
m_3 &= \alpha(\mu^3+3\mu\sigma^2) + (1-\alpha)(\nu^3+3\nu\tau^2) \\
m_4 &= \alpha(\mu^4+6\mu^2\sigma^2+3\sigma^4) + (1-\alpha)(\nu^4+6\nu^2\tau^2+3\tau^4) \\
m_5 &= \alpha(\mu^5+10\mu^3\sigma^2+15\mu\sigma^4) + (1-\alpha)(\nu^5+10\nu^3\tau^2+15\nu\tau^4) \\
m_6 &= \alpha(\mu^6+15\mu^4\sigma^2+45\mu^2\sigma^4+15\sigma^6) + (1-\alpha)(\nu^6+15\nu^4\tau^2+45\nu^2\tau^4+15\tau^6)
\end{aligned} \tag{5.0.1}$$

The variety $\mathrm{Sec}_2(\mathcal{G}_{1,6})$ is five-dimensional, so it is a hypersurface in $\mathbb{P}^6$. This is when the computation we did for Theorem 4.2.3 now becomes very relevant, and takes a new meaning as promised then.

Theorem 5.0.4. *The defining polynomial of* $\mathrm{Sec}_2(\mathcal{G}_{1,6})$ *is a sum of* 31154 *monomials of degree* 39. *This polynomial has degrees* $25, 33, 32, 23, 17, 12, 9$ *in the unknowns* $m_0, m_1, m_2, m_3, m_4, m_5, m_6$ *respectively.*

Proof. In the proof of Theorem 4.2.3, after checking that the polynomial vanishes at the parametrization (5.0.1), to pass from affine space $\mathbb{A}^6$ to projective space $\mathbb{P}^6$, we introduce the homogenizing variable m_0. This is done by replacing m_i with m_i/m_0 for $i = 1, 2, 3, 4, 5, 6$ and clearing denominators. The degrees for the unknowns remain unchanged, and the new degree for m_0 is also read off by inspection. □

We see in particular that m_6 can be recovered from m_1, m_2, m_3, m_4 and m_5 by solving a univariate equation of degree 9. This number rings a bell for us. Indeed, we should remember Pearson's result in [Pea94] that one can find the five parameters in (4.2.1) by solving an equation of degree 9 if the first five moments are given (as we carefully verified in Section 4.2). The two occurrences of the number 9 are equivalent, in light of Lazard's result [Laz04] that the parameters $\alpha, \mu, \nu, \sigma, \tau$ are rational functions in the first six moments $m_1, \ldots, m_6$.

Remark 5.0.5. The elimination in the proof above can be carried out by computing a Gröbner basis for the ideal that is obtained by adding (4.2.13) to the system (4.2.12). Such a Gröbner basis reveals that both p and s can be expressed as rational functions in the cumulants. This confirms Lazard's result [Laz04] that Gaussian mixtures for $k = 2$ and $n = 1$ are rationally identifiable from their moments up to order six. We stress that Lazard [Laz04] does much more than proving rational identifiability: he also provides a very detailed analysis of the real

structure and special fibers of the map $(\alpha, \mu, \nu, \sigma, \tau) \mapsto (m_1, m_2, m_3, m_4, m_5, m_6)$ in (5.0.1).

In order to understand $\mathrm{Sec}_k(\mathcal{G}_{n,d})$, we would really like to know its dimension, its degree and its defining equations. We did this description for $n = 1, k = 2, d = 6$ in the example and theorem above. By looking at the complexity of that case, we can already expect that this may be a very difficult problem.

Since the first fundamental invariant is the dimension, our aim now is to determine the dimension of the secant variety $\mathrm{Sec}_k(\mathcal{G}_{n,d})$. That dimension is always bounded above by the number of parameters, so we have

$$\dim\big(\mathrm{Sec}_k(\mathcal{G}_{n,d})\big) \;\leq\; \min\{\, N\,,\;\; kn(n+3)/2 \,+\, k-1\,\}\,. \tag{5.0.2}$$

The right hand side is the *expected dimension*.

Since Gaussian mixtures are identifiable [YS68], this secant variety eventually has the expected dimension:

$$\dim(\mathrm{Sec}_k(\mathcal{G}_{n,d})) \;=\; k \cdot \left[n + \binom{n+1}{2}\right] + k - 1 \qquad \text{for } d \gg 0. \tag{5.0.3}$$

If equality holds in (5.0.2), then $\mathrm{Sec}_k\big(\mathcal{G}_{n,d}\big)$ is *nondefective*. If this holds, and $N \geq \frac{1}{2}kn(n+3) + k - 1$, then the Gaussian mixtures are algebraically identifiable from their N moments of order $\leq d$. Here *algebraically identifiable* means that the map from the model parameters to the moments is generically finite-to-one. This means parameters can be recovered by solving a zero-dimensional system of polynomial equations. The term *rationally identifiable* is used if the map is generically one-to-one (identifying label-swapped parameters), so that the Gaussian mixture density is unique. In this section we only study algebraic identifiability. For rational identifiability see Remark 6.1.4.

Thus, the fundamental problem of determining the dimension of $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ translates into knowing if we can identify our model parameters from given moments up to order d. In other words, we want to know whether the method of moments can even succeed for given k, n, d.

The main result of this chapter contrasts the cases $n = 1$ and $n \geq 3$.

Theorem 5.0.6. *Equality holds in (5.0.2) for $n = 1$ and all values of d and k. Hence all moment varieties of mixtures of univariate Gaussians are algebraically identifiable. The same is false for $n \geq 3$, $d = 3$ and $k = 2$: here the right hand side of (5.0.2) exceeds the left hand side by two.*

Remark 5.0.7. The points on $\mathcal{G}_{n,d}$ where the covariance matrix is zero correspond to the classical *Veronese* varieties. Defective Veronese varieties are classified by the celebrated *Alexander-Hirschowitz Theorem* [AH00, BO08]. It may be relevant for our study since the Gaussian moment variety is a 'noisy' version of the former,

and Veronese varieties are naturally contained in corresponding Gaussian moment varieties. We will come back to these considerations in Chapter 6.

Our result for $d = 3$ is a Gaussian analogue of the infinite family ($d = 2$) in the Alexander-Hirschowitz classification (Theorem 6.2.7) of defective Veronese varieties. Many further defective cases for $d = 4$ are exhibited in Table 5.2.2 and Conjecture 5.2.10. Extensive computer experiments (up to $d = 24$) suggest that moment varieties are never defective for bivariate Gaussians ($n = 2$).

Conjecture 5.0.8. *Equality holds in (5.0.2) for $n = 2$ and all values of d and k. In particular, all moment varieties of mixtures of bivariate Gaussians are algebraically identifiable.*

The rest of the chapter is organized as follows. In the next section we focus on the case $n = 1$. We review what is known classically on defectivity of surfaces. Based on this, we then prove the first part of Theorem 5.0.6. In section 5.2 we study our problem for $n \geq 2$. We begin with the parametric representation of $\mathrm{Sec}_k(\mathcal{G}_{n,d})$, we next establish the second part of Theorem 5.0.6, and thereafter we study the defect and we examine higher moments. The last section discusses what little we know about the equations and degrees of the varieties $\mathrm{Sec}_k(\mathcal{G}_{n,d})$. Both Sections 5.2 and 5.3 feature many open problems.

5.1. One-dimensional Gaussians

The moments $m_0, m_1, m_2, \ldots, m_d$ of a Gaussian distribution on the real line are polynomial expressions in the mean μ and the variance σ^2. These expressions will be reviewed in Remark 5.1.1. They give a parametric representation of the Gaussian moment surface $\mathcal{G}_{1,d}$ in $\mathbb{P}^d$.

The $3 \times d$-matrix G_d has entries that are linear forms in $d+1$ unknowns $m_0, \ldots, m_d$. That matrix may be interpreted as a 3-dimensional tensor of format $3 \times d \times (d+1)$. That tensor can be turned into a $d \times (d+1)$ matrix whose entries are linear forms in three unknowns x, y, z. The result is what we call the *Hilbert-Burch matrix* of our surface $\mathcal{G}_{1,d}$. It equals

$$B_d \;=\; \begin{pmatrix} y & z & 0 & 0 & \cdots & 0 \\ x & y & z & 0 & \cdots & 0 \\ 0 & 2x & y & z & \cdots & 0 \\ 0 & 0 & 3x & y & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (d-1)x & y & z \end{pmatrix}. \tag{5.1.1}$$

Its maximal minors generate a Cohen-Macaulay ideal, defining a scheme Z_d of length $\binom{d+1}{2}$ supported at the point $(1:0:0)$. Consider the map defined by the

maximal minors of B_d,

$$\phi : \mathbb{P}^2 \dashrightarrow \mathbb{P}^d.$$

The base locus of the map ϕ is the scheme Z_d and its image is the surface $\mathcal{G}_{1,d}$.

Remark 5.1.1. The parametrization ϕ onto $\mathcal{G}_{1,d}$ is birational. It equals the familiar affine parametrization, as in (4.4.1), of the Gaussian moments in terms of mean and variance if we set

$$x = -\sigma^2\,, \;\; y = \mu \;\text{ and }\; z = 1. \tag{5.1.2}$$

The image of the line $\{x = -\sigma^2 z\}$, for fixed value of the variance σ^2, is a rational normal curve of degree d inside the Gaussian moment surface $\mathcal{G}_{1,d}$. It is defined by the 2×2-minors of a 2-dimensional space of rows in the matrix G_d. The singular line $\mathcal{L} \subset \mathcal{G}_{1,d}$ is the tangent line to this curve at the point $(0 : \cdots : 0 : 1)$. In particular, the image of the line $\{x = 0\}$ is the rational normal curve defined by the 2×2-minors of the last two rows of G_d.

We now come to our main question, namely whether there exist d and k such that $\mathcal{G}_{1,d}$ is k-defective in $\mathbb{P}^d$. Theorem 5.0.6 asserts that this is not the case. Equivalently, the dimension of $\mathrm{Sec}_k(\mathcal{G}_{1,d})$ is always equal to the minimum of d and $3k-1$, which is the upper bound in (5.0.2).

Curves can never be defective, but surfaces can. The prototypical example is the Veronese surface S in the space $\mathbb{P}^5$ of symmetric 3×3-matrices. Points on S are matrices of rank 1. The secant variety $\mathrm{Sec}_2(S)$ consists of matrices of rank ≤ 2. Its expected dimension is five whereas the true dimension of S is only four. This means that S is k-defective for $k = 2$.

The following well-known result on higher secant varieties of a variety X allows us to show that X is not k-defective for any k by proving this for one particular k (see [Ådl88]):

Proposition 5.1.2. *Let X be a k'-defective subvariety of $\mathbb{P}^d$ and $k > k'$. Then X is k-defective as long as $\mathrm{Sec}_k(X)$ is a proper subvariety of $\mathbb{P}^d$. In fact, the defectivity increases with k:*

$$(\dim(X)+1)\cdot k - 1 - \dim(\mathrm{Sec}_k(X)) \;>\; (\dim(X)+1)\cdot k' - 1 - \dim(\mathrm{Sec}_{k'}(X)). \tag{5.1.3}$$

Proof. By Terracini's Lemma, the dimension of the secant variety $\mathrm{Sec}_k(X)$ is the dimension of the span of the tangent spaces to X at k general points. Since X is k'-defective and $k' < k$, the linear span of $k - k'$ general tangent spaces to the affine cone over X must intersect the span of k' such general tangent spaces in a positive-dimensional linear space. The dimension of that intersection is the difference of the left hand side minus the right hand side in (5.1.3). □

Corollary 5.1.3. *If a surface $X \subset \mathbb{P}^d$ is defective, then X is k-defective for some $k \geq (d-2)/3$.*

Proof. We proceed by induction on k. If the surface X is $(k-1)$-defective and $k < (d-2)/3$, then $\dim(\mathrm{Sec}_k(X)) < 3k+2 < d$. So X is also k-defective, by Proposition 5.1.2. □

Our main geometric tool is Terracini's 1921 classification of all k-defective surfaces:

Theorem 5.1.4. *(Classification of k-defective surfaces) Let $X \subset \mathbb{P}^N$ be a reduced, irreducible, non-degenerate projective surface that is k-defective. Then $k \geq 2$ and either*

(1) X is the quadratic Veronese embedding of a rational normal surface Y in $\mathbb{P}^k$; or

(2) X is contained in a cone over a curve, with apex a linear space of dimension $\leq k-2$.

Furthermore, for general points $x_1, \ldots, x_k$ on X there is a hyperplane section tangent along a curve C that passes through these points. In case (1), the curve C is irreducible; in case (2), the curve C decomposes into k algebraically equivalent curves $C_1, \ldots, C_k$ with $x_i \in C_i$.

Proof. See [CC02, Theorem 1.3 (i),(ii)] and cases (i) and (ii) of the proof given there. □

Chiantini and Ciliberto offer a nice historical account of this theorem in the introduction to their article [CC02]. A modern proof follows from the more general result in [CC02, Theorem 1.1].

Corollary 5.1.5. *If the surface $X = \mathcal{G}_{1,d}$ is k-defective, then statement (2) in Theorem 5.1.4 holds.*

Proof. We need to rule out case (1) in Theorem 5.1.4. A rational normal surface is either a Hirzebruch surface or it is the cone over a rational curve. The former is smooth and the latter is singular at only one point. The same is true for the quadratic Veronese embedding of such a surface. By contrast, our surface $\mathcal{G}_{1,d}$ is singular along a line, by Lemma 4.4.5. Alternatively, a quadratic Veronese embedding of a surface contains no line. □

Our goal is now to rule out case (2) in Theorem 5.1.4. That proof will be much more involved. Our strategy is to set up a system of surfaces and morphisms between them, like this:

$$\begin{array}{ccccc} S_d & \to & \bar{S}_d & \subset & \mathbb{P}^{N_d} \\ \downarrow & & \downarrow & & \\ \mathbb{P}^2 & & \mathcal{G}_{1,d} & \subset & \mathbb{P}^d \end{array} \tag{5.1.4}$$

The second row in (5.1.4) represents the rational map $\phi : \mathbb{P}^2 \dashrightarrow \mathcal{G}_{1,d}$ that is given by the maximal minors of B_d. Above $\mathbb{P}^2$ sits a smooth surface S_d which we shall construct by a sequence of blow-ups from $\mathbb{P}^2$. It will have the property that ϕ lifts to a morphism on S_d. Curves of degree d in $\mathbb{P}^2$ specify a divisor class H_d on S_d. The complete linear system $|H_d|$ maps S_d onto a rational surface $\bar{S}_d$ in $\mathbb{P}^{N_d}$ where $N_d = \dim(|H_d|)$. The subsystem of $|H_d|$ given by the $d+1$ maximal minors of B_d, then defines the vertical map from $\bar{S}_d$ onto $\mathcal{G}_{1,d}$. Our plan is to use the intersection theory on S_d to rule out the possibility (2) in Theorem 5.1.4.

Lemma 5.1.6. *Suppose that we have a diagram as in (5.1.4) and $X = \mathcal{G}_{1,d}$ satisfies statement (2) in Theorem 5.1.4. Then, for any k general points $x_1, \ldots, x_k$ on the surface S_d, there exist linearly equivalent divisors $D_1 \ni x_1, \ldots, D_k \ni x_k$ and there exists a hyperplane section of $\mathcal{G}_{1,d}$ in $\mathbb{P}^d$, with pullback H_d to S_d, such that $H_d - 2D_1 - 2D_2 - \cdots - 2D_k$ is effective on S_d.*

Proof. By part (2) of Theorem 5.1.4, there exist algebraically equivalent curves $C_1, \ldots, C_k$ on X that contain the images of the respective points $x_1, \ldots, x_k$, and there is a hyperplane section H_X of X which contains and is singular along each C_i. Let $H \subset S_d$ be the preimage of H_X, and let $D_i \subset S_d$ be the preimage of C_i. Then $x_i \in D_i$ for $i = 1, \ldots, k$. Furthermore, the divisor H has multiplicity at least 2 along each D_i. Finally, since S_d is a rational surface, linear and algebraic equivalence of divisors coincide, and the lemma follows. □

We now construct the smooth surface S_d. Let V_d denote the $(d+1)$-dimensional vector space spanned by the maximal minors of the matrix B_d in (5.1.1). When d is odd these minors are

$$
\begin{aligned}
b_{d,0} &= z^d, \\
b_{d,1} &= yz^{d-1}, \\
b_{d,2} &= y^2z^{d-2} - xz^{d-1}, \\
b_{d,3} &= y^3z^{d-3} - 3xyz^{d-2}, \\
\cdots \quad &\cdots \qquad \cdots \quad \cdots \quad \cdots \quad \cdots \\
b_{d,d-1} &= y^{d-1}z - \tbinom{d-1}{2} xy^{d-3}z^2 + \ldots + a_{(\frac{d-3}{2},d-1)} x^{\frac{d-3}{2}} y^2 z^{\frac{d-1}{2}} + a_{(\frac{d-1}{2},d-1)} x^{\frac{d-1}{2}} z^{\frac{d+1}{2}}, \\
b_{d,d} &= y^d - \tbinom{d}{2} xy^{d-2}z + a_{(2,d)} x^2 y^{d-4} z^2 + \ldots + a_{(\frac{d-1}{2},d)} x^{\frac{d-1}{2}} y z^{\frac{d-1}{2}}.
\end{aligned}
$$

When d is even, the maximal minors of the Hilbert-Burch matrix B_d are

$$\begin{aligned}
b_{d,0} &= z^d, \\
b_{d,1} &= yz^{d-1}, \\
b_{d,2} &= y^2z^{d-2} - xz^{d-1}, \\
b_{d,3} &= y^3z^{d-3} - 3xyz^{d-2}, \\
\cdots \quad &\cdots \qquad \cdots \quad \cdots \quad \cdots \quad \cdots \\
b_{d,d-1} &= y^{d-1}z - \tbinom{d-1}{2}xy^{d-3}z^2 + \ldots + a_{(\frac{d-4}{2},d-1)}x^{\frac{d-4}{2}}y^3z^{\frac{d-2}{2}} + a_{(\frac{d-2}{2},d-1)}x^{\frac{d-2}{2}}yz^{\frac{d}{2}}, \\
b_{d,d} &= y^d - \tbinom{d}{2}xy^{d-2}z + a_{(2,d)}x^2y^{d-4}z^2 + \ldots + a_{(\frac{d}{2},d)}x^{\frac{d}{2}}z^{\frac{d}{2}}.
\end{aligned}$$

Here the $a_{(i,j)}$ are rational constants. The point $p = (1:0:0)$ is the only common zero of the forms $b_{d,0}, \ldots, b_{d,d}$. All forms are singular at p, with the following lowest degree terms:

$$z^d,\ yz^{d-1},\ z^{d-1},\ yz^{d-2},\ \ldots,\ z^{(d+1)/2},\ yz^{(d-1)/2} \qquad \text{when } d \text{ is odd;} \tag{5.1.5}$$

$$z^d,\ yz^{d-1},\ z^{d-1},\ yz^{d-2},\ \ldots,\ yz^{d/2},\ z^{d/2} \qquad \text{when } d \text{ is even.} \tag{5.1.6}$$

Consider a general form in V_d. Then its lowest degree term at p is a linear combination of $z^{(d+1)/2}$ and $yz^{(d-1)/2}$ when d is odd, and it is a scalar multiple of $z^{d/2}$ when d is even.

The forms $b_{d,0}, \ldots, b_{d,d}$ define a morphism $\phi : \mathbb{P}^2\backslash\{p\} \to \mathbb{P}^d$ that does not extend to p. Consider any map $\pi : S' \to \mathbb{P}^2$ that is obtained by a sequence of blow-ups at smooth points, starting with the blow-up of $\mathbb{P}^2$ at p. Let $E \subset S'$ be the preimage of p. The restriction of π to $S'\backslash E$ is an isomorphism onto $\mathbb{P}^2\backslash\{p\}$, and so ϕ naturally defines a morphism $S'\backslash E \to \mathbb{P}^d$.

We now define our surface S_d in (5.1.4). It is a minimal surface S' such that $S'\backslash E \to \mathbb{P}^d$ extends to a morphism $\tilde{\phi} : S' \to \mathbb{P}^d$. Here "minimal" refers to the number of blow-ups, and we do not claim S_d is the unique such minimal surface.

Let H_d be the strict transform on S_d of a curve in $\mathbb{P}^2$ defined by a general form in V_d. The complete linear system $|H_d|$ on S_d defines a morphism $S_d \to \mathbb{P}^{N_d}$, where $N_d = \dim|H_d|$. Let $\bar{S}_d \subset \mathbb{P}^N$ be the image. Then $\tilde{\phi} : S_d \to \mathbb{P}^d$ is the composition of $S_d \to \mathbb{P}^N$ and a linear projection to $\mathbb{P}^d$ whose restriction to $\bar{S}_d$ is finite. Thus we now have the diagram in (5.1.4).

Relevant for proving Theorem 5.0.6 are the first two among the blow-ups that lead to S_d. The map ϕ is not defined at p. More precisely, ϕ is undefined at p and at its tangent direction $\{z = 0\}$. Let $S_p \to \mathbb{P}^2$ be the blow-up at p, with exceptional divisor E_p. Let $S_{p,z} \to S_p$ be the blow-up at the point on E_p corresponding to the tangent direction $\{z = 0\}$ at p, with exceptional divisor E_z. To obtain S_d we need to blow up $S_{p,z}$ in s further points for some s.

Now, S_d is a smooth rational surface. Let L be the class of a line pulled back

to S_d, and let $E_p, E_z, F_1, \ldots, F_s$, be the classes of the exceptional divisors of each blow-up, pulled back to S_d. The divisor class group of S_d is the free abelian group with basis $L, E_p, E_z, F_1, \ldots, F_s$. The intersection pairing on this group is diagonal for this basis, with

$$L^2 \;=\; -E_p^2 \;=\; -E_z^2 \;=\; -F_1^2 \;=\; \cdots \;=\; -F_s^2 \;=\; 1. \tag{5.1.7}$$

The intersection of two curves on the smooth surface S_d, having no common components, is a nonnegative integer. It is computed as the intersection pairing of their classes using (5.1.7).

Lemma 5.1.7. *Consider the linear system $|H_d|$ on S_d that represents hyperplane sections of $\mathcal{G}_{1,d} \subset \mathbb{P}^d$, pulled back via the morphism $\tilde{\phi}$. Its class in the Picard group of S_d is given by*

$$\begin{aligned} H_d \;&=\; dL - \tfrac{d}{2}E_p - \tfrac{d}{2}E_z \;- c_1F_1 - c_2F_2 - \cdots - c_sF_s \qquad \textit{when } d \textit{ is even,} \\ H_d \;&=\; dL - \tfrac{d+1}{2}E_p - \tfrac{d-1}{2}E_z \;- c_1F_1 - c_2F_2 - \cdots - c_sF_s \quad \textit{when } d \textit{ is odd.} \end{aligned}$$

Here $c_1, c_2, \ldots, c_s$ are positive integers whose precise value will not matter to us.

Proof. The forms in V_d define the preimages in $\mathbb{P}^2$ of curves in $|H_d|$. The first three coefficients are seen from the analysis in (5.1.5) and (5.1.6). The general hyperplane in $\mathbb{P}^d$ intersects the image of the exceptional curve F_i in finitely many points. Their number is the coefficient c_i. □

Proof of the first part of Theorem 5.0.6. Suppose that $X = \mathcal{G}_{1,d}$ is k-defective for some k. By Corollary 5.1.3, we may assume that $3k + 2 \geq d$. By Corollary 5.1.5 and Lemma 5.1.6, the class of the linear system $|H_d|$ in the Picard group of the smooth surface S_d can be written as

$$H_d \;=\; A \,+\, 2kD,$$

where A is effective and D is the class of a curve on S_d that has no fixed component. According to Lemma 5.1.7, we can write

$$D \;=\; aL - b_pE_p - b_zE_z - \sum_{i=1}^{s} c_i'F_i,$$

where $a = D \cdot L$ is a positive integer and $b_p, b_z, c_1', \ldots, c_s'$ are nonnegative integers.

Assume first that $a \geq 2$. We have the following chain of inequalities:

$$0 \;\leq\; L \cdot A \;=\; L \cdot H_d - 2k(L \cdot D) \;=\; d - 2ka \;\leq\; d - 4k \;\leq\; 2 - k.$$

This implies $k \leq 2$. The case $k = 1$ being vacuous, we conclude that $k = 2$ and

hence $d \leq 8$. If $d \leq 5$, then $\mathrm{Sec}_2(\mathcal{G}_{1,d}) = \mathbb{P}^d$ is easily checked, by computing the rank of the Jacobian of the parametrization. For $d = 6$, we know from Theorem 5.0.4 that $\mathrm{Sec}_2(\mathcal{G}_{1,6})$ is a hypersurface of degree 39 in $\mathbb{P}^6$. If $d \in \{7, 8\}$, then the secant variety $\mathrm{Sec}_2(\mathcal{G}_{1,d})$ is also 5-dimensional, by the computation with cumulants in Proposition 5.3.1.

Next, suppose $a = D \cdot L = 1$. The divisor D is the strict transform on S_d of a line in $\mathbb{P}^2$. The multiplicity of this line at p is at most 1, i.e. $0 \leq D \cdot E_p \leq 1$. Furthermore, $D \cdot E_z = 0$ because D moves. Suppose that $D \cdot E_p = 0$ and d is even. Then we have $d \geq 4k$ because

$$d/2 \,=\, H_d \cdot E_p \,=\, A \cdot E_p \,\leq\, A \cdot L \,\leq\, d - 2k.$$

Since $d \leq 3k + 2$, this implies $k = 2$ and $d = 8$. This case has already been ruled out above. If $D \cdot E_p = 0$ and d is odd, then the same reasoning yields $(d+1)/2 = A \cdot E_p \leq d - 2k$. This implies $3k + 2 \geq d \geq 4k + 1$, which is impossible for $k \geq 2$.

It remains to examine the case $D \cdot E_p = 1$. Here, any curve linearly equivalent to D on S_d is the strict transform of a line in $\mathbb{P}^2$ passing through $p = (1 : 0 : 0)$. Through a general point in the plane there is a unique such line, so it suffices to show that the doubling of any line through p is not a component of any curve defined by a linear combination of the $b_{d,i}$. In particular, it suffices to show that y^2 is not a factor of any form in the vector space V_d.

To see this, we note that no monomial $x^r y^s z^t$ appears in more than one of the forms $b_{d,0}, b_{d,1}, \ldots, b_{d,d}$. Hence, in order for y^2 to divide a linear combination of $b_{d,0}, b_{d,1}, \ldots, b_{d,d}$, it must already divide one of the $b_{d,i}$. However, from the explicit expansions we see that y^2 is not a factor of $b_{d,i}$ for any i. This completes the proof of the first part in Theorem 5.0.6. □

5.2. Higher-dimensional Gaussians

We begin by recalling the general definition of the moment variety for Gaussian mixtures. The coordinates on $\mathbb{P}^N$ are the moments $m_{i_1 i_2 \cdots i_n}$. The variety $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ has the parametrization

$$\sum_{i_1,i_2,\ldots,i_n \geq 0} \frac{m_{i_1 i_2 \cdots i_n}}{i_1! i_2! \cdots i_n!} t_1^{i_1} t_2^{i_2} \cdots t_n^{i_n} = \sum_{\ell=1}^{k} \alpha_\ell \cdot \exp\Big(\sum_{r=1}^{n} t_r \mu_{\ell r} + \tfrac{1}{2} \sum_{i,j=1}^{n} \sigma_{\ell i j} t_i t_j \Big) \quad (5.2.1)$$

This is a formal identity of generating functions in n unknowns $t_1, \ldots, t_n$. The model parameters are the kn coordinates $\mu_{\ell i}$ of the mean vectors, the $k\binom{n+1}{2}$ entries $\sigma_{\ell i j}$ of the covariance matrices, and the k mixture parameters α_ℓ. The latter satisfy $\alpha_1 + \cdots + \alpha_k = 1$. This is a map from the space of model parameters into the affine space $\mathbb{A}^N$ that sits inside $\mathbb{P}^N$ as $\{m_{00\cdots 0} = 1\}$. We define $\mathrm{Sec}_k(\mathcal{G}_{n,d}) \subset \mathbb{P}^N$ as

the projective closure of the image of this map.

Remark 5.2.1. The affine Gaussian moment variety $\mathcal{G}_{n,d} \cap \mathbb{A}^N$ is isomorphic to an affine space (cf. 4.4.6). In particular it is smooth. Hence the singularities of $\mathcal{G}_{n,d}$ are all contained in the hyperplane at infinity. This means that the definition of $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ is equivalent to the usual definition of higher secant varieties: it is the closure of the union of all $(k-1)$-dimensional linear spaces that intersect $\mathcal{G}_{n,d}$ in k distinct smooth points.

In this section we focus on the case $d = 3$, that is, we examine the varieties defined by first, second and third moments of Gaussian distributions. The following is our main result.

Theorem 5.2.2. *The moment variety $\mathcal{G}_{n,3}$ is k-defective for $k \geq 2$. In particular, for $k = 2$, the model has two more parameters than the dimension of the secant variety, i.e. $n(n+3)+1 - \dim\big(\mathrm{Sec}_2(\mathcal{G}_{n,3})\big) = 2$. If $n \geq 3$ and we fix distinct first coordinates μ_{11} and μ_{21} for the two mean vectors, then the remaining parameters are identified uniquely. In each of these statements, the parameter k is assumed to be in the range where $\mathrm{Sec}_k(\mathcal{G}_{n,3})$ does not fill $\mathbb{P}^N$.*

This proves the second part of Theorem 5.0.6. We begin by studying the first interesting case.

Example 5.2.3. Let $n = d = 3$ and $k = 2$. In words, we consider moments up to order three for the mixture of two Gaussians in $\mathbb{R}^3$. This case is special because the number of parameters coincides with the dimension of the ambient space: $N = \frac{1}{2}kn(n+3) + k - 1 = 19$. The variety $\mathrm{Sec}_2(\mathcal{G}_{3,3})$ is the closure of the image of the map $\mathbb{A}^{19} \to \mathbb{P}^{19}$ that is given by the expansion of (5.2.1).

A direct computation shows that the 19×19-Jacobian matrix of this map has rank 17 for generic parameter values. Hence the dimension of $\mathrm{Sec}_2(\mathcal{G}_{3,3})$ equals 17. This is two less than the expected dimension of 19. We have here identified the smallest instance of defectivity.

Let $m = (m_{ijk})$ be a valid vector of moments. Thus m is a point in $\mathrm{Sec}_2(\mathcal{G}_{3,3})$. We assume that $m \notin \mathcal{G}_{3,3}$. Choose arbitrary but distinct complex numbers for μ_{11} and μ_{21}, while the other 17 model parameters remain unknowns. We note that, if $\mu_{11} = \mu_{21}$, then $m_{300} = 3m_{100}m_{200} - 2m_{100}^3$. This is not satisfied for a general choice of 19 model parameters.

What we see below is a system of 19 polynomial equations in 17 unknowns. We claim that this system has a unique solution over $\mathbb{C}$. Hence, if $\mu_{11}, \mu_{21} \in \mathbb{Q}$ and the left hand side vector m has its coordinates in $\mathbb{Q}$, then that unique solution has its coordinates in $\mathbb{Q}$.

$$
\begin{aligned}
m_{100} &= \alpha\mu_{11} + (1-\alpha)\mu_{21} \\
m_{010} &= \alpha\mu_{12} + (1-\alpha)\mu_{22} \\
m_{001} &= \alpha\mu_{13} + (1-\alpha)\mu_{23} \\
m_{200} &= \alpha(\mu_{11}^2 + \sigma_{111}) + (1-\alpha)(\mu_{21}^2 + \sigma_{211}) \\
m_{020} &= \alpha(\mu_{12}^2 + \sigma_{122}) + (1-\alpha)(\mu_{22}^2 + \sigma_{222}) \\
m_{002} &= \alpha(\mu_{13}^2 + \sigma_{133}) + (1-\alpha)(\mu_{23}^2 + \sigma_{233}) \\
m_{110} &= \alpha(\mu_{11}\mu_{12} + \sigma_{112}) + (1-\alpha)(\mu_{21}\mu_{22} + \sigma_{212}) \\
m_{101} &= \alpha(\mu_{11}\mu_{13} + \sigma_{113}) + (1-\alpha)(\mu_{21}\mu_{23} + \sigma_{213}) \\
m_{011} &= \alpha(\mu_{12}\mu_{13} + \sigma_{123}) + (1-\alpha)(\mu_{22}\mu_{23} + \sigma_{223}) \\
m_{300} &= \alpha(\mu_{11}^3 + 3\sigma_{111}\mu_{11}) + (1-\alpha)(\mu_{21}^3 + 3\sigma_{211}\mu_{21}) \\
m_{030} &= \alpha(\mu_{12}^3 + 3\sigma_{122}\mu_{12}) + (1-\alpha)(\mu_{22}^3 + 3\sigma_{222}\mu_{22}) \\
m_{003} &= \alpha(\mu_{13}^3 + 3\sigma_{133}\mu_{13}) + (1-\alpha)(\mu_{23}^3 + 3\sigma_{233}\mu_{23}) \\
m_{210} &= \alpha(\mu_{11}^2\mu_{12} + \sigma_{111}\mu_{12} + 2\sigma_{112}\mu_{11}) + (1-\alpha)(\mu_{21}^2\mu_{22} + \sigma_{211}\mu_{22} + 2\sigma_{212}\mu_{21}) \\
m_{201} &= \alpha(\mu_{11}^2\mu_{13} + \sigma_{111}\mu_{13} + 2\sigma_{113}\mu_{11}) + (1-\alpha)(\mu_{21}^2\mu_{23} + \sigma_{211}\mu_{23} + 2\sigma_{213}\mu_{21}) \\
m_{120} &= \alpha(\mu_{11}\mu_{12}^2 + \sigma_{122}\mu_{11} + 2\sigma_{112}\mu_{12}) + (1-\alpha)(\mu_{21}\mu_{22}^2 + \sigma_{222}\mu_{21} + 2\sigma_{212}\mu_{22}) \\
m_{102} &= \alpha(\mu_{11}\mu_{13}^2 + \sigma_{133}\mu_{11} + 2\sigma_{113}\mu_{13}) + (1-\alpha)(\mu_{21}\mu_{23}^2 + \sigma_{233}\mu_{21} + 2\sigma_{213}\mu_{23}) \\
m_{021} &= \alpha(\mu_{12}^2\mu_{13} + \sigma_{122}\mu_{13} + 2\sigma_{123}\mu_{12}) + (1-\alpha)(\mu_{22}^2\mu_{23} + \sigma_{222}\mu_{23} + 2\sigma_{223}\mu_{22}) \\
m_{012} &= \alpha(\mu_{12}\mu_{13}^2 + \sigma_{133}\mu_{12} + 2\sigma_{123}\mu_{13}) + (1-\alpha)(\mu_{22}\mu_{23}^2 + \sigma_{233}\mu_{22} + 2\sigma_{223}\mu_{23}) \\
m_{111} &= \alpha\left(\mu_{11}\mu_{12}\mu_{13} + \sigma_{112}\mu_{13} + \sigma_{113}\mu_{12} + \sigma_{123}\mu_{11}\right) \\
&\quad + (1-\alpha)(\mu_{21}\mu_{22}\mu_{23} + \sigma_{212}\mu_{23} + \sigma_{213}\mu_{22} + \sigma_{223}\mu_{21})
\end{aligned}
$$

By solving the first equation, we obtain the mixture parameter α. From the second and third equation we can eliminate μ_{12} and μ_{13}. Next, we observe that all 12 covariances σ_{ijk} appear linearly in our equations, so we can solve for these as well. We are left with a system of truly non-linear equations in only two unknowns, μ_{22} and μ_{23}. A direct computation now reveals that this system has a unique solution that is a rational expression in the given m_{ijk}.

Our computational argument therefore shows that each general fiber of the natural parametrization of $\mathrm{Sec}_2(\mathcal{G}_{3,3})$ is birational to the affine plane $\mathbb{A}^2$ whose coordinates are μ_{11} and μ_{21}. This establishes Theorem 5.2.2 for the special case of trivariate Gaussians ($n = 3$).

Remark 5.2.4. The second assertion in Theorem 5.2.2 holds for $n = 2$ because there are 11 parameters and $\mathrm{Sec}_2(\mathcal{G}_{2,3}) = \mathbb{P}^9$. However, the third assertion is not true for $n = 2$ because the general fiber of the parametrization map $\mathbb{A}^{11} \to \mathbb{P}^9$ is the union of three irreducible components. When μ_{11} and μ_{21} are fixed, then the fiber consists of three points and not one.

Proof of Theorem 5.2.2. Suppose $n \geq 4$ and let $m \in \mathrm{Sec}_2(\mathcal{G}_{n,3})\backslash\mathcal{G}_{n,3}$. Each moment $m_{i_1 i_2 \cdots i_n}$ has at most three non-zero indices. Hence, its expression in the model parameters involves at most three coordinates of the mean vectors and a

block of size at most three in the covariance matrices. Let μ_{11} and μ_{21} be arbitrary distinct complex numbers. Then we can apply the rational solution in Example 5.2.3 for any 3-element subset of $\{1, 2, \ldots, n\}$ that contains 1. This leads to unique expressions for all model parameters in terms of the moments $m_{i_1 i_2 \cdots i_n}$. In this manner, at most one system of parameters is recovered. Hence the third sentence in Theorem 5.2.2 is implied by the first two sentences. It is these two we shall now prove.

In the affine space $\mathbb{A}^N = \{m_{000} = 1\} \subset \mathbb{P}^N$, we consider the affine moment variety $G_n^A := \mathcal{G}_{n,3} \cap \mathbb{A}^N$. This has dimension $M = \frac{1}{2}n(n+3)$. The map from (5.2.1) that parametrizes the Gaussian moments is denoted $\rho : \mathbb{A}^M \to \mathbb{A}^N$. It is an isomorphism onto its image G_n^A.

Fix two points $p = (\mu, \sigma)$ and $p' = (\mu', \sigma')$ in $\mathbb{A}^M$. They determine the affine plane

$$A(p, p') \;=\; \big\{ (s\mu + (1-s)\mu',\, t\sigma + (1-t)\sigma') \mid s, t \in \mathbb{R} \big\} \;\subset\; \mathbb{A}^M.$$

Its image $\rho(A(p, p'))$ is a surface in $G_n^A \subset \mathbb{A}^N$. The restrictions $m_{i_1 \ldots i_n}(s, t)$ of the moments to this surface are polynomials in s, t with coefficients that depend on the points p, p'. Since $i_1 + \cdots + i_n \leq 3$, every moment $m_{i_1 \ldots i_n}(s, t)$ is a linear combination of the monomials $1, s, t, st, s^2, s^3$. Linearly eliminating these monomials, we obtain $N-5$ linear relations among the moments when restricted to the plane $A(p, p')$. These relations define the affine span of the surface $\rho(A(p, p'))$. This affine space is therefore 5-dimensional. We denote it by $\mathbb{A}^5_{p,p'}$.

The monomials $(b_1, b_2, b_3, b_4, b_5) = (s, t, st, s^2, s^3)$ serve as coordinates on $\mathbb{A}^5_{p,p'}$, modulo the affine-linear relations that define $\mathbb{A}^5_{p,p'}$. The image surface $\rho(A(p, p'))$ is therefore contained in the subvariety of $\mathbb{A}^5_{p,p'}$ that is defined by the 2×2-minors of the 2×4-matrix

$$\begin{pmatrix} 1 & b_2 & b_1 & b_4 \\ b_1 & b_3 & b_4 & b_5 \end{pmatrix} \;=\; \begin{pmatrix} 1 & t & s & s^2 \\ s & st & s^2 & s^3 \end{pmatrix}. \tag{5.2.2}$$

This variety is an irreducible surface, namely a scroll of degree 4. It hence equals $\rho(A(p, p'))$.

Let $\bar{\sigma}$ denote the covariance matrix with entries $\bar{\sigma}_{ij} = (\mu_i - \mu'_i)(\mu_j - \mu'_j)$. We define

$$\mathbb{A}^3_{p,p'} \;=\; \big\{ (\mu' + s(\mu - \mu'),\, \sigma' + t(\sigma - \sigma') + u\bar{\sigma}) \mid s, t, u \in \mathbb{R} \big\}.$$

Setting $u = 0$ shows that this 3-space contains the plane $A(p, p')$. We claim that

$$\rho(\mathbb{A}^3_{p,p'}) \;\subseteq\; \mathbb{A}^5_{p,p'}. \tag{5.2.3}$$

On the image $\rho(\mathbb{A}^3_{p,p'})$, each moment is a linear combination of the eight monomials

$1, s, s^2, s^3, t, st, u, su$. A key observation is that, by our choice of $\bar{\sigma}$, these expressions are actually linear combinations of the six expressions $1, s, s^2+u, s^3+3su, t, st$. Indeed, the coefficient of s^2 in the expansion of $(\mu_i' + s(\mu_i - \mu_i'))(\mu_j' + s(\mu_j - \mu_j'))$ matches the coefficient $\bar{\sigma}_{ij}$ of u in the expansion of second order moments. Likewise, s^2 and u have equal coefficients in the third order moments. Analogously, the coefficient of the monomial s^3 in the expansion of

$$(\mu_i' + s(\mu_i - \mu_i'))(\mu_j' + s(\mu_j - \mu_j'))(\mu_k' + s(\mu_k - \mu_k'))$$

is $(\mu_i - \mu_i')\bar{\sigma}_{jk} = (\mu_j - \mu_j')\bar{\sigma}_{ik} = (\mu_k - \mu_k')\bar{\sigma}_{ij}$, which coincides with the corresponding coefficient of $3su$ in the expansion of third order moments. From this we conclude that (5.2.3) holds.

Since ρ is birational, $\rho(\mathbb{A}^3_{p,p'})$ is a threefold in $\mathbb{A}^5_{p,p'}$. Since p and p' are arbitrary, these threefolds cover G_n^A. Through any point outside $\rho(\mathbb{A}^3_{p,p'})$ there is a 2-dimensional family of secant lines to $\rho(\mathbb{A}^3_{p,p'})$. The same holds for G_n^A. Hence the 2-defectivity of $\mathcal{G}_{n,3}$ is at least two.

To see that it is at most two, it suffices to find a point q in $\mathrm{Sec}_2(\mathcal{G}_{n,3})$ such that the variety of secant lines to $\mathcal{G}_{n,3}$ through q is 2-dimensional. Let $\mathcal{G}_{2,3}(1,2)$ denote the subvariety of $\mathcal{G}_{n,3}$ defined by setting all parameters other than $\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}$ to zero. The span of $\mathcal{G}_{2,3}(1,2) \cap \mathbb{A}^N$ is an affine 9-space $\mathbb{A}^9(1,2)$ inside $\mathbb{A}^N$. Consider a general point $q \in \mathbb{A}^9(1,2)$. Then $q \notin G_n^A$. We claim that any secant to G_n^A through q is contained in $\mathbb{A}^9(1,2)$.

A computation with `Macaulay2` [GS02] shows that this is the case when $n = 3$. Explicitly, if q is any point whose moment coordinates vanish except those that involve only $\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}$, then $\mu_3 = \sigma_{13} = \sigma_{23} = \sigma_{33} = 0$. Suppose now $n \geq 4$. Assume there exists a secant line through q that is not contained in $\mathbb{A}^9(1,2)$. Then we can find indices $1, 2, k$ such that the projection of that secant passes through the span of the corresponding $G_3^A \subset G_n^A$. In each case, the secant lands in $\mathbb{A}^9(1,2)$, so it must already lie in this subspace before any of the projections. This argument proves the claim.

In conclusion, we have shown that the 2-defectivity of the third order Gaussian moment variety $\mathcal{G}_{n,3}$ is precisely two. This completes the proof of Theorem 5.2.2. □

We offer some remarks on the geometry underlying the proof of Theorem 5.2.2, or more precisely, on the 2-dimensional family of secant lines through a general point q on the affine secant variety $\mathrm{Sec}_2(G_n^A)$. The *entry locus* Σ_q is the closure of the set of points $p \in G_n^A$ such that q lies on a secant line through p. This entry locus is therefore a surface. We identify the Zariski closure of this surface in $\mathbb{P}^N$.

Proposition 5.2.5. *The Zariski closure in $\mathcal{G}_{n,3}$ of the entry locus Σ_q of a general point $q \in \mathrm{Sec}_2(G_n^A)$ is the projection of a Del Pezzo surface of degree 6 into $\mathbb{P}^5$ that is singular along a line in the hyperplane at infinity.*

Proof. According to Example 5.2.3, the 2-dimensional family of secant lines through a general point $q \in \mathrm{Sec}_2(G_3^A)$ is irreducible and birational to the affine plane. If we consider G_3^A as a subvariety of G_n^A and $q \in \mathrm{Sec}_2(G_3^A)$, then we may argue as in the proof of Theorem 5.2.2 that any secant line to G_n^A through q, is a secant line to G_3^A. We conclude that the 2-dimensional family of secant lines through a general point $q \in \mathrm{Sec}_2(G_n^A)$ is irreducible.

On the other hand, if q is on the secant spanned by $p, p' \in G_n^A$, then, in the notation of the proof of Theorem 5.2.2, the point q lies in $\mathbb{A}^5_{p,p'}$. There is a 2-dimensional family of secant lines to $\rho(\mathbb{A}^3_{p,p'})$ through q. This family must coincide with the family of secant lines to G_n^A through q. The entry locus Σ_q therefore equals the double point locus of the projection

$$\pi_q : \rho(\mathbb{A}^3_{p,p'}) \to \mathbb{A}^4$$

from the point q. We shall identify this double point locus as a surface of degree 6. In fact, its Zariski closure in $\mathbb{P}^5$ is the projection of a Del Pezzo surface of degree 6 from $\mathbb{P}^6$.

Consider the maps

$$\begin{array}{rcccc} \tau & : & \mathbb{A}^3_{p,p'} \to \mathbb{A}^6 & : & (s,t,u) \mapsto (s,t,st,s^2,s^3+3su,u), \\ \pi & : & \mathbb{A}^6 \to \mathbb{A}^5_{p,p'} & : & (a_1,\ldots,a_6) \mapsto (a_1,a_2,a_3,a_4+a_6,a_5). \end{array}$$

The image $\tau(\mathbb{A}^3_{p,p'})$ in $\mathbb{A}^6$ is the 3-fold scroll defined by the 2×2 minors of the matrix

$$\begin{pmatrix} 1 & a_2 & a_1 & a_4+3a_6 \\ a_1 & a_3 & a_4 & a_5 \end{pmatrix}. \tag{5.2.4}$$

The composition $\pi \circ \tau$ is the restriction of ρ to $\mathbb{A}^3_{p,p'}$. Hence $\rho(\mathbb{A}^3_{p,p'})$ is also a *quartic threefold scroll*. To find its equations in $\mathbb{A}^5_{p,p'}$, we set $a_4 = b_4 - a_6$ and $a_i = b_i$ for $i \in \{1,2,3,5\}$, and then we eliminate a_6 from the ideal of 2×2-minors of (5.2.4). The result is the system

$$b_1b_2 - b_3 \;=\; 2b_1b_3^2 + b_2^2b_5 - 3b_2b_3b_4 \;=\; 2b_1^2b_3 + b_2b_5 - 3b_3b_4 \;=\; 2b_1^3 - 3b_1b_4 + b_5 \;=\; 0.$$

Let $X_{p,p'}$ be the Zariski closure of $\tau(\mathbb{A}^3_{p,p'})$ in $\mathbb{P}^6$. It is a threefold quartic scroll, defined by the 2×2 minors of the matrix

$$\begin{pmatrix} a_0 & a_2 & a_1 & a_4+3a_6 \\ a_1 & a_3 & a_4 & a_5 \end{pmatrix}. \tag{5.2.5}$$

The projection π, and the composition of π and the projection π_q from the point

$q \in \mathbb{A}^5_{p,p'}$, extend to projections

$$\bar{\pi} : X_{p,p'} \to \mathbb{P}^5 \qquad \text{and} \qquad \tilde{\pi} : X_{p,p'} \to \mathbb{P}^4.$$

By the double point formula [Ful13, Theorem 9.3], the double point locus $\Sigma_{\tilde{\pi}} \subset X_{p,p'}$ of $\tilde{\pi}$ is a surface of degree 6 anticanonically embedded in $\mathbb{P}^6$. This is the desired Del Pezzo surface.

Similarly, the double point locus of $\bar{\pi}$ is a plane conic curve in $X_{p,p'}$, that is mapped 2: 1 onto a line in $\mathbb{P}^5$. The plane conic curve is certainly contained in the double point locus $\Sigma_{\tilde{\pi}}$, so $\bar{\pi}(\Sigma_{\tilde{\pi}}) \subset \mathbb{P}^5$ is singular along a line. In the above coordinates, the conic is the intersection of $X_{p,p'}$ with the plane defined by $a_0 = a_1 = a_2 = a_3 = 0$, i.e. a conic in the hyperplane $\{a_0 = 0\}$ at infinity. The entry locus Σ_q is clearly contained in $\bar{\pi}(\Sigma_{\tilde{\pi}})$. In fact, the latter is the Zariski closure of the former in $\mathbb{P}^5$ and the proposition follows. □

We now come to the higher secant varieties of the Gaussian moment variety $\mathcal{G}_{n,3}$.

Corollary 5.2.6. *Let $k \geq 2$ and $n \geq 3k-3$. Then $\mathcal{G}_{n,3}$ is k-defective.*

Proof. This is immediate from Theorem 5.2.2 and Proposition 5.1.2. □

Based on computations, like those in Table 5.2.1, we propose the following conjecture.

Conjecture 5.2.7. *For any $n \geq 2$ and $k \geq 1$, we have*

$$\dim(\mathrm{Sec}_k(\mathcal{G}_{n,3})) \;=\; \frac{1}{6}k\left[k^2 - 3(n+4)k + 3n(n+6) + 23\right] - (n+2), \qquad (5.2.6)$$

for $k = 1, 2, \ldots, K$, where $K+1$ is the smallest integer such that the right hand side in (5.2.6) is larger than the ambient dimension $\binom{n+3}{3} - 1$.

For $k = 1$ this formula evaluates to $\dim(\mathcal{G}_{n,3}) = n(n+3)/2$, as desired. Conjecture 5.2.7 also holds for $k = 2$. This is best seen by rewriting the identity (5.2.6) as follows:

$$\frac{1}{2}kn(n+3)+k-1 - \dim\big(\mathrm{Sec}_k(\mathcal{G}_{n,3})\big) \;=\; \frac{1}{2}(k-1)(k-2)n - \frac{1}{6}(k-1)(k^2-11k+6).$$

This is the difference between the expected dimension and the true dimension of the kth secant variety. For $k = 2$ this equals 2, independently of n, in accordance with Theorem 5.2.2.

Conjecture 5.2.7 was verified computationally for $n \leq 15$. Table 5.2.1 illustrates all cases for $n \leq 10$. Here, $\exp = \min(\mathrm{par}, N)$ is the *expected dimension*, and $\delta = \exp - \dim$ is the *defect*.

n	k	d	par	N	exp	dim	δ	par-dim
3	2	3	19	19	19	17	2	2
4	2	3	29	34	29	27	2	2
5	2	3	41	55	41	39	2	2
5	3	3	62	55	55	51	4	11
6	2	3	55	83	55	53	2	2
6	3	3	83	83	83	71	12	12
6	4	3	111	83	83	82	1	29
7	2	3	71	119	71	69	2	2
7	3	3	107	119	107	94	13	13
7	4	3	143	119	119	111	8	32
8	2	3	89	164	89	87	2	2
8	3	3	134	164	134	120	14	14
8	4	3	179	164	164	144	20	35
8	5	3	224	164	164	160	4	64
9	2	3	109	219	109	107	2	2
9	3	3	164	219	164	149	15	15
9	4	3	219	219	219	181	38	38
9	5	3	274	219	219	204	15	70
10	2	3	131	285	131	129	2	2
10	3	3	197	285	197	181	16	16
10	4	3	263	285	263	222	41	41
10	5	3	329	285	285	253	32	76
10	6	3	395	285	285	275	10	120
11	2	3	155	363	155	153	2	2
11	3	3	233	363	233	216	17	17
11	4	3	311	363	311	267	44	44
11	5	3	389	363	363	307	56	82
11	6	3	467	363	363	337	26	130
11	7	3	545	363	363	358	5	187
12	2	3	181	454	181	179	2	2
12	3	3	272	454	272	254	18	18
12	4	3	363	454	363	316	47	47
12	5	3	454	454	454	366	88	88
12	6	3	545	454	454	405	49	140
12	7	3	636	454	454	434	20	202

Table 5.2.1.: Moment varieties of order $d = 3$ for mixtures of $k \geq 2$ Gaussians

We also undertook a comprehensive experimental study for higher moments of multivariate Gaussians. The following two examples are the two smallest defective cases for $d = 4$.

Example 5.2.8. Let $n = 8$ and $d = 4$. The Gaussian moment variety $\mathcal{G}_{8,4}$ is 11-defective. The expected dimension of $\mathrm{Sec}_{11}(\mathcal{G}_{8,4})$ equals the ambient dimension $N = 494$, but this secant variety is actually a hypersurface in $\mathbb{P}^{494}$. It would be very nice to know its degree.

Example 5.2.9. Let $n = 9$ and $d = 4$. The moment variety $\mathcal{G}_{9,4}$ is 12-defective but it is not 11-defective. Thus the situation is much more complicated than that in Theorem 5.2.2, where defectivity always starts at $k = 2$. We do not yet have any theoretical explanation for this.

Table 5.2.2 shows the first few defective cases for Gaussian moments of order $d = 4$. It suggests a clear pattern, resulting in the following conjecture. We verified this for $n \leq 14$.

Conjecture 5.2.10. *The Gaussian moment variety $\mathcal{G}_{n,4}$ is $(n+3)$-defective with defect $\delta_{n+3} = 1$ for $n \geq 8$. Furthermore, for all $r \geq 3$, the $(n+r)$-defect of $\mathcal{G}_{n,4}$ is equal to $\delta_{n+r} = \binom{r-1}{2}$, unless the number of model parameters exceeds the ambient dimension $\binom{n+4}{4} - 1$.*

5.3. Towards Equations and Degrees

Our problem is to study the higher secant variety $\mathrm{Sec}_k(\mathcal{G}_{1,d})$ of the moment surface $\mathcal{G}_{1,d} \subset \mathbb{P}^d$ whose equations were given in Proposition 4.4.2. The hypersurface $\mathrm{Sec}_2(\mathcal{G}_{1,6})$ was treated in Theorem 4.2.3. In the derivation of its equation in the previous section, we started out with introducing the new unknowns $s = \mu + \nu$ and $p = \mu\nu$. After introducing cumulant coordinates, the defining expressions for the moments m_4, m_5, m_6 in (4.2.1) turned into the three equations (4.2.12),(4.2.13) in $k_2, k_3, k_4, k_5, k_6, s, p$, and from these we then eliminated s and p.

The implicitization problem for $\mathrm{Sec}_2(\mathcal{G}_{1,d})$ when $d > 6$ can be approached with the same process. Starting from the moments, we derive polynomials in $k_2, k_3, \ldots, k_d, s$ and p that contain k_d linearly. The extra polynomial that contains k_7 linearly and is used for $\mathrm{Sec}_2(\mathcal{G}_{1,7})$ equals

$$\begin{aligned} &16p^3s^5 - 126k_2p^3s^3 + 42k_3p^2s^4 - 148p^4s^3 + 252k_2p^4s - 126k_3p^3s^2 \\ &+216p^5s + 315k_2k_3^2ps - 1260k_2k_3p^3 - 35k_3^3s^2 + 210k_3^2p^2s - 378k_3p^4 \\ &+189k_2k_5p^2 + 35k_3^3p + 315k_3k_4p^2 + 9k_7p^2. \end{aligned} \tag{5.3.1}$$

n	k	d	par	N	exp	dim	δ	par-dim
8	11	4	494	494	494	493	1	1
9	12	4	659	714	659	658	1	1
9	13	4	714	714	714	711	3	3
10	13	4	857	1000	857	856	1	1
10	14	4	923	1000	923	920	3	3
10	15	4	989	1000	989	983	6	6
11	14	4	1091	1364	1091	1090	1	1
11	15	4	1169	1364	1169	1166	3	3
11	16	4	1247	1364	1247	1241	6	6
11	17	4	1325	1364	1325	1315	10	10
12	15	4	1364	1819	1364	1363	1	1
12	16	4	1455	1819	1455	1452	3	3
12	17	4	1546	1819	1546	1540	6	6
12	18	4	1637	1819	1637	1627	10	10
12	19	4	1728	1819	1728	1713	15	15
12	20	4	1819	1819	1819	1798	21	21
13	16	4	1679	2379	1679	1678	1	1
13	17	4	1784	2379	1784	1781	3	3
13	18	4	1889	2379	1889	1883	6	6
13	19	4	1994	2379	1994	1984	10	10
13	20	4	2099	2379	2099	2084	15	15
13	21	4	2204	2379	2204	2183	21	21
13	22	4	2309	2379	2309	2281	28	28
13	23	4	2414	2379	2379	2378	1	36

Table 5.2.2.: A census of defective Gaussian moment varieties $d = 4$

The extra polynomial that contains k_8 linearly and is used for $\mathrm{Sec}_2(\mathcal{G}_{1,8})$ equals

$$\begin{gathered}20p^4s^6 + 336k_2p^4s^4 - 112k_3p^3s^5 + 124p^5s^4 - 3780k_2^2p^4s^2 + 2520k_2k_3p^3s^3 - 6048k_2p^5s^2 \\ -420k_3^2p^2s^4 + 2128k_3p^4s^3 - 2232p^6s^2 - 7560k_2^2k_3p^3s + 11340k_2^2p^5 + 2520k_2k_3^2p^2s^2 \\ -15120k_2k_3p^4s + 12096k_2p^6 - 280k_3^3ps^3 + 2940k_3^2p^3s^2 - 7056k_3p^5s + 3564p^7 \\ +1890k_2^2k_3^2p^2 + 5670k_2^2k_4p^3 - 420k_2k_3^3ps + 7560k_2k_3^2p^3 + 35k_3^4s^2 + 280k_3^3p^2s \\ -1260k_3^2p^4 + 756k_2k_6p^3 - 35k_3^4p + 1512k_3k_5p^3 + 945k_4^2p^3 + 27k_8p^3\end{gathered} \tag{5.3.2}$$

Proposition 5.3.1. *The ideals of the* 5*-dimensional varieties* $\mathrm{Sec}_2(\mathcal{G}_{1,7}) \cap \mathbb{A}^7$ *and* $\mathrm{Sec}_2(\mathcal{G}_{1,8}) \cap \mathbb{A}^8$ *in cumulant coordinates are obtained from (4.2.12), (4.2.13), (5.3.1) and (5.3.2) by eliminating* s *and* p*.*

The polynomials above represent a sequence of birational maps of the form

$\mathrm{Sec}_2(\mathcal{G}_{1,d}) \dashrightarrow \mathrm{Sec}_2(\mathcal{G}_{1,d-1})$, which allow us to recover all cumulants from earlier cumulants and the parameters p and s. In particular, by solving the equation (4.2.2) for p and then recovering s from the expression in Remark 4.2.2, we can invert the parametrization for any of the moment varieties $\mathrm{Sec}_2(\mathcal{G}_{1,d}) \subset \mathbb{P}^d$. If we are given m_1, m_2, m_3, m_4, m_5 from data then we expect $18 = 9 \times 2$ complex solutions $(\lambda, \mu, \nu, \sigma, \tau)$. The extra factor of 2 comes from label swapping between the two Gaussians. In that sense, the number 9 is the algebraic degree of the identifiability problem for $n = 1$ and $k = 2$.

We next move on to $k = 3$. There are now eight model parameters. These are mapped to $\mathbb{P}^8$ with coordinates $(m_0 : m_1 : \cdots : m_8)$, and we are interested in the degree of that map.

Working in cumulant coordinates, and using the Gröbner basis package `FGb` in `maple`, we computed the degree of that map. It turned out to be $1350 = 3! \cdot 225$.

Theorem 5.3.2. *The mixture model of $k = 3$ univariate Gaussians is algebraically identifiable from its first eight moments. The algebraic degree of this identifiability problem equals* 225.

We also computed a generalized Pearson polynomial of degree 225 for $k = 3$. Namely, we replace the three means μ_1, μ_2, μ_3 by their elementary symmetric polynomials $e_1 = \mu_1 + \mu_2 + \mu_3$, $e_2 = \mu_1\mu_2 + \mu_1\mu_3 + \mu_2\mu_3$ and $e_3 = \mu_1\mu_2\mu_3$. This is done by a derivation analogous to (4.2.7)-(4.2.12). This allows us to eliminate all model parameters other than e_1, e_2, e_3. The details and further computational results will be presented in a forthcoming article.

We compute a lexicographic Gröbner basis $\mathcal{G}$ for the above equations in the polynomial ring $\mathbb{R}[e_1, e_2, e_3]$, with generic numerical values of the eight moments $m_1, \ldots, m_8$. It has the expected shape

$$\mathcal{G} \;=\; \bigl\{ f(e_1), e_2 - g(e_1), e_3 - h(e_1) \bigr\}.$$

Here f, g, h are univariate polynomials of degrees $225, 224, 224$ respectively. In particular, f is the promised generalized Pearson polynomial of degree 225 for mixtures of three Gaussians.

For general k, the mixture model has $3k - 1$ parameters. By Theorem 5.0.6 we know it is algebraically identifiable, with a finite number of solutions. Based on what we know for $k = 2$ and $k = 3$, we offer the following conjecture concerning the degree of algebraic identifiability, along with our belief of when we can obtain rational identifiability.

Recall that the *double-factorial* is the product of the smallest odd positive integers:

$$(2k-1)!! \;\;=\;\; 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2k-1).$$

Conjecture 5.3.3. *When algebraically identifying a mixture of k univariate Gaussians by the moments of order $\leq 3k-1$, the degree of this identifiability problem equals $\big((2k-1)!!\big)^2$. Moreover, this model is rationally identifiable by the moments of order $\leq 3k$.*

The double-factorial part of the conjecture is nothing but a wild guess. We do not even know the degree of the hypersurface $\mathrm{Sec}_3(\mathcal{G}_{1,9}) \subset \mathbb{P}^9$.

Computations for $k = 4$ appear currently out of reach for Gröbner basis methods. If our wild guess is true then the expected number of complex solutions for the 11 moment equations whose solution identifies a mixture of $k = 4$ univariate Gaussians is $105^2 \times 4! = 264,600$.

Current work is using techniques in Numerical Algebraic Geometry [SVW96, BHSW06] to confirm or refute this number [AR16, Example 5.3]. The main approach is to try to exploit the symmetry of the solution set to these systems of equations, so that tracking the number of solutions becomes less costly computationally.

Example 5.3.4. Let $n = 1$, $k = 2$ and $d = 7$. The following results were obtained using methods from numerical algebraic geometry. The 5-dimensional variety $\mathrm{Sec}_2(\mathcal{G}_{1,7})$ has degree 105 in $\mathbb{P}^7$. The eight coordinate projections, defined algebraically by eliminating each one of $m_{07}, m_{16}, \ldots, m_{70}$ from the ideal of $\mathrm{Sec}_2(\mathcal{G}_{1,7})$, are hypersurfaces in $\mathbb{P}^6$. Their degrees are $85, 99, 104, 95, 78, 66, 48$ and 39 respectively. This suggests that there are no low degree generators in the ideal of $\mathrm{Sec}_2(\mathcal{G}_{1,7})$. In fact, a state-of-the-art Gröbner basis computation by Jean-Charles Faugère shows that the smallest degree of such a minimal generator is 25.

One would expect that it is even more difficult to describe the prime ideals of the secant varieties $\mathrm{Sec}_k(\mathcal{G}_{n,d})$ for $n \geq 2$, $k > 1$. With ideal generators out of reach, we first ask for the degrees of our secant varieties.

Conjecture 5.3.5. *For fixed k and n, the function $d \mapsto \deg \mathrm{Sec}_k(\mathcal{G}_{n,d})$ is a polynomial in d, starting from the smallest value of d where the secant variety does not fill the ambient space.*

The numerical `Macaulay2` [GS02] package `NumericalImplicitization.m2`, developed by Chen and Kileel [CK16], was very useful for us. It was able to compute the desired degrees in some interesting cases. These data points led us to Conjecture 5.3.5 and to the following result.

Proposition 5.3.6. *Suppose that Conjecture 5.3.5 holds for $k = 2$ and $n = 1$. Then, for all $d \geq 6$, the degree of the dth moment variety for mixtures of two univariate Gaussians equals*

$$\deg\ \mathrm{Sec}_2(\mathcal{G}_{1,d}) \ = \ \frac{(d+7)(d-4)(d-3)(d-2)}{8}. \tag{5.3.3}$$

Proof. Let X_d be a general variety defined by a Hilbert-Burch matrix B_d as in (5.1.1). Here 'general' means that the entries in B_d are generic linear forms in x, y, z. Using the double point formula in intersection theory [Ful13, Sec. 9.3] for a general projection $X_d \to \mathbb{P}^4$, we compute

$$\deg \operatorname{Sec}_2(X_d) \;=\; \frac{(d-4)(d-3)(d^2+5d-2)}{8}. \tag{5.3.4}$$

Since $\mathcal{G}_{1,d}$ is singular, the degrees of its secant varieties are lower than (5.3.4), with a correction term accounting for the singular line in Lemma 4.4.5. The assumption that Conjecture 5.3.5 holds in our case implies that $d \mapsto \operatorname{Sec}_2(\mathcal{G}_{1,d})$ is a polynomial function of degree at most 4. Our numerical computation shows that the degrees of $\operatorname{Sec}_2(\mathcal{G}_{1,d})$ for $d = 6, \ldots, 10$ are 39, 105, 225, 420 and 714. These are enough to interpolate, and we obtain the polynomial in (5.3.3). □

Remark 5.3.7. The zeroes of (5.3.3) at $d = 2, 3, 4$ were not part of the interpolation but they are not unexpected. Also, substituting $d = 5$ into (5.3.3) recovers our famous Pearson degree 9 for identifying mixtures of two univariate Gaussians! Using `NumericalImplicitization.m2`, we verified the correctness of (5.3.3) up to $d = 11$.

Following this train of thought, and using the Le Barz classification formulas in [LB87], we compute an analogous formula to (5.3.4) for trisecants ($k = 3$) of a general surface X_d:

$$\deg(\operatorname{Sec}_3(X_d)) \;=\; \frac{(d-6)(d^5+3d^4-57d^3-43d^2+752d-512)}{48}.$$

Conjecture 5.3.5 now suggests that $d \mapsto \deg \operatorname{Sec}_3(\mathcal{G}_{1,d})$ is a polynomial function of degree 6. Unfortunately, we do not yet have numerical evidence for this. For instance, we do not even know the degree of $\operatorname{Sec}_3(\mathcal{G}_{1,9})$. The formula yields the upper bound $\deg(\operatorname{Sec}_3(X_9)) = 2497$.

We close with two more cases with $n \geq 2$ for which we were able to compute the degrees.

Example 5.3.8. Let $n = 2$ and $d = 4$. The 5-dimensional moment variety $\mathcal{G}_{2,4}$ has degree 102 in $\mathbb{P}^{14}$. It is not defective. Its secant variety $\operatorname{Sec}_2(\mathcal{G}_{2,4})$ has dimension 11 and degree 538.

Example 5.3.9. We return to Example 5.2.3, so $n = d = 3$. The Gaussian moment variety $\mathcal{G}_{3,3}$ has dimension 9 and degree 130 in $\mathbb{P}^{19}$. We had found the number 130 already in Proposition 4.4.13. This variety is 2-defective. Its secant variety $\operatorname{Sec}_2(\mathcal{G}_{3,3})$ has dimension 17 and degree 79. We do not know its ideal generators. As in Example 5.3.4, we studied the degrees of its coordinate projections. The 20

coordinates on $\mathbb{P}^{19}$ come in seven symmetry classes. Representatives are

$$m_{000}, m_{100}, m_{200}, m_{110}, m_{300}, m_{201}, m_{111}.$$

By omitting these coordinates, one at a time, we obtain hypersurfaces in $\mathbb{P}^{18}$ whose degrees are

$$58, 63, 34, 42, 25, 34, 40$$

respectively. □

Our geometric study of secants of Gaussian moment varieties in this chapter suggests there is a rich theory behind them, with much yet to be discovered.

6. Submodels and Tensor Decomposition

So far we have dealt with general Gaussian mixture models. However, one often encounters certain submodels that assume natural constraints on the parameters. We already encountered examples of this in Chapter 2 with isotropic and homoscedastic mixtures, and in Chapter 4 with equal means. We gather these definitions now.

Definition 6.0.1. A Gaussian mixture with defining parameters $\{(\alpha_i, \mu_i, \Sigma_i)\}_{i=1\ldots,k}$ is said to be

- *homoscedastic* if all the covariances are equal: $\Sigma_i = \Sigma$ for all i.
- *isotropic* if the all the covariances are *spherical*: $\Sigma_i = \sigma_i I$ for all i.
- *centered* if all the means are equal: $\mu_i = \mu$ for all i.
- *uniform* if all the weights are equal: $\alpha_i = \frac{1}{k}$ for all i.

For example, a homoscedastic isotropic mixture is closely related to the *k-means clustering* problem, where we can think of each group or cluster as corresponding to a different component of the Gaussian mixture. See [Bis06, Chapter 9] for details. The main connection is that when updating parameters in a common algorithm for k-means, one assigns each data point to a cluster in a 'hard' definite manner, while the EM algorithm allows for a 'soft' assignment (the probabilities γ_i of (3.3.1) in Section 3.3). In fact, having the common $\sigma \to 0$ in the equation updates for the EM algorithm will give the equation updates for such k-means algorithm.

As another example, Srebro's conjecture [Sre07] asks whether a uniform homoscedastic isotropic mixture can have a likelihood function with local maxima that are not global in the infinite sample limit (with data sampled from the distribution). One can further assume the common variance is just the identity matrix. The hope is that in this 'simple' model, mixtures of the above type cannot exist, so the EM algorithm can be guaranteed to not get stuck in local maxima and converge to a global maximum (permutations of the true parameters). Unfortunately for such hopes, it was recently shown in [JZB+16] that there are examples for each $k \geq 3$ such that the conjecture is false. Even worse, EM will converge with random initialization to these 'bad' critical points with high probability. This kind

of result motivates further to look at alternative estimation methods, such as the method of moments. We explore recent strategies based on MOM in the coming section.

6.1. Machine Learning and MOM Reawakening

As we have mentioned, the EM algorithm introduced in 1977 by Dempster, Laird and Rubin [DLR77] popularized the use of Gaussian mixtures. In order to avoid the main drawback of getting stuck at local maxima, alternatives were sought by several people. Of worthy mention is the line pioneered by Dasgupta [Das99], with the use of *spectral methods* by Vempala and Wang. Their focus was on finding polynomial time algorithms for clustering.

The real method of moments reawakening comes from the work of Moitra, Kalai and Valiant [KMV10, MV10]. They realize that the polynomial equations from Pearson's method are much more tractable than EM. They propose to solve the parameter estimation problem in $n > 1$ by considering projections that reduce to solve the $n = 1$ case treated by Pearson. Furthermore, unaware that [Laz04] already proved that the first six moments really uniquely identify a mixture of two univariate Gaussians, they set out to prove this fact. Their result gives a bound on rational identifiability for $n = 1$. Since this is very relevant for us (cf. Conjecture 5.3.3), we present this result and their technique.

Proposition 6.1.1. *([KMV12, Section 4.2]) Let f and g be two mixtures of k Gaussians over $\mathbb{R}$. Then $f - g$ is either identically zero or has at most $4k-2$ zeros.*

Their proof is based on induction on k and the main ingredient is a theorem by Hummel and Gidas [HG84] in the context of the heat equation. The assertion is that if one convolves a Gaussian with an analytic function $f : \mathbb{R} \to \mathbb{R}$ with at most m zeros, then the resulting convolution still has at most m zeros.

With the proposition, we get the following important corollary.

Corollary 6.1.2. *If f and g are two univariate mixtures of k Gaussians that match in the first $4k - 2$ moments, then necessarily $f = g$. Thus, the minimum order d needed for rational identifiability is $\leq 4k - 2$.*

Remark 6.1.3. Note that for $k = 2$, we have $4 \times 2 - 2 = 6$, coinciding with what we know from Lazard [Laz04].

Proof. Assume $h = f - g$ is not identically zero. Let $p(x) : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree $4k - 2$ that matches the same sign as $h(x)$. Note this is possible since by Proposition 6.1.1, h has at most $4k - 2$ zeros. Thus, writing $p(x) = \sum_{j=1}^{4k-2} c_j x^j$

we have that

$$\begin{aligned} 0 \;<\; \int_{-\infty}^{\infty} p(x)h(x)dx \;&=\; \sum_{j=1}^{4k-2} \int_{-\infty}^{\infty} c_j x^j h(x)dx = \sum_{j=1}^{4k-2} c_j \int_{-\infty}^{\infty} x^j (f(x) - g(x))dx \\ &=\; \sum_{j=1}^{4k-2} c_j (m_j(f) - m_j(g)). \end{aligned}$$

Since the right hand side has to be positive, we conclude that there must exist a $j \in \{1, \ldots, 4k-2\}$ such that the jth moments $m_j(f)$ and $m_j(g)$ differ. □

Knowing this, they propose in [KMV12, Section 3] that to generalize Pearson's sixth moment test to $k > 2$, one should compute the first $4k-2$ sample moments and try to return the parameters that most closely match these. This gives a 'polynomially robust identifiable' algorithm for learning Gaussian mixtures. Indeed, their main Theorem 2 in [KMV12] shows that if two parameter sets for mixtures f and g differ by ϵ, then their first $4k-2$ moments have a $poly(\epsilon)$ discrepancy. Later Hardt and Price would prove optimality with respect to sample complexity [HP15]. They claim their result can be interpreted as showing that 'Pearson's original estimator is in fact an optimal solution to the problem he proposed'.

Remark 6.1.4. Let $d_{rat}(k)$ be the minimum order d needed so that mixtures of k univariate Gaussians are rationally identifiable from the first d moments. Recall that this means that generically there is at most one Gaussian mixture density with any sequence of such moments. Combining our algebraic identifiability result in Theorem 5.0.6 with the corollary above, we conclude that

$$3k - 1 \;\leq\; d_{rat}(k) \;\leq\; 4k - 2.$$

When the degree of algebraic identifiability is greater than 1 we can claim that the first inequality is strict. In this way, for $k = 2$ we recover $d_{rat}(2) = 6$, and for $k = 3$ we can say $8 < d_{rat}(3) \leq 10$ so that $d_{rat}(3)$ is either 9 or 10. We have conjectured in Conjecture 5.3.3 that $d_{rat}(k) = 3k$.

Recently, an MIT Summer Program for Undergraduate Research (SPUR) project [GBW16] proposed by Ankur Moitra explored the upper bound $4k-2$. The authors observe that there do exist pairs of distinct univariate Gaussian mixtures that have identical first $4k-3$ moments. However, their examples are from the equal-means submodel. After centering at $\mu = 0$, all the odd order moments are 0, so in this case one needs the first full $4k-2 = 2(2k-1)$ moments (the first $2k-1$ even moments) to determine the k unknown variances and the $k-1$ independent mixture weights. For the general unequal means model they also arrive at our conjecture of $3k$. □

At the same time that Kalai, Moitra and Valiant were proposing a return to the method of moments, Belkin and Sinha [BS10, BS15] realized that the nice polynomial structure extends to any distribution such that its moments are given by polynomials in the parameters. In terms of identifiability, Hilbert's Basis Theorem implies that there always exists a finite d_{rat} that will uniquely identify the parameters from sufficiently enough moments. Another important consequence of the semialgebraic nature of the moments and parameters is that one can obtain again polynomial $poly(\epsilon)$ discrepancies in the moments from ϵ discrepancy of the parameters. This is a nice application of the Tarski-Seidenberg Theorem from Real Algebraic Geometry [BCR13, Section 1.4]: semialgebraic sets are closed under projections.

Hsu and Kakade follow the method of moments revival trend and consider the learning of mixtures of isotropic Gaussians from the moments up to order $d = 3$ [HK13]. In order to solve the moment equations, they propose to find orthogonal decompositions of the second and third order moment tensors that reveal the mixture parameters. See Remark 4.3.2. More precisely,

Theorem 6.1.5. *(Theorem 1, [HK13]) Assume the means $\mu_i \in \mathbb{R}^n$ of an isotropic Gaussian mixture span a k-dimensional subspace and all $\alpha_i > 0$. Then the average variance $\bar{\sigma} := \sum_{i=1}^k \alpha_i \sigma_i$ is the smallest eigenvalue of the covariance matrix $\mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$. Furthermore, let $v \in \mathbb{R}^n$ be any unit norm eigenvector corresponding to the eigenvalue $\bar{\sigma}$. Define*

$$M_1 := \mathbb{E}[x(v^T(x - \mathbb{E}[x]))^2] \in \mathbb{R}^n,$$
$$M_2 := \mathbb{E}[x \otimes x] - \bar{\sigma} I \in \mathbb{R}^{n \times n},$$
$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \sum_{i=1}^n \big(M_1 \otimes e_i \otimes e_i + e_i \otimes M_1 \otimes e_i + e_i \otimes e_i \otimes M_1\big) \in \mathbb{R}^{n \times n \times n}$$

Then

$$M_1 = \sum_{i=1}^k \alpha_i \ \sigma_i \mu_i, \qquad M_2 = \sum_{i=1}^k \alpha_i \ \mu_i \otimes \mu_i, \qquad M_3 = \sum_{i=1}^k \alpha_i \ \mu_i \otimes \mu_i \otimes \mu_i.$$

One key aspect is that an orthogonal tensor decomposition when $d > 2$ is unique [BDHR15]. The main technique to find such a tensor decomposition is via the tensor power method, which extends the usual power method for matrices. The behavior is quite different from the matrix case: any eigenvector is a stable fixed point of the power map (not only the one corresponding to the largest eigenvalue). We refer to [Rob16, ASS17] for details.

While this method will cover a range of situations, note that the critical hypothesis that the means span a k-dimensional subspace in $\mathbb{R}^n$ implies $k \leq n$. Thus,

it is not applicable even in the one-dimensional case $n = 1$ (where all mixtures are automatically isotropic).

Why would we expect the mixtures in Theorem 6.1.5 to be identifiable from moments up to order $d = 3$? An n dimensional isotropic mixture of k components has $k(n+1)+(k-1)$ unknown parameters. On the other hand, there are $\binom{n+3}{3}-1$ moments up to order 3. So if $k \leq n$:

$$k(n+1)+(k-1) \leq n^2+2n-1 < \frac{n^3+6n^2+11n}{6}.$$

That is, we have $O(n^2)$ parameters versus $O(n^3)$ moments, with growing difference as $n \to \infty$. The same holds for the simpler submodel of homoscedastic isotropic mixtures with $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$ that they also consider (see [AGH$^+$14, Theorem 3.2]). The tensor decomposition method is applicable to some other models with latent variables, as explored in [AGH$^+$14].

More recently, Ge, Huang and Kakade discuss the possible polynomial learning complexity for general mixture of Gaussians without assuming isotropic, in [GHK15]. However, as their title *'Learning Mixtures of Gaussians in High Dimensions'* suggests, their method works in a very large ambient space where $n \geq 100k^2$. They measure the moments of orders $d = 3, 4$ and 6 and rely on a smoothed analysis setting [ST04] to estimate perturbed parameters.

It is not uncommon to find in these sorts of methods that identifiability assumptions are made. Based on the polynomial nature of the moments of many common distributions, in [WCL15] the authors propose semidefinite programming (SDP) approximations with linear moment constraints to approach learning parameters from any mixture model of a polynomial family (in the sense of Belkin and Sinha [BS15]). However, for mixtures of k univariate Gaussians [WCL15, Example 2.1] they propose to take only the first 6 moments and state that their framework can recover the parameters. They note that the 6 moments they use have been shown to be sufficient for $k = 2$ (incorrectly attributed to Pearson [Pea94]) and assume throughout that in their models, the number of observation functions that they consider uniquely identify the model (up to permutation of the components). It is very important for us to test the validity of these assumptions.

6.2. Veronese Subvarieties and the Alexander-Hirschowitz Theorem

In Remark 5.0.7 we observed that if we set $\sigma_{ij} = 0$ for all $1 \leq i, j \leq n$ in the parametrization of the Gaussian moment variety $\mathcal{G}_{n,d}$, then we recover the familiar Veronese variety $\mathcal{V}_{n,d}$ in Algebraic Geometry. Such a choice degenerates the Gaussian distribution to a Dirac point mass. Points on its secant variety $\mathrm{Sec}_k(\mathcal{V}_{n,d})$

represent moments of finitely supported signed measures on $\mathbb{R}^n$.

Our interest in these varieties is two-fold. On the one hand, as subvarieties of our somewhat complex moment varieties, they may shed some light on their properties. On the other hand, statistically they correspond to some of the submodels in the previous section, as we will see in Proposition 6.2.6.

Example 6.2.1. The moment variety $\mathcal{G}_{2,4}$ contains the *quartic Veronese surface* $\nu_4(\mathbb{P}^2)$. While we know from Proposition 4.4.11 that the former is not determinantal, the latter is. This surface is defined by 75 binomial quadrics in $\mathbb{P}^{14}$, included in the 2×2-minors of the matrix

$$\begin{pmatrix} m_{00} & m_{01} & m_{02} & m_{10} & m_{11} & m_{20} \\ m_{01} & m_{02} & m_{03} & m_{11} & m_{12} & m_{21} \\ m_{02} & m_{03} & m_{04} & m_{12} & m_{13} & m_{22} \\ m_{10} & m_{11} & m_{12} & m_{20} & m_{21} & m_{30} \\ m_{11} & m_{12} & m_{13} & m_{21} & m_{22} & m_{31} \\ m_{20} & m_{21} & m_{22} & m_{30} & m_{31} & m_{40} \end{pmatrix}. \tag{6.2.1}$$

As observed in [CCM$^+$16, Section 4.3], this is just a linear coordinate space in cumulant coordinates:

$$\begin{aligned} \nu_4(\mathbb{P}^2) \cap \mathbb{A}^{14} &= V(k_{20}, k_{11}, k_{02}, k_{30}, k_{21}, k_{12}, k_{03}, k_{40}, k_{31}, k_{22}, k_{13}, k_{04}) \\ &= V(k_{20}, k_{11}, k_{02}) \cap \mathcal{G}_{2,4}. \end{aligned}$$

The secant variety $\mathrm{Sec}_2(\nu_4(\mathbb{P}^2))$ comprises all ternary quartics of tensor rank ≤ 2. It has dimension 5 and degree 75 in $\mathbb{P}^{14}$, and its homogeneous prime ideal is minimally generated by 148 cubics, namely the 3×3-minors of the 6×6 Hankel matrix in (6.2.1). Also this ideal becomes much simpler when passing from moments to cumulant coordinates. Here, the ideal of $\mathrm{Sec}_2(\nu_4(\mathbb{P}^2)) \cap \mathbb{A}^{14}$ is generated by 36 binomial quadrics, like $k_{31}^2 - k_{22}k_{40}$ and $k_{30}k_{31} - k_{21}k_{40}$, along with seven trinomial cubics like $2k_{20}^3 - k_{30}^2 + k_{20}k_{40}$ and $2k_{11}k_{20}^2 - k_{21}k_{30} + k_{11}k_{40}$.

Example 6.2.2. The hypersurface in $\mathbb{P}^6$ described in Theorem 5.0.4 contains a familiar threefold, namely the determinantal variety $\mathrm{Sec}_2(\nu_6(\mathbb{P}^1))$ defined by the 3×3-minors of the 4×4-Hankel matrix

$$\begin{pmatrix} m_0 & m_1 & m_2 & m_3 \\ m_1 & m_2 & m_3 & m_4 \\ m_2 & m_3 & m_4 & m_5 \\ m_3 & m_4 & m_5 & m_6 \end{pmatrix}. \tag{6.2.2}$$

This can be seen by setting $\sigma = \tau = 0$ in the parametrization (4.2.1). Indeed, if the variances tend to zero then the Gaussian mixture converges to a mixture of

the point distributions, supported at the means μ and ν. The first $d+1$ moments of point distributions form the rational normal curve in $\mathbb{P}^d$, consisting of Hankel matrices of rank 1. Their kth mixtures specify a secant variety of the rational normal curve, consisting of Hankel matrices of rank k.

We now present the main result of this chapter. It is a complete classification of algebraic identifiability for homoscedastic Gaussian mixtures with known variance.

Theorem 6.2.3. *Consider the statistical model given by the homoscedastic mixture of $k > 1$ Gaussians in $\mathbb{R}^n$ with known covariance Σ. Let $d > 1$ such that the number of moments is at least the number of parameters: $\binom{n+d}{d} \geq (n+1)k$. Then the model is always algebraically identifiable from the moments up to order d, except in the following cases:*

- $d = 2$
- $d = 3$, $n = 4$, $k = 7$
- $d = 4$, $n = 2$, $k = 5$
- $d = 4$, $n = 4$, $k = 14$

Remark 6.2.4. For the homoscedastic (isotropic) case of Theorem 6.1.5, as it appears in [AGH$^+$14, Theorem 3.2], Theorem 6.2.3 applies as soon as the variance is estimated from the covariance matrix (that is, from the moments of order $d = 1$ and $d = 2$). Since we have in this case $d = 3$ and $k \leq n$, no exceptions apply and we can confirm algebraic identifiability for these submodels.

In order to prove Theorem 6.2.3, we first define the corresponding moment subvariety and then prove that it is actually isomorphic to the Veronese. That is, $\mathrm{Sec}_k(\mathcal{V}_{n,d})$ not only represents mixtures of point mass distributions but also homoscedastic Gaussian mixtures with a fixed covariance. Even if the mixture is not homoscedastic, as long as the covariances Σ_i are known (also a statistically relevant model), there is an algebraic object representing the mixture moments, namely the *join* of Veronese varieties.

Definition 6.2.5. Let $\mathcal{G}^S_{n,d} \subset \mathcal{G}_{n,d} \subset \mathbb{P}^N$ the homoscedastic Gaussian moment variety with fixed covariance $S \succ 0$. We say that the corresponding secant moment variety $\mathrm{Sec}_k(\mathcal{G}^S_{n,d}) \subset \mathrm{Sec}_k(\mathcal{G}_{n,d})$ is *algebraically identifiable* when the map from the model parameters $\{\alpha_i, \mu_i\}_{i=1,\dots,n}$ to the moments up to order d is generically finite-to-one.

Proposition 6.2.6. *The moment variety $\mathcal{G}^S_{n,d}$ is isomorphic to $\mathcal{V}_{n,d}$ under a linear change of coordinates in $\mathbb{P}^N$. In particular,* $\dim \mathrm{Sec}_k(\mathcal{G}^S_{n,d}) = \dim \mathrm{Sec}_k(\mathcal{V}_{n,d})$, *for all $n, d, k \geq 1$.*

Proof. The result follows by inspecting the moment generating function (4.1.1):

$$\sum_{i_1,i_2,\ldots,i_n\geq 0} \frac{m_{i_1 i_2 \cdots i_n}}{i_1! i_2! \cdots i_n!} t_1^{i_1} t_2^{i_2} \cdots t_n^{i_n} \;=\; \exp\Big(\mu_1 t_1 + \mu_2 t_2 + \ldots + \mu_n t_n\Big) \cdot \exp\left(\frac{1}{2}\sum_{i,j=1}^{n} s_{ij} t_i t_j\right)$$

All s_{ij} are constants so we may multiply the moment generating function by $\exp\left(-\frac{1}{2}\sum_{i,j=1}^{n} s_{ij} t_i t_j\right)$. This will cause a linear change of coordinates in the $m_{i_1 i_2 \cdots i_n}$ on the left hand side, while the expansion of the right hand side becomes precisely the Veronese embedding. □

The heart behind Theorem 6.2.3 is the Alexander-Hirschowitz Theorem [AH00].

Theorem 6.2.7 (Alexander-Hirschowitz)**.** *[BO08, Theorem 1.2] The secant variety* $\mathrm{Sec}_k(\mathcal{V}_{n,d})$ *has the expected dimension* $\min\left\{k(n+1), \binom{n+d}{d}\right\} - 1$ *for all values of* $k, n, d \geq 1$ *except for the following cases:*

- $d = 2,\ 2 \leq k \leq n$
- $n = 2,\ d = 4,\ k = 5$
- $n = 3,\ d = 4,\ k = 9$
- $n = 4,\ d = 3,\ k = 7$
- $n = 4,\ d = 4,\ k = 14$

Proof. (of Theorem 6.2.3) By Proposition 6.2.6, considering defectivity of $\mathcal{G}_{n,d}^S$ is equivalent to considering defectivity of $\mathcal{V}_{n,d}$. Given that we have more moments that parameters by our choice of d, algebraic identifiability will fail if and only if the secant variety $\mathrm{Sec}_k(\mathcal{V}_{n,d})$ is defective. We now go through the possible cases:

- All cases for $d = 2$ are defective. Indeed, since $(n+1)k \leq \binom{n+2}{2}$ we must have $k \leq \frac{n+2}{2} \leq n$ for all $n > 1$ (and $n = 1$ is impossible since $k > 1$), so the first sequence of exceptions applies.
- For $n = 2, d = 4, k = 5$ we have $\binom{n+d}{d} = \binom{2+4}{4} = 15 = 5(2+1) = k(n+1)$ so the exception applies.
- For $n = 3, d = 4, k = 9$ we have $\binom{n+d}{d} = \binom{3+4}{4} = 35 < 36 = 9(3+1) = k(n+1)$ so we do *not* consider this exception.
- For $n = 4, d = 3, k = 7$ we have $\binom{n+d}{d} = \binom{4+3}{3} = 35 = 7(4+1) = k(n+1)$ so the exception applies.
- Finally, for $n = 4, d = 4, k = 14$ we have $\binom{n+d}{d} = \binom{4+4}{4} = 70 = 14(4+1) = k(n+1)$ so the last exception applies. □

We end this section with a couple of explorations of the centered submodel, that is, equal means. In that case we get the secant varieties of varieties of powers of quadratic forms.

Proposition 6.2.8. *The equal-means submodel of* $\mathrm{Sec}_2(\mathcal{G}_{2,3}) = \mathbb{P}^9$ *has dimension* 5 *and degree* 16. *It is identical to the Gaussian moment variety* $\mathcal{G}_{2,3}$ *in Proposition 4.4.7 so the mixtures add nothing new in* $\mathbb{P}^9$. *The equal-variances submodel of* $\mathrm{Sec}_2(\mathcal{G}_{2,3})$ *has dimension* 7 *and degree* 15 *in* $\mathbb{P}^9$. *Its ideal is Cohen-Macaulay and is generated by the maximal minors of the* 6×5*-matrix*

$$\begin{pmatrix} 0 & 0 & m_{00} & m_{10} & m_{01} \\ 0 & m_{10} & m_{20} & m_{30} & m_{21} \\ m_{01} & 0 & m_{02} & m_{12} & m_{03} \\ 0 & m_{00} & 2m_{10} & 2m_{20} & 2m_{11} \\ m_{00} & 0 & 2m_{01} & 2m_{11} & 2m_{02} \\ m_{10} & m_{01} & 2m_{11} & 2m_{21} & 2m_{12} \end{pmatrix}. \tag{6.2.3}$$

Since the number of parameters for these submodels are 9 *and* 8 *respectively, both of these models are not identifiable.*

This proposition is proved by a direct computation. That the equal-means submodel of $\mathrm{Sec}_2(\mathcal{G}_{2,3})$ equals $\mathcal{G}_{2,3}$ is not so surprising, since the parametrization of the latter is linear in the variance parameters s_{11}, s_{12}, s_{22}. This holds for all moments up to order 3. The same is no longer true for $d \geq 4$. On the other hand, it was gratifying to see an occurrence, in the matrix (6.2.3), of the *Hilbert-Burch Theorem* for Cohen-Macaulay ideals of codimension 2.

Example 6.2.9. ([AFS16, Example 21]) The following concrete example was worked out with some input from Giorgio Ottaviani. Consider the mixture of two bivariate Gaussians that are centered at the origin. This model has 7 parameters: there is one mixture parameter, and each Gaussian has a 2×2 covariance matrix, with three unknown entries. We consider the variety $\mathcal{V}$ that is parametrized by all moments of order exactly $d = 6$. This variety has only dimension 5. It lives in the $\mathbb{P}^6$ with coordinates $m_{06}, m_{15}, \ldots, m_{60}$. This hypersurface has degree 15. Its points are the binary octics that are sums of the third powers of two binary quadrics. Thus, this is the secant variety of a linear projection of the third Veronese surface from $\mathbb{P}^9$ to $\mathbb{P}^6$.

The polynomial that defines $\mathcal{V}$ has 1370 monomials of degree 15 in the seven unknowns $m_{06}, m_{15}, \ldots, m_{60}$. In fact, this is the unique (up to scaling) invariant of binary sextics of degree 15. It is denoted I_{15} in Faa di Bruno's book [dB76, Table IV^{10}], where a determinantal formula was given. A quick way to compute $\mathcal{V}$ by elimination is as follows. Start with the variety $\mathrm{Sec}_2(\nu_3(\mathbb{P}^2))$ of symmetric $3 \times 3 \times 3$-tensors of rank ≤ 2. This is defined by the maximal minors of a Hankel matrix

of size 3×6. It has degree 15 and dimension 5 in $\mathbb{P}^9$. Now project into $\mathbb{P}^6$. This projection has no base points, so the image is a hypersurface of degree 15.

Example 6.2.10. ([AFS16, Example 22]) Consider $n = k = 2$ mixtures of two bivariate Gaussians that are centered at zero. As in [GHK15] let us consider the moments of orders $d = 3, 4, 6$. The odd order moment $d = 3$ is zero, but let us examine the corresponding variety for $d = 4$ and $d = 6$. This lives in the $\mathbb{P}^{12}$ with coordinates

$$m_{00}, m_{40}, m_{31}, m_{22}, m_{13}, m_{04}, m_{60}, m_{51}, m_{42}, m_{33}, m_{24}, m_{15}, m_{06}.$$

We start with the variety X that is parametrized by the 4th and 6th powers of binary quadrics. This variety has dimension three and degree 27 in $\mathbb{P}^{12}$. We are interested in the secant variety $\mathrm{Sec}_2(X)$. This secant variety has the expected dimension 7, so the model is algebraically identifiable. We do not know whether $\mathrm{Sec}_2(X)$ is rationally identifiable. A relation of lowest degree is the following quartic:

$$\begin{aligned}
&6m_{15}m_{22}m_{31}^2 - 10m_{13}m_{24}m_{31}^2 - 2m_{06}m_{31}^3 + 10m_{04}m_{31}^2m_{33} - 9m_{15}m_{22}^2m_{40} + 15m_{13}m_{22}m_{24}m_{40} \\
&\quad +2m_{13}m_{15}m_{31}m_{40} + 3m_{06}m_{22}m_{31}m_{40} - 5m_{04}m_{24}m_{31}m_{40} - 10m_{13}^2m_{33}m_{40} - m_{06}m_{13}m_{40}^2 \\
&\quad +m_{04}m_{15}m_{40}^2 + 10m_{13}^2m_{31}m_{42} - 15m_{04}m_{22}m_{31}m_{42} + 5m_{04}m_{13}m_{40}m_{42} - 6m_{13}^2m_{22}m_{51} \\
&+9m_{04}m_{22}^2m_{51} - 2m_{04}m_{13}m_{31}m_{51} - m_{04}^2m_{40}m_{51} + 2m_{13}^3m_{60} - 3m_{04}m_{13}m_{22}m_{60} + m_{04}^2m_{31}m_{60}
\end{aligned}$$

□

The takeaway from the last examples should be that there are many possibilities to explore. One can fix a particular order of moments, as in Example 6.2.9 or one can take moments of two orders as in Example 6.2.10. Another instance could be moments of order $d = 2$ and $d = 3$ as considered in [HK13]. It would be interesting to determine the algebraic relations for general restricted sets of moments. Geometrically, even for small values of k we would expect to obtain interesting varieties.

7. Conclusion

Throughout this thesis we presented Gaussian mixture models and parametric inference for them from an algebraic perspective, and successfully obtained relevant results tying back to the statistical applications by applying algebraic techniques.

On the maximum number of modes that a mixture of k Gaussians in $\mathbb{R}^n$ can achieve, we proved a lower bound (Theorem 2.3.1) that matches for $n = 2$ a conjecture by Bernd Sturmfels (Conjecture 2.2.1). We also gave the first upper bound (assuming finiteness of the modes) in terms of n, k (Theorem 2.4.4), thus answering a question first posed in [RL05].

On the maximum likelihood approach to estimate the Gaussian mixture parameters from data, we proved that these functions are transcendental (Theorem 3.1.2), therefore surpassing a finite algebraic complexity. As a consequence, there is no notion of ML degree and furthermore there is no bound on the number of critical points that the log-likelihood of a Gaussian mixture can have (Theorem 3.3.1).

We revisited Pearson's method of moments and asked in general about identifiability of parameters from a set of moments up to order d. We distinguished between algebraic identifiability and rational identifiability and proceeded to successfully apply algebraic techniques to shed light on this problem. Indeed, we showed that the first $3k - 1$ moments suffice to algebraically identify a mixture of univariate Gaussians (Theorem 5.0.6). This was possible thanks to the language of secant varieties of moment varieties and the concept of nondefectivity. Further, we saw that for $k = 3$ the corresponding identifiability degree to Pearson's 9 is 225, and conjectured how these grow with k (Conjecture 5.3.3).

For higher dimensional Gaussians, we warned that there are cases where method of moments will fail to recover the parameters from the expected number of moments (Theorem 5.0.6). These cases are linked to defective secant varieties and we explored all the small instances of defectivity (Tables 5.2.1 and 5.2.2).

The learning problem of Gaussian mixtures and their submodels remains an active research topic in Computer Science and Machine Learning. We drew some parallels between results in these areas and the results presented here (Remarks 6.1.4 and 6.2.4). Further, we applied the celebrated Alexander-Hirschowitz Theorem of Algebraic Geometry to solve the algebraic identifiability problem for homoscedastic Gaussian mixtures with known variance (Theorem 6.2.3).

In summary, the study of moments of mixtures of Gaussians leads to many

interesting projective varieties, whose geometry is still largely unexplored. We believe there is still plenty of fertile ground for investigations by algebraic geometers. On the statistical side, it is most interesting to understand the fibers of the natural parameterization of the variety $\mathrm{Sec}_k(\mathcal{G}_{n,d})$, which includes the method of moments. Problem 3 (cf. Section 1.3) serves as a guiding question. In the case of algebraic identifiability, we are always interested in finding the algebraic degree of the parametrization, and in effective methods for solving for the model parameters.

We hope that by adding mixtures of Gaussians to the models studied in Algebraic Statistics, the interest for both these models and this area will spread even more and further connections will be developed.

Bibliography

[A+97] Aldrich, John et al.: *R.A. Fisher and the making of maximum likelihood 1912-1922.* Statistical Science, 12(3):162–176, 1997.

[ABB+17] Améndola, Carlos, Nathan Bliss, Isaac Burke, Courtney R Gibbons, Martin Helmer, Serkan Hoşten, Evan D Nash, Jose Israel Rodriguez and Daniel Smolkin: *The maximum likelihood degree of toric varieties.* arXiv preprint arXiv:1703.02251, 2017.

[Ådl87] Ådlandsvik, Bjørn: *Joins and higher secant varieties.* Mathematica Scandinavica, 61:213–222, 1987.

[Ådl88] Ådlandsvik, Bjørn: *Varieties with an extremal number of degenerate higher secant varieties.* Journal für die reine und angewandte Mathematik, 392:16–26, 1988.

[ADS15] Améndola, Carlos, Mathias Drton and Bernd Sturmfels: *Maximum likelihood estimates for Gaussian mixtures are transcendental.* In *International Conference on Mathematical Aspects of Computer and Information Sciences*, pages 579–590. Springer, 2015.

[AEH17] Améndola, Carlos, Alexander Engström and Christian Haase: *Maximum number of modes of Gaussian mixtures.* arXiv preprint arXiv:1702.05066, 2017.

[AFS16] Améndola, Carlos, Jean-Charles Faugère and Bernd Sturmfels: *Moment varieties of Gaussian mixtures.* Journal of Algebraic Statistics, 7:14–28, 2016.

[AGH+14] Anandkumar, Animashree, Rong Ge, Daniel J Hsu, Sham M Kakade and Matus Telgarsky: *Tensor decompositions for learning latent variable models.* Journal of Machine Learning Research, 15(1):2773–2832, 2014.

[AH00] Alexander, James and André Hirschowitz: *An asymptotic vanishing theorem for generic unions of multiple points.* Inventiones Mathematicae, 140(2):303–325, 2000.

[AHR13] Alexandrovich, Grigory, Hajo Holzmann and Surajit Ray: *On the number of modes of finite mixtures of elliptical distributions.* In *Algorithms from and for Nature and Life*, pages 49–57. Springer, 2013.

[AR16] Améndola, Carlos and Jose Israel Rodriguez: *Solving parameterized polynomial systems with decomposable projections.* arXiv preprint arXiv:1612.08807, 2016.

[ARS17] Améndola, Carlos, Kristian Ranestad and Bernd Sturmfels: *Algebraic identifiability of Gaussian mixtures.* International Mathematics Research Notices, 2017.

[ASS17] Abo, Hirotachi, Anna Seigal and Bernd Sturmfels: *Eigenconfigurations of tensors.* Algebraic and Geometric Methods in Discrete Mathematics, 685:1, 2017.

[Bak90] Baker, Alan: *Transcendental number theory.* Cambridge University Press, 1990.

[BCHY09] Benaglia, Tatiana, Didier Chauveau, David Hunter and Derek Young: *mixtools: An R package for analyzing finite mixture models.* Journal of Statistical Software, 32(6):1–29, 2009.

[BCR13] Bochnak, Jacek, Michel Coste and Marie-Françoise Roy: *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[BCS13] Bürgisser, Peter, Michael Clausen and Amin Shokrollahi: *Algebraic complexity theory*, volume 315. Springer Science & Business Media, 2013.

[BDHR15] Boralevi, Ada, Jan Draisma, Emil Horobet and Elina Robeva: *Orthogonal and unitary tensor decomposition from an algebraic perspective.* arXiv:1512.08031, to appear in Israel Journal of Mathematics, 2015.

[Beh70] Behboodian, Javad: *On the modes of a mixture of two normal distributions.* Technometrics, 12(1):131–139, 1970.

[BHSPR07] Buot, Max-Louis G, Serkan Hoşten and Donald St. P. Richards: *Counting and locating the solutions of polynomial systems of maximum likelihood equations, II: The Behrens-Fisher problem.* Statistica Sinica, pages 1343–1354, 2007.

[BHSW06] Bates, D.J., J.D. Hauenstein, A.J. Sommese and C.W. Wampler: *Bertini: Software for numerical algebraic geometry*. Available at `bertini.nd.edu` with permanent doi: dx.doi.org/10.7274/R0H41PB5, 2006.

[Bis06] Bishop, Christopher M: *Pattern recognition and machine learning*. Springer, 2006.

[BO08] Brambilla, Maria Chiara and Giorgio Ottaviani: *On the Alexander–Hirschowitz theorem*. Journal of Pure and Applied Algebra, 212(5):1229–1251, 2008.

[BR10] Bhattacharya, Rabi N and R Ranga Rao: *Normal approximation and asymptotic expansions*. SIAM, 2010.

[BS10] Belkin, Mikhail and Kaushik Sinha: *Polynomial learning of distribution families*. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[BS15] Belkin, Mikhail and Kaushik Sinha: *Polynomial learning of distribution families*. SIAM Journal on Computing, 44(4):889–911, 2015.

[BV06] Bruns, Winfried and Udo Vetter: *Determinantal rings*, volume 1327. Springer, 2006.

[CC02] Chiantini, Luca and Ciro Ciliberto: *Weakly defective varieties*. Transactions of the American Mathematical Society, 354(1):151–178, 2002.

[CCK+06] Chang, Ee-Chien, Sung Woo Choi, Do Yong Kwon, Hyungju Park and Chee K Yap: *Shortest path amidst disc obstacles is computable*. International Journal of Computational Geometry & Applications, 16(05n06):567–590, 2006.

[CCM+16] Ciliberto, Ciro, Maria Angelica Cueto, Massimiliano Mella, Kristian Ranestad and Piotr Zwiernik: *Cremona linearizations of some classical varieties*. In *From Classical to Modern Algebraic Geometry*, pages 375–407. Springer, 2016.

[CHKS06] Catanese, Fabrizio, Serkan Hoşten, Amit Khetan and Bernd Sturmfels: *The maximum likelihood degree*. American Journal of Mathematics, 128(3):671–697, 2006.

[CK16] Chen, Justin and Joe Kileel: *Numerical implicitization for Macaulay2*. arXiv preprint arXiv:1610.03034, 2016.

[CMM12] Chen, WC, R Maitra and V Melnykov: *EMCluster: EM Algorithm for model-based clustering of finite mixture Gaussian distribution*. R Package, URL http://cran. rproject. org/package= EMCluster, 2012.

[CPPY06] Choi, Sung Woo, Sung-il Pae, Hyungju Park and Chee K Yap: *Decidability of collision between a helical motion and an algebraic motion*. page 69, 2006.

[CPW03a] Carreira-Perpinán, MÁ and CKI Williams: *An isotropic Gaussian mixture can have more modes than components*. Institute for Adaptive and Neural Computation, 2003.

[CPW03b] Carreira-Perpinan, Miguel and Christopher Williams: *On the number of modes of a Gaussian mixture*. In *Scale Space Methods in Computer Vision*, pages 625–640. Springer, 2003.

[Das99] Dasgupta, Sanjoy: *Learning mixtures of Gaussians*. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[dB76] Bruno, Francesco Fàa di: *Thèorie des formes binaires*. Brero, 1876.

[DLR77] Dempster, Arthur P, Nan M Laird and Donald B Rubin: *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B (methodological), pages 1–38, 1977.

[DS+98] Diaconis, Persi, Bernd Sturmfels et al.: *Algebraic algorithms for sampling from conditional distributions*. The Annals of Statistics, 26(1):363–397, 1998.

[DSS08] Drton, Mathias, Bernd Sturmfels and Seth Sullivant: *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

[EFR13] Edelsbrunner, Herbert, Brittany Terese Fasy and Günter Rote: *Add isotropic Gaussian kernels at own risk: more and more resilient modes in higher dimensions*. Discrete & Computational Geometry, 49(4):797–822, 2013.

[Eis64] Eisenberger, Isidore: *Genesis of bimodal distributions*. Technometrics, 6(4):357–363, 1964.

[Eis13] Eisenbud, David: *Commutative algebra: with a view toward algebraic geometry*, volume 150. Springer Science & Business Media, 2013.

[FR03] Fraley, Chris and Adrian E Raftery: *Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST*. Journal of Classification, 20(2):263–286, 2003.

[Ful13] Fulton, William: *Intersection theory*, volume 2. Springer Science & Business Media, 2013.

[GBW16] Gandhi, Kavish and Yonah Borns-Weil: *Moment-based learning of mixture distributions*. MIT Summer Program for Undergraduate Research (SPUR), 2016.

[GDP+12] Gross, Elizabeth, Mathias Drton, Sonja Petrović et al.: *Maximum likelihood degree of variance component models*. Electronic Journal of Statistics, 6:993–1016, 2012.

[Gel15] Gelfond, Aleksandr Osipovich: *Transcendental and algebraic numbers*. Courier Dover Publications, 2015.

[Gha15] Ghassabeh, Youness Aliyari: *A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel*. Journal of Multivariate Analysis, 135:1–10, 2015.

[GHK15] Ge, Rong, Qingqing Huang and Sham M Kakade: *Learning mixtures of Gaussians in high dimensions*. In *Proceedings of the forty-seventh annual ACM Symposium on Theory of Computing*, pages 761–770. ACM, 2015.

[GMS+06] Geiger, Dan, Christopher Meek, Bernd Sturmfels et al.: *On the toric algebra of graphical models*. The Annals of Statistics, 34(3):1463–1492, 2006.

[GS98] Gupta, Lalit and Thotsapon Sortrakul: *A Gaussian-mixture-based image segmentation algorithm*. Pattern Recognition, 31(3):315–325, 1998.

[GS02] Grayson, Daniel R and Michael E Stillman: *Macaulay 2, a software system for research in algebraic geometry*, 2002.

[Her05] Herbrich, Ralf: *Minimising the Kullback–Leibler divergence*. Microsoft, Tech. Rep., 2005.

[HG84] Hummel, Robert A and Basilis C Gidas: *Zero crossings and the heat equation*. New York University. Courant Institute of Mathematical Sciences. Computer Science Department, 1984.

[HK13] Hsu, Daniel and Sham M Kakade: *Learning mixtures of spherical Gaussians: moment methods and spectral decompositions*. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[HMMR15] Hennig, Christian, Marina Meila, Fionn Murtagh and Roberto Rocci: *Handbook of cluster analysis*. CRC Press, 2015.

[HP15] Hardt, Moritz and Eric Price: *Tight bounds for learning a mixture of two Gaussians*. In *Proceedings of the forty-seventh annual ACM symposium on the theory of computing*, pages 753–760. ACM, 2015.

[HS14] Huh, June and Bernd Sturmfels: *Likelihood geometry*. In *Combinatorial algebraic geometry*, pages 63–117. Springer, 2014.

[JZB+16] Jin, Chi, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright and Michael I Jordan: *Local maxima in the likelihood of Gaussian mixture models: structural results and algorithmic consequences*. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

[Kho91] Khovanskii, A.G.: *Fewnomials*, volume 88. American Mathematical Soc., 1991.

[KMV10] Kalai, Adam Tauman, Ankur Moitra and Gregory Valiant: *Efficiently learning mixtures of two Gaussians*. In *Proceedings of the forty-second ACM Symposium on the Theory of Computing*, pages 553–562. ACM, 2010.

[KMV12] Kalai, Adam Tauman, Ankur Moitra and Gregory Valiant: *Disentangling gaussians*. Communications of the ACM, 55(2):113–120, 2012.

[Laz04] Lazard, Daniel: *Injectivity of real rational mappings: the case of a mixture of two Gaussian laws*. Mathematics and Computers in Simulation, 67(1):67–84, 2004.

[LB87] Le Barz, P: *Formules pour les trisécantes des surfaces algébriques*. L'Ens. Math, 33(198):1–66, 1987.

[MP04] McLachlan, Geoffrey and David Peel: *Finite mixture models*. John Wiley & Sons, 2004.

[MSS65] Milnor, John Willard, L Siebenmann and J Sondow: *Lectures on the h-cobordism theorem*, volume 963. Princeton University Press Princeton, NJ, 1965.

[MSUZ16] Michałek, Mateusz, Bernd Sturmfels, Caroline Uhler and Piotr Zwiernik: *Exponential varieties*. Proceedings of the London Mathematical Society, 112(1):27–56, 2016.

[MV10] Moitra, Ankur and Gregory Valiant: *Settling the polynomial learnability of mixtures of gaussians*. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[Pea94] Pearson, Karl: *Contributions to the mathematical theory of evolution*. Philosophical Transactions of the Royal Society of London. A, 185:71–110, 1894.

[PP+08] Petersen, Kaare Brandt, Michael Syskind Pedersen et al.: *The matrix cookbook*. Technical University of Denmark, 7:15, 2008.

[PRW00] Pistone, Giovanni, Eva Riccomagno and Henry P Wynn: *Algebraic statistics: Computational commutative algebra in statistics*. CRC Press, 2000.

[PS05] Pachter, Lior and Bernd Sturmfels: *Algebraic statistics for computational biology*, volume 13. Cambridge university press, 2005.

[Ree85] Reeds, James A: *Asymptotic number of roots of Cauchy location likelihood equations*. The Annals of Statistics, pages 775–784, 1985.

[RL05] Ray, Surajit and Bruce G Lindsay: *The topography of multivariate normal mixtures*. Annals of Statistics, pages 2042–2065, 2005.

[Rob16] Robeva, Elina: *Orthogonal decomposition of symmetric tensors*. SIAM Journal on Matrix Analysis and Applications, 37(1):86–102, 2016.

[RR95] Reynolds, Douglas A and Richard C Rose: *Robust text-independent speaker identification using Gaussian mixture speaker models*. IEEE transactions on Speech and Audio Processing, 3(1):72–83, 1995.

[RR12] Ray, Surajit and Dan Ren: *On the upper bound of the number of modes of a multivariate normal mixture.* Journal of Multivariate Analysis, 108:41–52, 2012.

[RW84] Redner, Richard A and Homer F Walker: *Mixture densities, maximum likelihood and the EM algorithm.* SIAM review, 26(2):195–239, 1984.

[Sot11] Sottile, Frank: *Real solutions to equations from geometry*, volume 57. American Mathematical Society Providence, RI, 2011.

[Sre07] Srebro, Nathan: *Are there local maxima in the infinite-sample likelihood of Gaussian mixture estimation?* Lecture Notes in Computer Science, 4539:628, 2007.

[ST04] Spielman, Daniel A and Shang-Hua Teng: *Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time.* Journal of the ACM (JACM), 51(3):385–463, 2004.

[Ste11] Steele, Russell: *organized by Russell Steele, Bernd Sturmfels, and Sumio Watanabe Workshop Summary.* 2011.

[Stu16] Sturmfels, Bernd: *Tensors and their eigenvectors.* Notices of the AMS, 63(6), 2016.

[SU10] Sturmfels, Bernd and Caroline Uhler: *Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry.* Annals of the Institute of Statistical Mathematics, 62(4):603–638, 2010.

[SVW96] Sommese, Andrew J, Jan Verschelde and Charles W Wampler: *Numerical algebraic geometry.* In *The Mathematics of Numerical Analysis, volume 32 of Lectures in Applied Mathematics.* Citeseer, 1996.

[Tei63] Teicher, Henry: *Identifiability of finite mixtures.* The annals of Mathematical statistics, pages 1265–1269, 1963.

[U+12] Uhler, Caroline et al.: *Geometry of maximum likelihood estimation in Gaussian graphical models.* The Annals of Statistics, 40(1):238–261, 2012.

[VdV00] Vaart, Aad W Van der: *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Wal13] Wallace, Benjamin: *On the critical points of Gaussian mixtures.* Master's thesis, Queen's University, Ontario, Canada, 2013.

[Wat04] Watanabe, Sumio: *Kullback information of normal mixture is not an analytic function.* IEICE Technical Report, 2004:41–46, 2004.

[Wat09] Watanabe, Sumio: *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009.

[WCL15] Wang, Sida, Arun Tejasvi Chaganty and Percy S Liang: *Estimating mixture models via mixtures of polynomials.* In *Advances in Neural Information Processing Systems*, pages 487–495, 2015.

[Wil05] Willink, R: *Normal moments and Hermite polynomials.* Statistics & Probability Letters, 73(3):271–275, 2005.

[Wu83] Wu, CF Jeff: *On the convergence properties of the EM algorithm.* The Annals of statistics, pages 95–103, 1983.

[YS68] Yakowitz, Sidney J and John D Spragins: *On the identifiability of finite mixtures.* The Annals of Mathematical Statistics, pages 209–214, 1968.

[Zwi15] Zwiernik, Piotr: *Semialgebraic statistics and latent tree models.* CRC Press, 2015.

List of Figures

List of Tables

Appendix

The following pages contain computer code in the language `R`, which was used in some chapters of this thesis. They explore several aspects of univariate Gaussian mixtures.

- In Appendix A we give a short note about correct sampling from Gaussian mixtures and present basic code for achieving this.
- In Appendix B we present an implementation of the EM algorithm for Gaussian mixtures.
- In Appendix C we provide a general Pearson's method of moments implementation for fitting a mixture of two univariate Gaussians to data.

A. Sampling

We present a basic script to sample from a Gaussian mixture given as input its parameters and the sample size.

One important pitfall to avoid is that to sample from

$$f_X = \alpha_1 f_{X_1} + \ldots + \alpha_k f_{X_k}$$

where $X_i \sim N(\mu_i, \Sigma_i)$ are the Gaussian components, it does **not** suffice to sample from each X_i individually and then setting $X = \alpha_1 X_1 + \ldots + \alpha_k X_k$. This is wrong and would instead create a Gaussian distribution again (Gaussian convolution).

Instead, the right way of sampling from a mixture is to sample first from a discrete distribution with probabilities given by the vector of α. This will indicate the component X_i and now we can sample from this Gaussian. In other words, one samples from X_i with probability α_i.

The function below, 'sampleMixt.R' relies on `R`'s function `rnorm` that samples from a single univariate Gaussian. It also plots the corresponding data density.

```
sampleMixt  <- function(N,means,variances,alphas){
  k = length(means);
  class <- sample(1:k,prob=alphas,size=N,replace=TRUE);
  data <- rnorm(n=N,mean=means[class],sd=sqrt(variances[class]));
  plot(density(data),main="Gaussian mixture density")
  return(data);
}
```

If we want to sample from a multivariate Gaussian mixture, `rnorm` does not work. However, there are `R` packages that will have the corresponding multivariate version. Indeed, we can use the function '`mvrnorm`' from the package `MASS` or the function '`rmvnorm`' from the package `mvtnorm`.

B. EM Algorithm

The main function is 'EMalgorithm.R', which receives as input a vector of data and parameter starting values. It calls 'Estep.R' and 'Mstep.R'. The former is the Expectation step, which takes the input and assigns a probability to each data point. Then the latter, being the Maximization step, takes this probability vector and computes the maximum likelihood estimates for the parameters, updating their values. The process repeats until convergence.

```
EMalgorithm <- function(x,params){
  TOL = 1e-8;
  iter = 1;
  ITMAX = 100;
  change = 1;
  while(change > TOL && iter<ITMAX){
    #print(iter)
    iter = iter + 1;
    result = Estep(x,params);
    probs = result$probs;
    loglik = result$loglik;
    #print(probs)
    #print(loglik)
    paramsnew = Mstep(x,probs);
    #print(params)
    change = sqrt(sum((paramsnew - params)^2));
    params = paramsnew;
  }
  #print(probs)
  return(list(params = params, probs = probs, loglik = loglik))
}
```

```
Estep <- function(x,params){
  alpha=params[1];
  mu1=params[2];
  mu2=params[3];
  sigma1=params[4];
```

```
  sigma2=params[5];
  probs = alpha*dnorm(x,mu1,sigma1)/(alpha*dnorm(x,mu1,sigma1)
                                + (1-alpha)*dnorm(x,mu2,sigma2));
  loglik = sum(log(alpha*dnorm(x,mu1,sigma1)
                               + (1-alpha)*dnorm(x,mu2,sigma2)));
  return(list(probs = probs, loglik = loglik))
}
```

```
Mstep <- function(x,probs){
  N=length(x);
  N1=sum(probs);
  N2=N-N1;
  alpha = N1/N;
  mu1 = sum(probs*t(x))/N1;
  mu2 = sum((1-probs)*t(x))/N2;
  sigma1 = sqrt(sum(probs*t((x-mu1)^2))/N1);
  sigma2 = sqrt(sum((1-probs)*t((x-mu2)^2))/N2);
  params = c(alpha,mu1,mu2,sigma1,sigma2)
  return(params)
}
```

C. Pearson's MOM

The main function is 'PearsonMOM.R', which receives as input a vector with the 6 (sample) moments m1,m2,m3,m4,m5,m6. Then it will apply Pearson's method of moments and produce the estimates. It calls the function 'toCumulants.R' (which just transforms moments into cumulants) and in some cases it also calls 'mixtmoms.R' which computes the moments for a mixture of two Gaussians with specified parameters (this is particularly useful for generating input for tests).

```
PearsonMOM <- function(moms){
  cums = toCumulants(moms);
  k1 = cums[1];
  k2 = cums[2];
  k3 = cums[3];
  k4 = cums[4];
  k5 = cums[5];
  k6 = cums[6];
  sols = 0;
  if (k3==0 & k5==0){
      if(k4==0){
        print(cat("Data resembles a single Gaussian with mean ",k1,"
                      and variance ",k2,". No honest mixture.","\n"));
        sols =1;
      }
      else {
        if(k4>0){
          print(cat("Data is consistent only with an equal means model."
                                                                , "\n"));}
        else{
          print(cat("Data displays symmetry, different means alternative
                                                 still explored.","\n"));}
        e2 = k2^2 - k4/3 + k2*k6/(5*k4);
        e1 = 2*k2 + k6/(5*k4);
        d = e1^2 - 4*e2;
        if ((e1>0 & e2>0) & d>0 ){
          v1 = (e1 - sqrt(d))/2;
          v2 = (e1 + sqrt(d))/2;
```

```
        a = (v2 - k2)/ (v2 - v1);
        print(cat("(alpha,mu1,mu2,sigma1,sigma2)=
                          ",c(a,k1,k1,sqrt(v1),sqrt(v2)),"\n"));
        sols = 1;
      }
      else {
        print(cat("Negative variance found, discarding equal means."
                                                          , "\n"));}
    }
  }
rootsp = polyroot(c(-8*k3^6,-32*k3^4*k4,-21*k3^2*k4^2-24*k3^3*k5,
           96*k3^4+9*k4^3-36*k3*k4*k5,148*k3^2*k4-6*k5^2,
           24*k3*k5+30*k4^2,12*k3^2,28*k4,0,8));
rootsord = rootsp[sort.list(abs(Im(rootsp)))];
print(cat("Pearson's polynomial roots: ", rootsord,"\n"));
rootsreal = subset(rootsord, abs(Im(rootsord))<0.01 & Re(rootsord)<0);
print(cat("Pearson's polynomial appears to have ",length(rootsreal),"
           negative real roots.", "\n"));
if (length(rootsreal)>0){
  p = Re(rootsreal);
  s = (2*k3^3+6*k3*k4*p+3*k5*p^2-8*k3*p^3)/(p*(4*k3^2+3*k4*p+2*p^3));
  m1 = (s-sqrt(s^2-4*p))/2;
  m2 = (s+sqrt(s^2-4*p))/2;
  R1 = p + k2;
  R2 = (k3/p + s)/3;
  var1 = R1 - m1*R2;
  var2 = R1 - m2*R2;
  sigma1 = sqrt(ifelse(var1>=0,var1,NA));
  sigma2 = sqrt(ifelse(var2>=0,var2,NA));
  alpha = m2 / (m2-m1);
  mu1 = m1 + k1;
  mu2 = m2 + k1;
  sixth = Inf*p;
  for (i in 1:length(rootsreal)){
    if (is.na(sigma1[i]) | is.na(sigma2[i])){
      print(cat("Negative variance found, removing root.", "\n"));}
    else{
      print(cat("(alpha,mu1,mu2,sigma1,sigma2)=
      ", c(alpha[i],mu1[i],mu2[i],sigma1[i],sigma2[i]),"\n"));
      sixth[i] = abs(mixtmoms(c(alpha[i],mu1[i],mu2[i],
                                sigma1[i],sigma2[i]))[6] - moms[6]);
```

```
        sols = sols + 1; }
    }
  }
  if(sols > 1){
    j = which.min(sixth);
    print(cat("Of the ",sols," statistically meaningful solutions, the
closest to the sample's sixth moment is",
      "(alpha,mu1,mu2,sigma1,sigma2)=
         ", c(alpha[j],mu1[j],mu2[j],sigma1[j],sigma2[j]), "\n"));
  } else{
    if(sols == 1){
      print("Unique statistically meaningful solution found.")}
        else {
          print("No solutions, the data does not come from a mixture
of two Gaussians.");}
    }
}
```

```
toCumulants <- function(moms){
  m1 = moms[1];
  m2 = moms[2];
  m3 = moms[3];
  m4 = moms[4];
  m5 = moms[5];
  m6 = moms[6];
  k1 = m1;
  k2 = m2 - m1^2;
  k3 = m3 -3*m1*m2 + 2*m1^3;
  k4 = m4 - 4*m1*m3 - 3*m2^2 + 12*m1^2*m2 - 6*m1^4;
  k5 = m5 -5*m1*m4 -10*m2*m3 + 20*m1^2*m3 + 30*m1*m2^2 - 60*m1^3*m2
          + 24*m1^5;
  k6 = m6 - 6*m1*m5 - 15*m2*m4 + 30*m1^2*m4 - 10*m3^2 + 120*m1*m2*m3
          -120*m1^3*m3 +30*m2^3 -270*m1^2*m2^2 + 360*m1^4*m2 -120*m1^6;
  cums = c(k1,k2,k3,k4,k5,k6);
  return(cums)
}
```

```
mixtmoms <- function(params){
  alpha = params[1];
  mu1 = params[2];
```

```
  mu2 = params[3];
  sigma1 = params[4];
  sigma2 = params[5];
  m1 = alpha*mu1 + (1-alpha)*mu2;
  m2 = alpha*(mu1^2 + sigma1^2) + (1-alpha)*(mu2^2 + sigma2^2);
  m3 = alpha*(mu1^3 + 3*mu1*sigma1^2)
             + (1-alpha)*(mu2^3 + 3*mu2*sigma2^2);
  m4 = alpha*(mu1^4 + 6*mu1^2*sigma1^2 + 3*sigma1^4)
             + (1-alpha)*(mu2^4 + 6*mu2^2*sigma2^2 + 3*sigma2^4);
  m5 = alpha*(mu1^5 + 10*mu1^3*sigma1^2 + 15*mu1*sigma1^4)
             + (1-alpha)*(mu2^5 + 10*mu2^3*sigma2^2 + 15*mu2*sigma2^4);
  m6 = alpha*(mu1^6 + 15*mu1^4*sigma1^2 + 45*mu1^2*sigma1^4 + 15*sigma1^6)
             + (1-alpha)*(mu2^6 + 15*mu2^4*sigma2^2 + 45*mu2^2*sigma2^4
                                + 15*sigma2^6);
  moms = c(m1,m2,m3,m4,m5,m6);
  return(moms)
}
```