$\mathcal{M}$-Decomposability and Elliptical Unimodal Densities

by

CHIA KANG-KIANG NICHOLAS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy
in

Statistics

in the

DEPARTMENT OF STATISTICAL SCIENCE,

SCHOOL OF MULTIDISCIPLINARY SCIENCES

of the

GRADUATE UNIVERSITY FOR ADVANCED STUDIES

Fall 2006

$\mathcal{M}$-Decomposability and Elliptical Unimodal Densities

by

CHIA KANG-KIANG NICHOLAS

Abstract

In data analysis and engineering applications, one often comes across unknown densities which are complex and multimodal. In such situations, it is natural and intuitive to break up the original density into a mixture of simpler, structurally less complex densities, so as to facilitate analysis and modelling. In this thesis, we demonstrate that it is possible to *decompose* a multimodal density into simpler densities via the novel concept of $\mathcal{M}$-decomposability. The letter $\mathcal{M}$ derives from "multimodal" or "mixture".

For clarity of presentation, this thesis is divided into two parts. Part one consists of Chapters 1 to 4, and solely considers densities in one-dimension. In Chapter 2, we introduce the notion of *$\mathcal{M}$-decomposability* in one-dimension. We say that a density f is $\mathcal{M}$-decomposable if it is possible to rewrite f as a mixture of two densities g and h such that the sum of the standard deviations of g and h is less than the standard deviation of f. If f does not satisfy the above condition, we say that f is $\mathcal{M}$-undecomposable. To clarify matters, we then provide examples to illustrate the concept of $\mathcal{M}$-decomposability. We also derive a theorem that states that "All uniform densities in one-dimension are $\mathcal{M}$-undecomposable" (Theorem 2.1). In Chapter 3, we demonstrate that unimodal densities in one-dimension can be approximated to an arbitrary level of accuracy using a specially constructed mixture of uniform densities. In Chapter 4, we make use of Theorem 2.1 and the representation in Chapter 3 to derive a theorem which states that "All symmetric unimodal densities in one-dimension are $\mathcal{M}$-undecomposable" (Theorem 4.1).

The second part of the thesis builds up on the results derived in the first and extends to d-dimensions. To avoid confusion of notation, we provide a fresh set of

notations in Chapter 5 and a list of theorems and definitions to apply to the second part of the paper. In Chapters 6 and 7, we provide the theoretical aspects of $\mathcal{M}$-decomposability in d-dimensions. In Chapter 6, we define the uniform density in d-dimensions to be the elliptical uniform. To extend the definition of $\mathcal{M}$-undecomposability to apply d-dimensions, the "standard deviation" that appears in the first part is replaced by the "square-root of the determinant of covariance" of the underlying density. This step is crucial to the future development of $\mathcal{M}$-decomposability in d-dimensional. We derive a theorem that says that "All elliptical uniform densities in d-dimension are $\mathcal{M}$-undecomposable". In Chapter 7, we extend Theorem 4.1 derived in Chapter 4 to d-dimensions, *i.e.*, "All elliptical unimodal densities in d-dimension are $\mathcal{M}$-undecomposable" (Theorem 7.2).

In Chapter 8, we derive a theorem which links $\mathcal{M}$-decomposability with Kullback-Leibler divergence. This provides justification of using $\mathcal{M}$-decomposability in a number of statistical applications, namely clustering and density estimation. Simulation examples of both clustering and density estimation are provided in the chapter. On top of that, we also demonstrate the application of $\mathcal{M}$-decomposability to real data cluster analysis, using the Iris dataset as test data. The results not only show that $\mathcal{M}$-decomposability can be used to improve cluster analysis and density estimation, but also suggest that $\mathcal{M}$-decomposability is a viable criterion for cluster discrimination. Concluding remarks are given in Chapter 9.

Contents

List of Figures

List of Tables

Acknowledgements

Pursuing (in the truest sense of the word) a PhD is a challenging task in life. Everyone who is lucky enough to have gone through the process can verify that. The subset of the above population who survives this far knows that he or she is not alone in this world. At this moment, I can only feel a sense of gratitude and debt towards the many people around me. Compiled below is an incomplete list of people to whom I owe my thanks.

First and foremost, Junji Nakano of the Institute of Statistical Mathematics (ISM) tops the list. Junji provided invaluable advice and insight to my work. He also had to endure the uncertainty of my outputs and my attendance. However, as a great thesis advisor, he always has the right words (and actions) to steer me towards the right direction when everything seems to go the wrong way. He deserves all the credit for getting me here from three years ago.

My gratitude also goes to Dominic Savio Lee (presently with the University of Canterbury at Christchurch, New Zealand) who initiated me to the world of statistics and encouraged me, against the likelihoods, to pursue a PhD. Dominic painstakingly taught me about research, in particular, Bayesian statistics, when we were with DSO National Laboratories in Singapore. I proudly proclaim myself to be Dominic's first ever disciple, a permanant distinction that will never stand to be challenged!

The ultimate misfortune of being my first teachers befalls upon my parents, Seow-Huat and Soy-Ngoo, who were school teachers themselves. It was they who had me mesmerized in the pure beauty and elegance of mathematics and shaped my future. My brother, Kang-Ngee, acts as my eternal rival and friend. I continue to benefit from their unconditional love and support.

Katuomi Hirano, Yoshiyasu Tamura, Satoshi Kuriki from the ISM and Kunio Shimizu from Keio University took valuable time off their research work and invested

energy in trying to understand my script. They also provided expert advice and patiently went through several iterations to improve the manuscript. Without them, my work would remain unrefined.

John Copas of the University of Warwick generously took time to decipher the initial version of my script and provided insightful comments and suggestions for improvement during his recent short visit to the ISM. Eguchi Shinto of the ISM is also gratefully acknowledged for his time. Richard William Farebrother of the University of Manchester kindly supplied his proof of a theorem (the extension of Euler's rotation theorem to n-dimension), which was crucial to the proof of Lemma 6.1, an important building block of the main result of the thesis.

Yukito Iba taught me that the life of a mathematician begins at forty, a view which seriously challenges the judgement of John Charles Fields. Taichi Morichika taught me that life is unconditionally beautiful.

An American writer once said "No man should publish a book without first reading it to a woman." To this, I would add a corollary that includes publishing a PhD thesis. I happen to be extremely fortunate to be blessed with many members of the fairer (and wiser!) sex who generously lend their ears and hearts. They include

- My mother-in-law, Haruko "Helen" Tachibana. At the beginning of 2003, when I commenced my PhD, jobless and homeless, Helen provided a roof over my head, put food into my stomach and hope in my heart. Back then, I was struggling to get out of one of the *local minimas* of my life. Thanks, mum!

- Nahoko Kono, my friend and regular lunch-mate. Nahoko is the unfortunate victim of my perpectual braggings and naggings. Unwillingly and unnecessarily exposed to the world of mathematics, she, however, does not fail to find words of courage and strength to keep me going. She even read Simon Singh's "Fermat's Last Theorem" to find me inspirations.

- "Saint" Kiyoi Watari, the guardian angel for PhD students at the Institute of Statistical Mathematics. Her biblical quotations never fail to provide a source of divine energy when I needed it most.
- Finally, my wife and soulmate Ayako, who does not fail to have faith in me, even during times when I lose faith in myself. Without her unlimited love and understanding, I would not have gone this far; Even if I did, it will not be worthwhile.

Last but not least, our son Jun who recently turns three, taught me the pricelessness of silence when he is around, and the invaluability of companionship when he is not. I shall try to make up for the time I should have spent with him.

Curriculum Vitæ

CHIA KANG-KIANG NICHOLAS

Education

1993	University of Tokyo B.S., Electronics Engineering
2006	Graduate University for Advanced Studies Ph.D., Statistics

Personal

Born	November 9, 1968, Singapore

Selected Journal Publications / Conference Presentations

Lee and Chia (2002) “A Particle Algorithm for Sequential Bayesian Parameter Estimation and Model Selection”, *Special Issue on Monte Carlo Methods of Statistical Signal Processing, IEEE Transactions on Signal Processing (Feb 2002).* **This journal paper has been cited 9 times as of December 2006.**

Lee, Liew, Chia and Cheng (2002) “Bayesian Algorithms for the Passive Location of a Stationary Emitter by a Moving Platform”, *invited paper at Special Session, EUSIPCO 2002, Toulouse, France*

Skills

Languages	Chinese, English, Japanese
Programming	*C^{++}, C, MATLAB, Fortran* 90/95
Operating Systems	Unix, Linux, MacOS X, Windows

Chapter 1

Introduction

In this thesis, we introduce the notion of $\mathcal{M}$-decomposability in probability density functions and present $\mathcal{M}$-decomposability as an alternative non-parametric approach to statistical analysis. When dealing with an unknown density with possibly complex underlying structure, it is natural and intuitive to break up the original density into a mixture of simpler, structurally less complex densities, so as to facilitate analysis and modelling. We demonstrate that the above can be achieved via the concept of $\mathcal{M}$-decomposability. Conceptually, $\mathcal{M}$-decomposability is a non-parametric approach to data analysis and statistical modelling, which is based on a natural strategy in the spirit of divide-and-conquer. One important aspect of $\mathcal{M}$-decomposability is that it can either be applied as a standalone tool, or provide support to improve existing methods in data analysis and modeling. As such, the implementation of $\mathcal{M}$-decomposability can have important consequences in statistics and scientific applications. To the best of our knowledge, the very concept of $\mathcal{M}$-decomposability is the first of its kind in the literature of statistics.

1.1 Existing Parametric and Nonparametric Approaches to Data Analysis

Representing a complicated density via parametric or semi-parametric models has become routine in statistical data analysis. The *finite mixture model* method is probably the most commonly used parametric or semi-parametric approach, and is treated in great detail in [*McLachlan and Basford*, 1988], [*McLachlan and Peel*, 2000] and many others. The idea is to attempt to model a given sample as a mixture of parametric densities (usually unimodal, often Gaussian!), where the parameters of the mixture components (location, scale and number of underlying components) are usually derived using maximum likelihood estimation. Kernel density estimation is a popular semi/non-parametric approach to data analysis, and has been covered by [*Scott*, 1992], [*Silverman*, 1986], [*Wand and Jones*, 1995] among many others. Here, the problem of parameter estimation in the mixture model approach is being transformed to that of *bandwidth estimation*. It is well known that bandwidth estimation works best for densities which are approximately elliptical unimodal, and is problematic with densities comprising of undulating profiles.

1.2 $\mathcal{M}$-Decomposability and Modality

The notion of $\mathcal{M}$-decomposability derived in the thesis is closely related to the aspects of *modality* of a probability density function. Multimodal densities are, by definition, structurally complex and it is both natural and desirable to have multimodal densities represented as mixtures of structurally simpler, unimodal densities as far as the possibility arises. The ideas introduced in this thesis can be implemented either as a standalone tool to locate modally simpler densities within a multimodal density, or as a supplement to further improve existing statistical methodologies like mixture models and kernel density estimation.

As $\mathcal{M}$-decomposability is a novel idea in this thesis, we build its theoretical foundations from scratch. One important class of statistical distributions is the class of symmetric unimodal distributions, among which the Gaussian is perhaps the most commonly used. Unimodality and symmetric unimodality have been previously investigated by [*Anderson*, 1955] and [*Ibragimov*, 1956] among many others. (See, for example, the references on pg 8870 accompanying the section on *Unimodality* in [*Kotz et al.*, 2005]. The class of symmetric unimodal distributions is a more general and flexible class of distributions than the Gaussian (and many others with specific functional forms), without the strongly assumptive functional constraints.

A complimentary class of the symmetric unimodal distributions is the class of multimodal distributions. In this paper, we attempt to quantify the fundamental differences between the densities of unimodal and multimodal distributions. Intuitively, it is possible to express a multimodal density as a mixture of functionally simpler, unimodal ones, such that the sum of the a certain measure of "scatter" of each unimodal density component is less than that of the original density. One possible measure of scatter is to consider the "hypervolume" of the covariance matrix of the density. Fig 1.1 shows that a bimodal density has a larger "scatter" than its mixture components. In this visual example, it is clear that the original density can be expressed as a mixture of two densities with simpler structure. On the other hand, it may be difficult to achieve the same for unimodal densities. The main result of this paper is developed from this relatively simple observation.

1.3 Layout of Thesis

For clarity, this thesis is presented in the following chronological order. From Chapters 2 to 4, we only consider probability density functions in one-dimension. The extension from one-dimension to a more general scenario in d-dimensions is presented from Chapters 5 onwards.

In Chapter 2, we introduce the notion of $\mathcal{M}$-*decomposability* of probability density functions in one-dimension. The prefix '$\mathcal{M}$' in "$\mathcal{M}$-decomposability" can mean both '*multimodal*' and '*mixture*'. Examples are provided to illustrate the concept of $\mathcal{M}$-decomposability. In Chapter 3, we demonstrate that unimodal densities in one-dimension can be approximated using a specially constructed mixture of uniform densities. In Chapter 4, we derive an inequality on symmetric unimodal densities in one-dimension. This ends the first part of the thesis.

The second part of the thesis builds up on the results derived in the first and extends to d-dimensions. To avoid confusion of notation, we provide a fresh set of notations in Chapter 5 and a list of theorems and definitions to apply to the second part of the paper. In Chapters 6 and 7, the theoretical aspects of $\mathcal{M}$-decomposability in d-dimensions are documented. All d-dimensional extensions to theorems and lemmas derived in Chapters 2 and 4 are proven in Chapters 6 and 7. In Chapter 6, we define the uniform density in d-dimensions to be the elliptical uniform. This step is crucial to the future development of $\mathcal{M}$-decomposability in d-dimensional.

In Chapter 8, we derive a theorem which provides justification of using $\mathcal{M}$-decomposability in a number of statistical applications, namely clustering and density estimation. Simulation examples of both clustering and density estimation are provided in the chapter. On top of that, we also demonstrate the application of $\mathcal{M}$-decomposability to real data cluster analysis, using the Iris dataset as test data. Concluding remarks are given in Chapter 9.

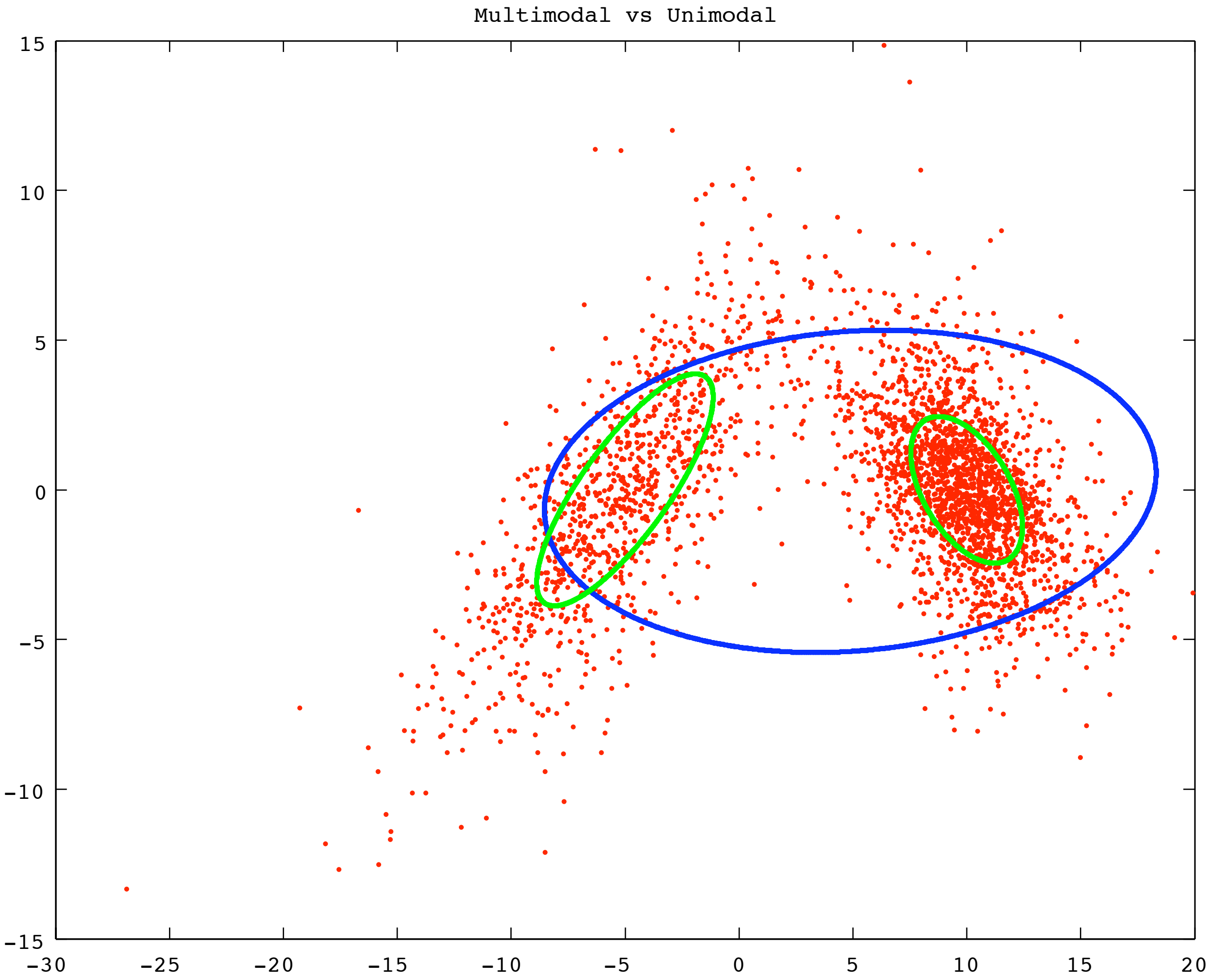

Figure 1.1. Sample from multimodal density. Blue ellipse denotes covariance structure of original density; green ellipses denote covariance structures of each mixture component.

Chapter 2

$\mathcal{M}$-Decomposability

The following notations apply to the first part of this thesis, from Chapter 2 to Chapter 4. We denote the mean and the standard deviation of a one-dimensional density f by μ_f and σ_f respectively. The density of the uniform distribution on the support $[a, b]$ is denoted by $\mathcal{U}(\cdot\,|a,\, b)$ for $(a < b)$. As for unimodality in one-dimension, we say that f is *unimodal* with mode m if there exists a real number m such that f is non-decreasing on $(-\infty, m)$ and non-increasing on (m, ∞). If f does not satisfy the above, we say that f is *multimodal*. Our definition of unimodality is commonly used in textbooks and is comparable with the definitions given in [*Dharmadhikari and Joag-Dev*, 1987] and [*Kotz et al.*, 2005]. If we also have $f(m-x) = f(m+x)$ on top of unimodality, we say that f is *symmetric unimodal* with mode m.

2.1 Definitions

A density f can always be written as a two-component mixture, *i.e.* in the form

$$f(x) = \alpha\, g(x) + (1-\alpha)\, h(x) \tag{2.1}$$

where $0 < \alpha < 1$. Conventionally, g and h are known as the component densities of f. In general, the number of component densities are not limited to two. In this paper, however, the focus is on the decomposition of a density into two components. We define *decomposition pairs* of a probability density function f as follows:

Definition 2.1 (Decomposition Pair) *Given a probability density function f, a pair of densities $\{g, h\}$ satisfying Eq (2.1) is defined as a decomposition pair of f.*

It is clear that there exist infinitely many possible decomposition pairs for a given f.

Definition 2.2 ($\mathcal{M}$-Decomposability) *For a given probability density function f, if there exists a decomposition pair $\{g, h\}$ such that*

$$\sigma_f > \sigma_g + \sigma_h \,,$$

then f is defined to be $\mathcal{M}$-decomposable. Otherwise, f is $\mathcal{M}$-undecomposable. If, for all decomposition pairs $\{g, h\}$,

$$\sigma_f < \sigma_g + \sigma_h \,,$$

then f is strictly $\mathcal{M}$-undecomposable.

2.2 Examples

Example 2.1 (Mixture Density of 2 Gaussians) *Let p be a mixture of two Gaussians such that*

$$p(x) = 0.5\mathcal{N}(x|-m,\,1) + 0.5\mathcal{N}(x|m,\,1)\,.$$

Here, $\mathcal{N}(\cdot|\mu, \sigma)$ denotes the density of the Gaussian with mean μ and standard deviation σ, and $m \geq 0$. The original density p has a standard deviation σ_p which *is* $\sqrt{1+m^2}$. *One possible decomposition pair* $\{q, r\}$ *is easily obtained by setting* $q(x) = \mathcal{N}(x|-m, 1)$ *and* $r(x) = \mathcal{N}(x|m, 1)$, *yielding* $\sigma_q + \sigma_r = 2$. *If* $m > \sqrt{3}$, *then* $\sigma_p > \sigma_q + \sigma_r$ *and accordingly* p *is* $\mathcal{M}$*-decomposable. Fig 2.2 shows the densities of* p *with* $m = 3$ *and* $m = \sqrt{3}$. *The density of* p *with* $m = 3$ *is an example of an* $\mathcal{M}$*-decomposable density.*

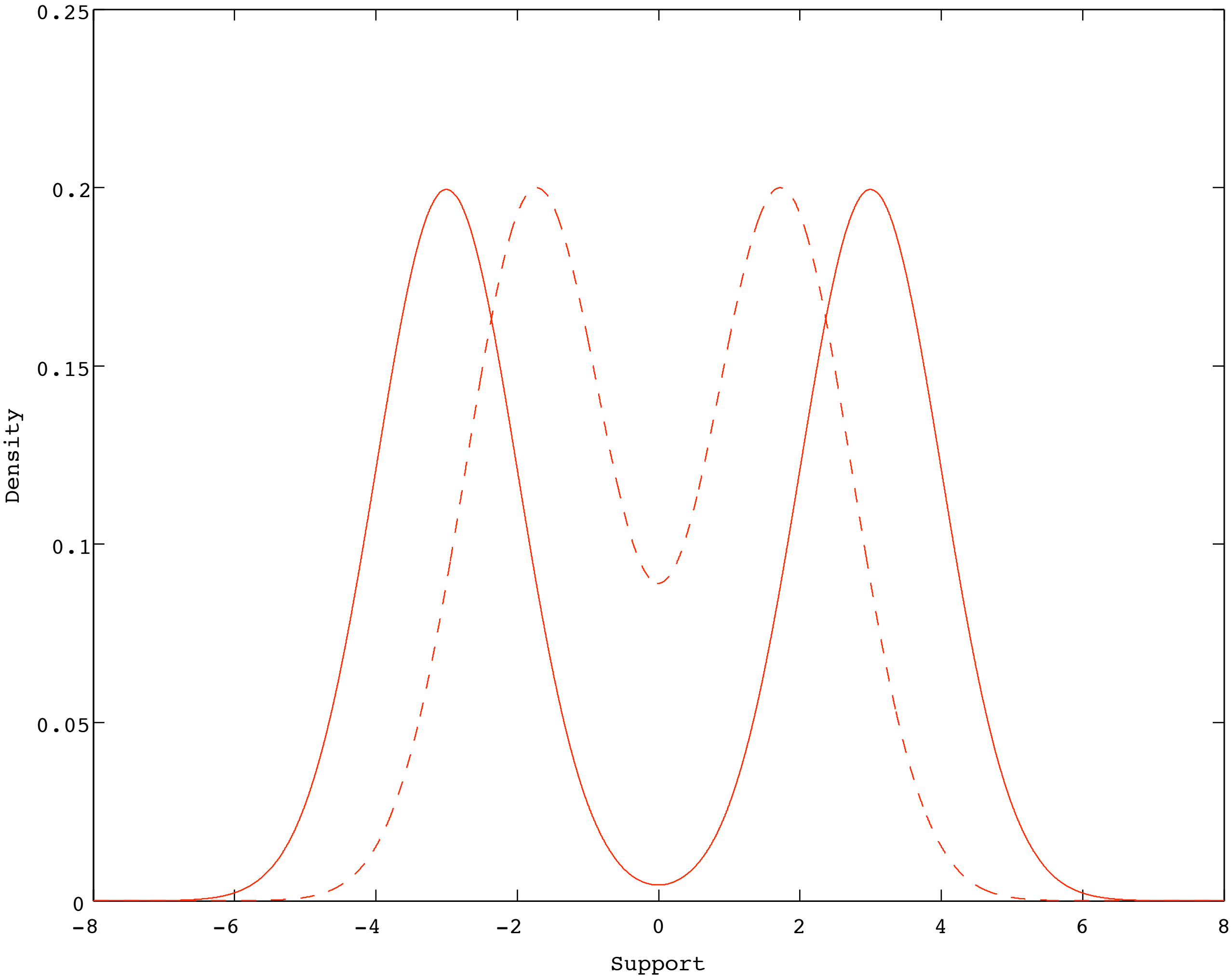

Figure 2.1. Densities in Example 2.1: p with $m = 3$ and $\sqrt{3}$; denoted by solid and broken lines respectively.

From the above argument, a density is likely to be $\mathcal{M}$-decomposable if it is a mixture of two distantly located densities. In Example 2.1, p is $\mathcal{M}$-decomposable for all $m > \sqrt{3}$ by considering the given decomposition pair $\{q, r\}$. It is actually possible to find another decomposition pair $\{q^*, r^*\}$ of p such that $\sigma_{q^*} + \sigma_{r^*} < 2$. For example, set q^* to be p truncated above 0 (hence r^* is p truncated below 0). Then, regardless of m, we must have $\sigma_{q^*} = \sigma_{r^*} < 1$. We are therefore able to conclude that when $m = \sqrt{3}$, p is $\mathcal{M}$-decomposable as well. For $0 < m < \sqrt{3}$, it is difficult to determine the $\mathcal{M}$-decomposability of p.

Next, we present a class of $\mathcal{M}$-decomposable density.

Theorem 2.1 *All uniform densities are $\mathcal{M}$-undecomposable.*

To prove Theorem 2.1, we need to establish the following lemma first.

Lemma 2.1 (Density with Minimum Variance) *Let f be a probability density function such that $f(x) \le M_f < \infty$ for all x. Then*

$$\sigma_f \ge \frac{1}{M_f\sqrt{12}} \,.$$

Identity holds if and only if f is $\mathcal{U}(\,\cdot\,|t, t + 1/M_f)$ for real t's.

Proof We prove Lemma 2.1 in the spirit of *Chebyshev's inequality*. Set $\mu_f = 0$ without loss of generality. Let the density of u be

$$u(x) = \mathcal{U}(x| -\frac{1}{2M_f}, \frac{1}{2M_f}) \,.$$

Therefore, $\mu_u = 0$ and $\sigma_u = 1/(M_f\sqrt{12})$. It is also clear that

$$f(x) \begin{cases} \le u(x) & \text{when } |x| \le \frac{1}{2\,M_f}\,; \\ \ge u(x) & \text{when } |x| > \frac{1}{2\,M_f}\,. \end{cases}$$

Since $\mu_f = \mu_u = 0$, we obtain

$$\begin{aligned}
\sigma_f^2 - \sigma_u^2 &= \int x^2 \left\{ f(x) - u(x) \right\} dx \\
&= \int_{|x| \le \frac{1}{2\,M_f}} x^2 \{ \underbrace{f(x) - u(x)}_{\le 0} \}\, dx + \int_{|x| > \frac{1}{2\,M_f}} x^2 \{ \underbrace{f(x) - u(x)}_{\ge 0} \}\, dx \quad (*) \\
&\ge \frac{1}{4\,M_f^2} \int \left\{ f(x) - u(x) \right\} dx = 0\,.
\end{aligned}$$

Therefore, $\sigma_f^2 \ge \sigma_u^2$ and hence $\sigma_f \ge \sigma_u$. Identity holds if and only if both terms of $(*)$ equal to 0, that is $f(x) = u(x)$. ∎

Using Lemma 2.1, we are ready to prove Theorem 2.1.

Proof [Proof of Theorem 2.1] Let u be a uniform density. We need to prove that for any decomposition pair $\{v, w\}$ of u,

$$\sigma_u \le \sigma_v + \sigma_w\,.$$

Without loss of generality, set $\max(u) = M$ and therefore, and $\sigma_u = 1/(M\sqrt{12})$. Since $u(x) = \alpha\, v(x) + (1-\alpha)\, w(x)$, we have

$$v(x) \le \frac{u(x)}{\alpha} \le \frac{M}{\alpha}; \qquad w(x) \le \frac{u(x)}{1-\alpha} \le \frac{M}{1-\alpha}. \tag{2.2}$$

Using Lemma 2.1, the standard deviations of v and w must satisfy

$$\sigma_v \ge \frac{\alpha}{M\,\sqrt{12}} = \alpha\,\sigma_u; \qquad \sigma_w \ge \frac{1-\alpha}{M\,\sqrt{12}} = (1-\alpha)\,\sigma_u; \tag{2.3}$$

yielding

$$\sigma_v + \sigma_w \ge \sigma_u\,. \quad ∎ \tag{2.4}$$

Remark For identity in Eq (2.4) to hold, equality has to hold for both cases in Eq (2.3). From Lemma 2.1, this occurs if and only if v is uniform with $\max(v) = M/\alpha$ and w is uniform with $\max(w) = M/(1-\alpha)$. The original density u can be written as $u(x) = \mathcal{U}(x|a, b)$ where $b = a + 1/M$. Identity holds in Eq (2.4) if and only if v and w are such that $v(x) = \mathcal{U}(x|a, c)$ and $w(x) = \mathcal{U}(x|c, b)$ where $a < c < b$.

The uniform distribution forms a natural divider between unimodal and multimodal distributions. When the density is cup-shaped with depression occurring near the centre, we have a multimodal distribution. On the other hand, if the density is bell-shaped, with the mode located around the middle, an unimodal distribution is formed. Intuitively, unimodal densities are more likely to be $\mathcal{M}$-undecomposable. In the next example, we investigate the $\mathcal{M}$-decomposability of a skewed unimodal density.

Example 2.2 (L-Shaped Density) *Let the probability density function p be*

$$p(x) = 0.1\,\mathcal{U}(x|0,1) + 0.9\,\mathcal{U}(x|0,9)\,,$$

as depicted in Fig 2.2. The standard deviation of p is $\sigma_p = \sqrt{2257/300} > 2.742$. One can also write p as $p(x) = 0.2\,q(x) + 0.8\,r(x)$, where $q(x) = \mathcal{U}(x|0,1)$ and $r(x) = \mathcal{U}(x|1,9)$. Now, we can easily compute $\sigma_q = \sqrt{1/12} < 0.289$ and $\sigma_r = \sqrt{16/3} < 2.310$. Hence, $\sigma_q + \sigma_r < 2.599 < \sigma_p$ and thus p is $\mathcal{M}$-decomposable.

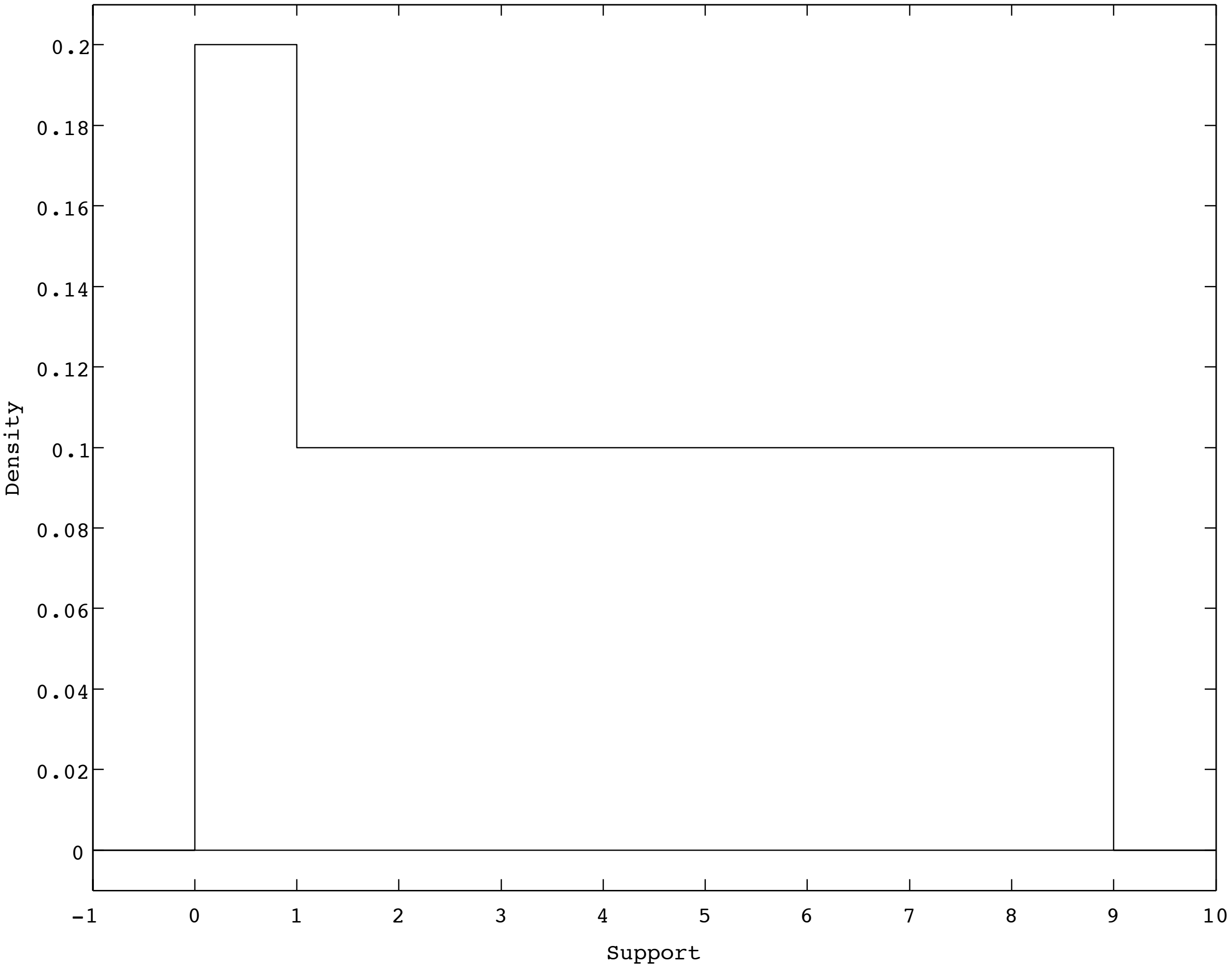

Figure 2.2. Density p which is shown in Example 2.2.

Thus, we have a skewed unimodal density p which is $\mathcal{M}$-decomposable. As such, we conclude that not all unimodal densities are $\mathcal{M}$-undecomposable.

Chapter 3

Representation of Unimodal Densities

From Theorem 2.1, all uniform densities are $\mathcal{M}$-undecomposable. We shall proceed to show that the class of $\mathcal{M}$-undecomposable densities can be extended to include symmetric unimodal densities. For that purpose, we need to represent symmetric unimodal densities via a mixture of uniform densities in a special way presented in this section.

3.1 Approximation via Uniforms

Theorem 3.1 (Representation of Unimodal Densities via Uniforms) *Let f be an unimodal density whose k^{th} moment is finite and is equal to M, where k is even. Then, for all $\epsilon > 0$, it is possible to construct $g_n = \sum_{i=1}^{n} \omega_i u_i$, a mixture of uniforms such that*

$$|\int_{-\infty}^{\infty} x^k \, g_n(x) \, dx - M| < \epsilon \, .$$

Here, each u_i is the density of the uniform on the interval $I_{i,n}$ satisfying $I_{1,n} \supseteq I_{2,n} \supseteq \ldots \supseteq I_{n,n}$, and the weight ω_i, corresponding to u_i, is proportional to the length of the interval $I_{i,n}$.

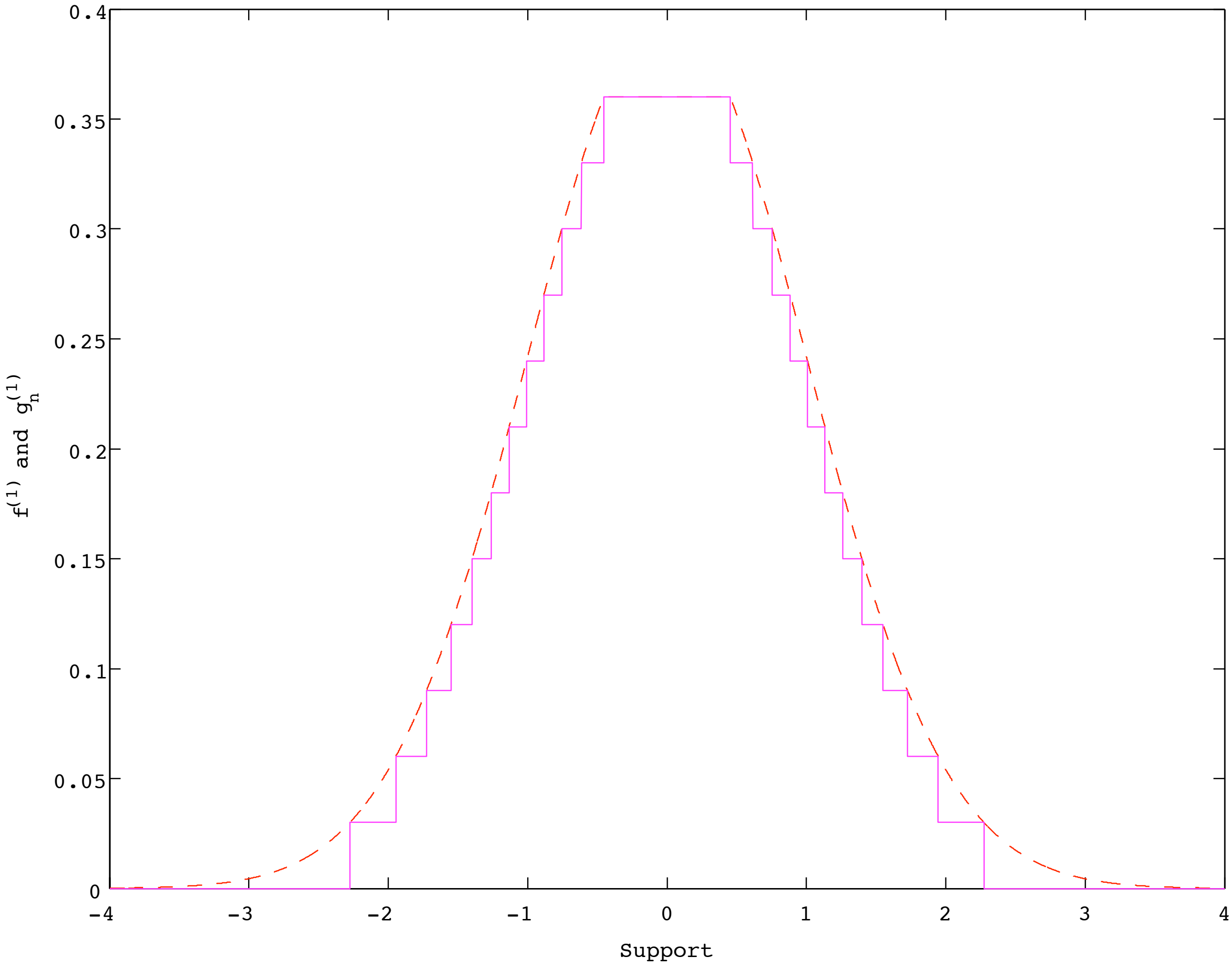

Figure 3.1. $f^{(1)}$ dominating $g_n^{(1)}$. Functions $f^{(1)}$ and $g_n^{(1)}$ denoted by broken and solid lines respectively.

Proof In this proof, all integrals are evaluated from $-\infty$ to ∞. As both $f(x)$ and $x^k f(x)$ are non-negative and integrable, we can define the following functions on non-negative values of y, for a given f:

$$p(y) = \int \min\{f(x), y\}\, dx\,; \qquad q(y) = \int x^k \min\{f(x), y\}\, dx\,. \tag{3.1}$$

Then, both p and q are increasing with $p(0) = q(0) = 0$. If f is unbounded, then p and q are strictly increasing for all y with $\lim_{y\to\infty} p(y) = 1$ and $\lim_{y\to\infty} q(y) = M$. If f is bounded such that $\max(f) = F$, then p and q are strictly increasing for $0 \le y \le F$ and $p(F) = 1$ and $q(F) = M$.

We can rewrite f as a sum of two positive functions in the form

$$f(x) = f^{(1)}(x) + f^{(2)}(x) \tag{3.2}$$

where $f^{(1)}(x) = \min\{f(x), Y\}$ and Y is positive. For a given $\epsilon_1 > 0$, it is possible to choose Y such that

$$1 - \epsilon_1 < \int f^{(1)}(x)\, dx = p(Y) < 1, \tag{3.3}$$

$$M - \epsilon_1 < \int x^k f^{(1)}(x)\, dx = q(Y) < M\,. \tag{3.4}$$

The above "slicing" ensures that the function $f^{(1)}$ is bounded from above by Y. Let $h = Y/n$. Define two sets of real numbers $\{a_{n,1}, \ldots, a_{n,n}\}$ and $\{b_{n,1}, \ldots, b_{n,n}\}$ by

$$a_{n,j} = \inf\{x|f(x) \ge jh\} \quad \text{and} \quad b_{n,j} = \sup\{x|f(x) \ge jh\}\,.$$

Let $I_{n,j}$ denote the interval $[a_{n,j}, b_{n,j}]$ and let $u_{n,j}$ be the density of the uniform on the interval $I_{n,j}$. By construction, a's are monotone non-decreasing and b's are monotone non-increasing, ensuring that $I_{n,1} \supseteq I_{n,2} \supseteq \ldots \supseteq I_{n,n}$. Setting

$$\omega_{n,j} = \frac{b_{n,j} - a_{n,j}}{\sum_{i=1}^{n}(b_{n,i} - a_{n,i})}\,,$$

we create a density g_n such that $g_n(x) = \sum_{j=1}^{n} \omega_{n,j}\, u_{n,j}(x)$. Next, rewrite g_n as a sum of two positive functions in the form of

$$g_n(x) = g_n^{(1)}(x) + g_n^{(2)}(x) \tag{3.5}$$

where $g_n^{(1)}(x) = \sum_{j=1}^{n} (b_{n,j} - a_{n,j})\, h\, u_{n,j}(x)$. Here, all three functions g_n, $g_n^{(1)}$ and $g_n^{(2)}$ are proportional to one another. Each uniform component $(b_{n,j} - a_{n,j})\, h\, u_{n,j}$ of $g_n^{(1)}$ has thickness h. As depicted in Fig 3, we have constructed $g_n^{(1)}$ such that it is dominated everywhere by $f^{(1)}$. Unimodality of f ensures that

$$0 \leq f^{(1)}(x) - g_n^{(1)}(x) \leq \min(f(x), h) \leq h\,.$$

It is then possible to choose n (and hence h) such that

$$\int |g_n^{(1)}(x) - f^{(1)}(x)|\, dx = \int \{f^{(1)}(x) - g_n^{(1)}\}\, dx = p(h) < \epsilon_1 \tag{3.6}$$

$$\int |x^k\, g_n^{(1)}(x) - x^k\, f^{(1)}(x)|\, dx = \int x^k\, \{f^{(1)}(x) - g_n^{(1)}(x)\}\, dx = q(h) < \epsilon_1\,. \tag{3.7}$$

Applying the triangle inequality on integrals twice, we have

$$\begin{aligned} |\int x^k\, g_n(x)\, dx - M| &\leq \int |x^k\, g_n(x) - x^k\, f(x)|\, dx \\ \leq \int |x^k\, g_n^{(1)}(x) - x^k\, f^{(1)}(x)|\, dx &+ \int |x^k\, f^{(2)}(x)|\, dx + \int |x^k\, g_n^{(2)}(x)|\, dx\,. \end{aligned} \tag{3.8}$$

The first term on the last inequality is less than ϵ_1, from Eq (3.7). The second term is also less than ϵ_1, ensured by Eqs (3.2) and (3.4). To quantify the third term, note that from Eqs (3.5) and (3.6),

$$\int g_n^{(2)}(x)\, dx = 1 - \int g_n^{(1)}(x)\, dx < 1 - \int f^{(1)}(x)\, dx + \epsilon_1 < 2\,\epsilon_1$$

and therefore, $\int g_n^{(1)}(x)\, dx > 1 - 2\,\epsilon_1$. Furthermore, since $g_n^{(1)}$ and $g_n^{(2)}$ are proportional,

$$g_n^{(2)}(x) < \frac{2\,\epsilon_1}{1 - 2\,\epsilon_1} \times g_n^{(1)}(x) < \frac{2\,\epsilon_1}{1 - 2\,\epsilon_1} \times f(x)$$

and hence

$$\int x^k\, g_n^{(2)}(x)\, dx < \frac{2\,\epsilon_1}{1-2\,\epsilon_1} \times \int x^k\, f(x)\, dx = \frac{2\,\epsilon_1}{1-2\,\epsilon_1} \times M\,.$$

Choosing $\epsilon_1 < 1/4$, the right side of Eq (3.8) becomes less than $2\epsilon_1(1{+}2M)$. Therefore, starting with any $\epsilon > 0$ and setting

$$\epsilon_1 = \min\{\frac{1}{4}, \frac{\epsilon}{2(1+2M)}\}\,,$$

we attain $|\int x^k\, g_n(x)\, dx - M| < \epsilon$ with the constructed g_n. ▮

Chapter 4

Symmetric Unimodal Densities

In Chapter 2, we proved Theorem 2.1, which states that all uniform densities are $\mathcal{M}$-undecomposable. In this chapter, we shall extend $\mathcal{M}$-undecomposable densities to include a more general class of densities.

Theorem 4.1 (Inequality on Symmetric Unimodal Densities) *Let f be a symmetric unimodal density with finite variance. Then, for any decomposition pair $\{g, h\}$,*

$$\sigma_f \leq \sigma_g + \sigma_h \,.$$

Proof From Theorem 3.1, it is possible to approximate f as a mixture of uniform components as shown below, such that the variances converge:

$$f(x) = \frac{k_1}{k_1 + \ldots + k_n} \mathcal{U}(x| - k_1, k_1) + \ldots + \frac{k_n}{k_1 + \ldots + k_n} \mathcal{U}(x| - k_n, k_n) \,. \tag{4.1}$$

Without loss of generality, we have set the mean m to 0. As f and all uniforms appearing in Eq (4.1) above have means fixed at 0, the variance of f is computed to be

$$\sigma_f^2 = \int x^2 \, f(x) \, dx = \frac{k_1^3 + \ldots + k_n^3}{3\,(k_1 + \ldots + k_n)} \,. \tag{4.2}$$

As a result of f being decomposed into a mixture of two densities g and h, each uniform component is consequently broken up into a mixture of two densities as well. The i^{th} uniform component becomes

$$\mathcal{U}(x|-k_i, k_i) = \alpha_i\, v_i(x) + (1-\alpha_i)\, w_i(x)\,, \tag{4.3}$$

where α_i's are real numbers such that $0 \leq \alpha_i \leq 1$. Here, we allow *some but not all of* α_i's to assume the trivial values of 0 or 1 to ensure the generality of the separation of f. Using Eqs (4.1) and (4.3), we can rewrite f in terms of u_i's and v_i's as follows:

$$\begin{aligned} f(x) = \{ &\frac{k_1\,\alpha_1}{k_1+\ldots+k_n}\, v_1(x) + \ldots + \frac{k_n\,\alpha_n}{k_1+\ldots+k_n}\, v_n(x)\} \\ + \{ &\frac{k_1\,(1-\alpha_1)}{k_1+\ldots+k_n}\, w_1(x) + \ldots + \frac{k_n\,(1-\alpha_n)}{k_1+\ldots+k_n}\, w_n(x)\}\,. \end{aligned}$$

Assigning $\alpha\, g(x)$ and $(1-\alpha)\, h(x)$ the first and second terms respectively, we have

$$\begin{aligned} \alpha\, g(x) &= \frac{k_1\,\alpha_1}{k_1+\ldots+k_n}\, v_1(x) + \ldots + \frac{k_n\,\alpha_n}{k_1+\ldots+k_n}\, v_n(x)\,, \\ (1-\alpha)\, h(x) &= \frac{k_1\,(1-\alpha_1)}{k_1+\ldots+k_n}\, w_1(x) + \ldots + \frac{k_n\,(1-\alpha_n)}{k_1+\ldots+k_n}\, w_n(x)\,, \end{aligned}$$

or more compactly,

$$g(x) \;\propto\; k_1\alpha_1\, v_1(x) + \ldots + k_n\alpha_n\, v_n(x) = l_1\, v_1(x) + \ldots + l_n\, v_n(x)\,, \tag{4.4}$$

$$h(x) \;\propto\; k_1(1-\alpha_1)\, w_1(x) + \ldots + k_n(1-\alpha_n)\, w_n(x) = m_1\, w_1(x) + \ldots + m_n\, w_n(x)$$

where $l_i \equiv k_i\alpha_i$ and $m_i \equiv k_i(1-\alpha_i)$. Note that

$$l_i + m_i = k_i \tag{4.5}$$

for all $1 \leq i \leq n$. By the choice of α's, we circumvent the trivial situation where $g(x) = 0$ or $h(x) = 0$ as at least one l must be neither 0 nor 1. The same applies to m's.

Next, using Eq (4.3) and Theorem 2.1, we obtain

$$\sigma_{v_i} \geq \frac{k_i\alpha_i}{\sqrt{3}} = \frac{l_i}{\sqrt{3}} \quad \text{and} \quad \sigma_{w_i} \geq \frac{k_i(1-\alpha_i)}{\sqrt{3}} = \frac{m_i}{\sqrt{3}}\,.$$

From Eq (4.4), μ_g, the mean of g can be expressed in terms of means of v_i's as

$$\mu_g = \frac{l_1\,\mu_{v_1} + \ldots + l_n\,\mu_{v_n}}{l_1 + \ldots + l_n}\,.$$

Consequently, the variance of g becomes

$$\begin{aligned}
\sigma_g^2 &= \int x^2\, g(x)\, dx - \mu_g^2 \\
&= \frac{l_1\sigma_{v_1}^2 + \ldots + l_n\sigma_{v_n}^2}{l_1 + \ldots + l_n} + \{\frac{l_1\mu_{v_1}^2 + \ldots + l_n\mu_{v_n}^2}{l_1 + \ldots + l_n} - (\frac{l_1\,\mu_{v_1} + \ldots + l_n\,\mu_{v_n}}{l_1 + \ldots + l_n})^2\} \qquad (4.6) \\
&\geq \frac{l_1\sigma_{v_1}^2 + \ldots + l_n\sigma_{v_n}^2}{l_1 + \ldots + l_n} \geq \frac{l_1^3 + \ldots + l_n^3}{3\,(l_1 + \ldots + l_n)}\,.
\end{aligned}$$

The first inequality in Eq (4.6) is the result of Jensen's inequality, ensuring that

$$\frac{l_1\mu_{v_1}^2 + \ldots + l_n\mu_{v_n}^2}{l_1 + \ldots + l_n} \geq (\frac{l_1\,\mu_{v_1} + \ldots + l_n\,\mu_{v_n}}{l_1 + \ldots + l_n})^2\,. \qquad (4.7)$$

Similarly, the variance of h can be bounded from below as

$$\sigma_h^2 \geq \frac{m_1^3 + \ldots + m_n^3}{3\,(m_1 + \ldots + m_n)}\,, \qquad (4.8)$$

yielding,

$$\sigma_g + \sigma_h \geq \frac{1}{\sqrt{3}} \cdot \{\,(\frac{l_1^3 + \ldots + l_n^3}{l_1 + \ldots + l_n})^{\frac{1}{2}} + (\frac{m_1^3 + \ldots + m_n^3}{m_1 + \ldots + m_n})^{\frac{1}{2}}\,\}.$$

From Eq (4.2), we have

$$\sigma_f = \frac{1}{\sqrt{3}} \cdot (\frac{k_1^3 + \ldots + k_n^3}{k_1 + \ldots + k_n})^{\frac{1}{2}}.$$

Therefore, Lemma 4.1 which follows immediately below is a sufficient condition for the inequality $\sigma_f \leq \sigma_g + \sigma_h$ to hold. We are now only left with proof of Lemma 4.1 to complete the proof of Theorem 4.1.

Lemma 4.1 *Let a_i, b_i, c_i be sequences of non-negative real numbers such that for all i, $a_i = b_i + c_i$ and $a_i > 0$. Then the following inequality holds for any positive integer n:*

$$(\frac{a_1^3 + \ldots + a_n^3}{a_1 + \ldots + a_n})^{\frac{1}{2}} \leq (\frac{b_1^3 + \ldots + b_n^3}{b_1 + \ldots + b_n})^{\frac{1}{2}} + (\frac{c_1^3 + \ldots + c_n^3}{c_1 + \ldots + c_n})^{\frac{1}{2}}\,.$$

Equality holds if and only if the sequences a_i, b_i and c_i are linearly dependent.

Proof We prove the inequality in the spirit of [*Hardy et al.*, 1988] and [*Pòlya and Szegö*, 1972].

Set $\mathbf{x} \equiv [x_1, \cdots, x_n]^T$, $\mathbf{y} \equiv [y_1, \cdots, y_n]^T$ and $\mathbf{z} \equiv [z_1, \cdots, z_n]^T$ and similarly for $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Let $\mathbf{x} = t\,\mathbf{y} + (1-t)\,\mathbf{z}$, *i.e.* $x_i = t\,y_i + (1-t)\,z_i$ for all i. Furthermore, define the function ψ as follows:

$$\psi(\mathbf{x}) = (\frac{x_1^3 + \ldots + x_n^3}{x_1 + \ldots + x_n})^{\frac{1}{2}} \tag{4.9}$$

and set $\phi(t) = \psi(t\,\mathbf{y} + (1-t)\,\mathbf{z}) \equiv \psi(\mathbf{x})$ where $0 \leq t \leq 1$. It suffices to prove that $\phi''(t) \geq 0$ for $0 \leq t \leq 1$. This is an immediate consequence of Jensen's inequality as $\phi''(t) \geq 0$ implies $\phi(t) \leq t\,\phi(0) + (1-t)\,\phi(1)$. Setting $t = 1/2$, we have

$$\psi(\frac{\mathbf{y}}{2} + \frac{\mathbf{z}}{2}) \leq \frac{1}{2}\,\psi(\mathbf{y}) + \frac{1}{2}\,\psi(\mathbf{z}).$$

Denoting by $\mathbf{y} = \mathbf{b}$, $\mathbf{z} = \mathbf{c}$, this becomes $2\,\psi(\mathbf{a}/2) \leq \psi(\mathbf{b}) + \psi(\mathbf{c})$. Using Eq (4.9),

$$\psi(\frac{\mathbf{a}}{2}) = (\frac{1}{2})^{(3-1)\cdot\frac{1}{2}} \cdot \psi(\mathbf{a}) = \frac{1}{2} \cdot \psi(\mathbf{a}).$$

Therefore $\phi''(t) \geq 0$ implies $\psi(\mathbf{a}) \leq \psi(\mathbf{b}) + \psi(\mathbf{c})$ as required. Equality holds if and only if $\phi''(t) = 0$.

We define ϕ as

$$\phi(t) = \psi(\mathbf{x}) = (\Sigma\,x_i^3)^{\frac{1}{2}}\,(\Sigma\,x_j)^{-\frac{1}{2}}. \tag{4.10}$$

Differentiating once with respect to t, we have

$$\begin{aligned}\phi'(t) = (\frac{3}{2}) \cdot \phi(t) \cdot [\Sigma\,x_i^3]^{-1} \cdot [\Sigma\,x_k^2\,(y_k - z_k)]\\ -(\frac{1}{2}) \cdot \phi(t) \cdot [\Sigma x_j]^{-1} \cdot [\Sigma(y_k - z_k)]\,.\end{aligned} \tag{4.11}$$

Differentiating again with respect to t, we have

$$\begin{aligned}
\phi''(t) = {} & (\frac{3}{2}) \cdot \phi'(t) \cdot [\Sigma\, x_i^3]^{-1} \cdot [\Sigma\, x_k^2\, (y_k - z_k)] \\
& + (\frac{3}{2}) \cdot \phi(t) \cdot (-1) \cdot [\Sigma\, x_i^3]^{-2} \cdot (3) \cdot [\Sigma\, x_k^2\, (y_k - z_k)]^2 \\
& + (\frac{3}{2}) \cdot \phi(t) \cdot [\Sigma\, x_i^3]^{-1} \cdot (2) \cdot [\Sigma\, x_k\, (y_k - z_k)^2] \\
& - (\frac{1}{2}) \cdot \phi'(t) \cdot [\Sigma x_j]^{-1} \cdot [\Sigma(y_k - z_k)] \\
& - (\frac{1}{2}) \cdot \phi(t) \cdot (-1) \cdot [\Sigma x_j]^{-2} \cdot [\Sigma(y_k - z_k)]^2\,.
\end{aligned} \tag{4.12}$$

After some rearrangements, we have

$$\begin{aligned}
\frac{\phi''(t)}{\phi(t)} = {} & \frac{3}{4} \cdot \underbrace{\{\, [\Sigma x_j]^{-1} \cdot [\Sigma(y_k - z_k)] - [\Sigma x_i^3]^{-1} \cdot [\Sigma x_k^2\, (y_k - z_k)]\, \}^2}_{A} \\
& + (3) \cdot [\Sigma x_i^3]^{-2} \cdot \underbrace{\{\, [\Sigma x_i^3] \cdot [\Sigma x_j (y_j - z_j)^2] - [\Sigma x_k^2\, (y_k - z_k)]^2\, \}}_{B}\,.
\end{aligned} \tag{4.13}$$

Here, term A is a square and therefore $A \geq 0$. To prove that $B \geq 0$, set $p_i^2 = x_i^3$ and $q_j^2 = x_j (y_j - z_j)^2$, and therefore we obtain

$$B = [\Sigma p_i^2] \cdot [\Sigma q_j^2] - [\Sigma p_k\, q_k]^2 \geq 0, \tag{4.14}$$

as an immediate consequence of Cauchy-Schwarz's inequality. As such, $\phi''(t)/\phi(t) \geq 0$ and therefore $\phi''(t) \geq 0$, due to the non-negativeness of x_i, y_i and z_i.

Next, for $B = 0$ to hold in Eq (4.14), there must exist a real number s such that $p_i = s\, q_i$ for all i, implying that $x_i = s\,(y_i - z_i)$. When this happens, term A in Eq (4.13) becomes 0 as well. Combining with the initial condition $x_i = t\, y_i + (1-t)\, z_i$, we have $(s - t)\, y_i = (s - t + 1)\, z_i$, *i.e.* the sequence y_i and z_i (and hence b_i and c_i) must be linearly dependent to ensure that $A = B = 0$, resulting in $\phi''(t) = 0$. Hence Lemma 4.1 is proven and that consequently proves Theorem 4.1. ▮

The following theorem spells the condition for equality in Theorem 4.1 to hold.

Theorem 4.2 *In Theorem* 4.1, $\sigma_f = \sigma_g + \sigma_h$ *holds if and only if* f *is uniform and* $f(x) = \mathcal{U}(x|a, b)$, $g(x) = \mathcal{U}(x|a, c)$, $h(x) = \mathcal{U}(x|c, b)$ *where* $a < c < b$.

Proof To ensure that $\sigma_f = \sigma_g + \sigma_h$, identities must hold in Eqs (4.6) and (4.8). In Eq (4.6), identity in the first inequality is achievable only if $\mu_{v_1} = \ldots = \mu_{v_n}$. Similarly, we must have $\mu_{w_1} = \ldots = \mu_{w_n}$. As for the second inequality in Eq (4.6), identity holds if and only if $v_i(x) = \mathcal{U}(x| - k_i, K_i)$ and $w_i = \mathcal{U}(x|K_i, k_i)$ for all i. When this occurs, we have $|\mu_{v_i} - \mu_{w_i}| = l_i + m_i = k_i$ (or $\mu_{v_i} - \mu_{w_i} = \pm k_i$) for all i. The only possible solution is $k_1 = \ldots = k_n$ and $K_1 = \ldots = K_n$. Hence, the necessary condition is that f is uniform with the prescribed decomposition. The sufficient condition is trivial. ∎

The results in this chapter can be summarized as follows: "The uniform density is $\mathcal{M}$-undecomposable. All other symmetric unimodal densities with finite variances are strictly $\mathcal{M}$-undecomposable."

Chapter 5

Transition from One to d-Dimensions

In Chapter 2, the notion of $\mathcal{M}$-decomposability in one-dimension was introduced. Then, through Chapter 4, we proved that all symmetric unimodal densities in one-dimension with finite variances are $\mathcal{M}$-undecomposable.

The chapters that proceed further contribute to $\mathcal{M}$-decomposability both in the theoretical and practical aspects. Theoretically, we extend the concept of $\mathcal{M}$-decomposability to apply to d-dimensions, where d may assume any positive integer value. As the main result of this thesis, we prove that, using a suitable extension, the inequality on symmetric unimodal densities derived originally for one-dimension applies to higher dimensions. Furthermore, we provide a theoretical justification for using $\mathcal{M}$-decomposability for statistical applications, *e.g.* clustering, mode-finding and density estimation.

5.1 Notations and Theorems for d-Dimensions

For the proceeding chapters, we adopt the following notations. We assume that all probability density functions are defined on the d-dimensional support. $\mathcal{R}^d$ denotes the $(d \times 1)$ vector of real numbers, The $(d \times 1)$ mean vector of a probability density function f is denoted by μ_f, and Σ_f denotes the $(d \times d)$ covariance matrix of f. For any square matrix M, the determinant of M is represented by $|M|$. The zero matrix or zero vector is denoted by $\mathbf{0}$; the identity matrix of order d is denoted by $\mathbf{I}_d$. A real $d \times d$ matrix Q is said to be *orthogonal* if the product of Q and its transpose is equal to the identity matrix, *i.e.* if

$$Q \cdot Q^T = \mathbf{I}_d \,.$$

From the above, it is clear that all orthogonal matrices take 1 or -1 as determinants. The subset of orthogonal matrices, whose determinant is 1, is known as *special orthogonal*. We denote the group of $d \times d$ orthogonal matrices and special orthogonal matrices by $\mathcal{O}(d)$ and $\mathcal{SO}(d)$ respectively. The indicator function, $\mathbb{I}_A$, is defined as follows:

$$\mathbb{I}_A = \begin{cases} 1 & \text{if } A \text{ is true;} \\ 0 & \text{otherwise.} \end{cases}$$

On top of the above, the following list of definitions and theorems are used in the proceeding chapters. Proofs of theorems listed in this chapter can be found in many statistics textbooks and are hence omitted.

Definition 5.1 (Spherical Densities) *We say that f is* spherical *if there exists $\mathbf{x}_0 \in \mathcal{R}^d$ such that*

$$|\mathbf{x}_1 - \mathbf{x}_0| = |\mathbf{x}_2 - \mathbf{x}_0| \quad \Rightarrow \quad f(\mathbf{x}_1) = f(\mathbf{x}_2)$$

Remark One alternative equivalent definition is that there exists a function $\tilde{f}$ such that

$$f(\mathbf{x}) = \tilde{f}(|\mathbf{x} - \mathbf{x}_0|)\,.$$

Definition 5.2 (Undirectional Densities) *We say that f is* undirectional *if the covariance matrix of f is a multiple of the identity matrix, i.e. there exists $k > 0$ such that*

$$\Sigma_f = k\,\mathbf{I}_d\,.$$

Definition 5.3 (Uniform Densities) *We say that f is* uniform *if there exists a subset $A \subset \mathcal{R}^d$ such that*

$$f(\mathbf{x}) \propto \mathbb{I}_{\mathbf{x} \in A}\,.$$

The constant of proportionality, which has been omitted for clarity, is computed as

$$(\int_{\mathbf{x} \in A} \mathbb{I}_{\mathbf{x} \in A}\, d\mathbf{x})^{-1}\,.$$

This term can be physically interpreted as the reciprocal of the hypervolume of the subset A.

If f is of the form

$$f(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mathbf{x}_0)^T\, M\, (\mathbf{x}-\mathbf{x}_0) < r^2}\,, \tag{5.1}$$

where M is an positive semidefinite symmetric matrix, then we say f is elliptical uniform. *If, on top of that, M is undirectional, then $M = k\,\mathbf{I}_d$ and*

$$f(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mathbf{x}_0)^T\, (\mathbf{x}-\mathbf{x}_0) < r'^2} = \mathbb{I}_{|\mathbf{x}-\mathbf{x}_0| < r'}\,, \tag{5.2}$$

we say f is spherical uniform. *($r'^2 = r^2/k$).*

Theorem 5.1 (Change of Variables) *Let f_x be a probability density function defined on $\mathbf{x} \in \mathcal{R}^d$. Let M be an invertible $d \times d$ matrix and $\mathbf{y} = M\,\mathbf{x}$ define a linear transformation of $\mathbf{x}$. Then $\mathbf{y}$ has a probability density function given by*

$$f_y(\mathbf{y}) = f_x(\mathbf{x})/|M| = f_x(M^{-1}\,\mathbf{y})/|M|\,. \tag{5.3}$$

Under the linear transformation M, we have

$$\mu_{f_y} = M \cdot \mu_{f_x} \tag{5.4}$$

and

$$\Sigma_{f_y} = M \cdot \Sigma_{f_x} \cdot M^T \, . \tag{5.5}$$

Corollary 5.1 (Orthogonal Invariance and Undirectionality) *Let f be an undirectional density and $\mathcal{O}(d)$ denote the group of $d \times d$ orthogonal matrices. If an operation $H \in \mathcal{O}(d)$ is applied onto the support, the covariance matrix of the transformed density remains invariant;*

$$H \cdot \Sigma_f \cdot H^T = H \cdot (k \, \mathbf{I}_d) \cdot H^T = k \, H \, H^T = k \, \mathbf{I}_d = \Sigma_f \, .$$

Definition 5.4 (Hyperspherical Coordinates) *We define a coordinate system in a d-dimensional Euclidean space ($d > 2$) in which the coordinates consist of a radial coordinate r and $d-1$ angular coordinates $\phi_1, \ldots, \phi_{d-1}$. If x_i are the Cartesian coordinates, then we may define*

$$\begin{aligned}
x_1 &= r \, \cos(\phi_1) \\
x_2 &= r \, \sin(\phi_1) \, \cos(\phi_2) \\
x_3 &= r \, \sin(\phi_1) \, \sin(\phi_2) \, \cos(\phi_3) \\
&\vdots \\
x_{d-1} &= r \, \sin(\phi_1) \cdots \sin(\phi_{d-2}) \, \cos(\phi_{d-1}) \\
x_d &= r \, \sin(\phi_1) \cdots \sin(\phi_{d-2}) \, \sin(\phi_{d-1}) \, .
\end{aligned}$$

The hyperspherical element can be derived from the Jacobian transformation:

$$\begin{aligned}
dx_1 \cdots dx_d &= |\det \frac{\partial(x_i)}{\partial(r, \phi_i)}| \, dr \, d\phi_1 \cdots d\phi_{d-1} \\
&= r^{d-1} \, \sin^{d-2}(\phi_1) \, \sin^{d-3}(\phi_2) \cdots \sin(\phi_{d-2}) \, dr \, d\phi_1 \cdots d\phi_{d-1} \, .
\end{aligned}$$

Integrating the element of hypervolume, we obtain the next theorem.

Theorem 5.2 (Volume of d-Dimensional Hypersphere) *The volume of a d-dimensional hypersphere with radius r is given as*

$$V_{d,r} = \frac{\pi^{\frac{d}{2}}\, r^d}{\Gamma(\frac{d}{2}+1)}\,, \tag{5.6}$$

where Γ is the Gamma function.

Theorem 5.3 (Covariance of Spherical Uniform Densities) $u_{d,r}(\mathbf{x})$ *is a spherical uniform density defined on $\mathcal{R}^d$ with radius r centred at $\mathbf{c}$ and whose probability density function is given by*

$$u_{d,r}(\mathbf{x}) = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}} r^d}\ \mathbb{I}_{\|\mathbf{x}-\mathbf{c}\|<r}$$

using the indicator function.

The covariance of $u_{d,r}(\mathbf{x})$ is given as follows:

$$\Sigma_{u_{d,r}} = \frac{r^2}{(d+2)}\mathbf{I}_d,$$

and therefore

$$|\Sigma_{u_{d,r}}| = [\frac{r^2}{(d+2)}]^d\,|\mathbf{I}_d| = \frac{r^{2d}}{(d+2)^d}.$$

Using Theorem 5.2, an alternative expression of the above in terms of hyper-volume is as follows:

$$|\Sigma_{u_{d,r}}| = \frac{[\Gamma(\frac{d}{2}+1)]^2}{(d+2)^d\,\pi^d}\,V_{d,r}^2 \propto V_{d,r}^2.$$

The constant of proportionality depends only on the dimension d.

The next two inequalities are taken from [*Cover and Thomas*, 1988]. K_1 and K_2 are non-negative definite symmetric $d \times d$ matrices.

Theorem 5.4 (Theorem 2, Cover and Thomas)

$$|K_1 + K_2| \geq |K_1|\,. \tag{5.7}$$

Theorem 5.5 (Minkowski's Inequality)

$$|K_1 + K_2|^{\frac{1}{d}} \geq |K_1|^{\frac{1}{d}} + |K_2|^{\frac{1}{d}} \,. \tag{5.8}$$

Identity holds if and only if K_1 and K_2 are proportional to each other.

The next theorem, which is related to the representation of special orthogonal matrices ($\mathcal{SO}(d)$), is brought to the attention to the author from [*Bernstein*, 2005]

Theorem 5.6 *Let $A \in \mathcal{R}^{n\times n}$, where $n \geq 2$. Then $A \in \mathcal{SO}(d)$ if and only if there exist m such that $1 \leq m \leq d(d-1)/2$, $\theta_1, \ldots, \theta_m \in \mathcal{R}$, and $j_1, \ldots, j_m, k_1, \ldots, k_m \in \{1, \ldots, d\}$ such that*

$$A = \prod_{i=1}^{m} P(\theta_i, j_i, k_i),$$

where

$$P(\theta, j, k) \equiv I_d + (\cos\theta - 1)(E_{j,j} + E_{k,k}) + (\sin\theta)(E_{j,k} - E_{k,j})$$

and $E_{i,j}$ denotes the $n \times n$ matrix with one at the (i, j)-th element and zero everywhere else.

The proof is given in [*Farebrother and Wrobel*, 2002].

Remark $P(\theta, j, k)$ is a *plane* or *Givens rotation.*

Remark Theorem 5.6 is an extension of Euler's rotation theorem, which is the case when $n = 3$.

Chapter 6

$\mathcal{M}$-Decomposability in d-Dimensions

The definition of $\mathcal{M}$-decomposability given in Chapter 2 involves only the standard deviation of the probability density functions involved. Here, we present an updated version which generalizes to d-dimensions where $d \geq 1$.

We begin with *decomposition pairs.* As finite mixture models apply to d-dimensional densities in general, Definition 2.1, which is introduced in Chapter 2, applies henceforth to densities in d-dimensions as well.

6.1 Definitions

In the first part of the thesis, we are concerned with the standard deviations of the densities involved. In one-dimension, the standard deviation is a natural measure of scatter of a given distribution. When considering higher dimensions, it is possible to consider the square-root of the determinant of the covariance structure of the distribution involved as a corresponding measure of scatter. Henceforth, we shall define the above measure the *pseudo-volume* of the distribution. When a linear transformation is applied to the axes such that the support space is magnified, the pseudo-volume of the new distribution is increased by the same ratio. Hence, "pseudo-volume" is a consistent measure of volume of scatter.

Definition 6.1 (Pseudo-volume) *Let f be a probability density function. Define the pseudo-volume of f as $|\Sigma_f|^{\frac{1}{2}}$, the square-root of the determinant of the covariance matrix of f.*

Remark In one-dimension, the pseudo-volume is simply the standard deviation.

Definition 6.2 ($\mathcal{M}$-Decomposability in d-Dimensions) *For a given probability density function f, if there exists a decomposition pair $\{g, h\}$ such that*

$$|\Sigma_f|^{\frac{1}{2}} > |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}} ,$$

then f is defined to be $\mathcal{M}$-decomposable. Otherwise, f is $\mathcal{M}$-undecomposable. If for all decomposition pairs $\{g, h\}$,

$$|\Sigma_f|^{\frac{1}{2}} < |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}} ,$$

then f is strictly $\mathcal{M}$-undecomposable.

Remark It is trivial to verify that when $d = 1$, Definition 6.2 coincides with Definition 2.2 presented in Chapter 2. In d-dimensions, the definition of $\mathcal{M}$-decomposability can be described compactly using pseudo-volumes.

In the proceeding chapters of this thesis, our goal is to show that corresponding to the one-dimensional case, we have a theorem that says that all "symmetric unimodal densities" in d-dimension are $\mathcal{M}$-undecomposable. The uniform density is trivially defined in one-dimension, but in higher dimensions, the uniform density may assume many different shapes. As the fundamental building block of the d-dimensional extension of symmetric unimodal densities, we define the elliptical uniform as the corresponding d-dimensional uniform density in this thesis. We prove below that all elliptical uniforms are $\mathcal{M}$-undecomposable.

6.2 Elliptical Uniform

Theorem 6.1 (Inequality on Elliptical Uniform Densities) *All elliptical uniform densities defined on $\mathcal{R}^d$ are $\mathcal{M}$-undecomposable for $d = 1$ and strictly $\mathcal{M}$-undecomposable for $d \geq 2$.*

The proof of Theorem 6.1 proceeds the following lemma.

Lemma 6.1 (Density with Minimum Pseudo-volume) *Let f be a probability density function defined on $\mathbf{x} \in \mathcal{R}^d$ such that $f(\mathbf{x}) \leq M_f$ for all $\mathbf{x}$. Then*

$$|\Sigma_f|^{\frac{1}{2}} \geq \frac{\Gamma(\frac{d}{2}+1)}{M_f\,[\pi\,(d+2)]^{\frac{d}{2}}}\,.$$

Identity holds if and only if f is elliptical uniform with $\max(f) = M_f$*.*

Remark When $d = 1$, we recover $\sigma_f \geq 1/(M_f\sqrt{12})$, the result obtained in Lemma 2.1.

Proof We shall denote by u, the density of elliptical uniforms that satisfy $\max(f) = \max(u)$. Without loss of generality, we set $\mu_u = \mu_f = \mathbf{0}$. Our goal is to prove that $|\Sigma_f| \geq |\Sigma_u|$, with identity holding if and only if $f = u$. As f may assume any analytical form, it is generally non-trivial to compare the determinant of covariances of f and u. To circumvent the difficulties incurred by the incompatibility of functional forms, we adopt the following strategy by creating proxies of f and u, both of which are *spherical.* The steps are detailed as follows.

1. Construct undirectional densities (see Definition 5.2) f^w and u^w where $|\Sigma_{f^w}| = |\Sigma_f|$ and $|\Sigma_{u^w}| = |\Sigma_u|$.
2. Construct spherical densities (see Definition 5.1) f^s and u^s where $\Sigma_{f^s} = \Sigma_{f^w}$ and $\Sigma_{u^s} = \Sigma_{u^w}$.
3. Consequently, we have $|\Sigma_{f^s}| = |\Sigma_f|$ and $|\Sigma_{u^s}| = |\Sigma_u|$. Therefore, an equivalent statement of our goal is $|\Sigma_{f^s}| \geq |\Sigma_{u^s}|$.

First, we construct undirectional densities that conform to the conditions set in item 1. Being covariance matrices, both Σ_f and Σ_u must be positive definite and hence expressable using eigenvalues and eigenvectors as

$$\Sigma_f = P \cdot D \cdot P^T .$$

(For simplicity, we only show the case of Σ_f as the same arguments apply to Σ_u.) Here $P \in \mathcal{SO}(d) \subset \mathcal{O}(d)$ and hence satisfies $|P| = 1$ and $P\,P^T = \mathbf{I}_d$, whereas D is a diagonal matrix with diagonal elements $\{\lambda_1^2, \cdots, \lambda_d^2\}$. All λ's are positive. We consider the following linear transformation on the support space. The matrix representation of the linear transformation is of the form $Q = D_l\,P^T$. The first matrix D_l is diagonal with diagonal elements $\{l_1, \cdots, l_d\}$, each l_i satisfying

$$l_i = \frac{\lambda_1^{\frac{1}{d}} \cdots \lambda_d^{\frac{1}{d}}}{\lambda_i} > 0 .$$

As the result of the linear transformation Q on the support space, the original covariance matrix Σ_f is transformed to

$$\Sigma_{f^w} = Q \cdot \Sigma_f \cdot Q^T = D_l \cdot D \cdot D_l^T = (\lambda_1^{\frac{2}{d}} \cdots \lambda_d^{\frac{2}{d}}) \cdot \mathbf{I}_d ,$$

using Eq (5.5) of Theorem 5.1. Hence Σ_{f^w} is undirectional. Next, it is easy to check that the determinants of covariance of both the original covariance matrix Σ_f and the transformed covariance matrix Σ_{f^w} are equal to $\lambda_1^2 \cdots \lambda_d^2$. The determinant of Q is calculated as

$$|Q| = |D_t| = l_1 \cdots l_d = 1\,.$$

Using Eq (5.3) of Theorem 5.1, since $|Q| = 1$, the maximum densities of f and f^w must have the same values, *i.e.* $\max(f^w) = \max(f)$. Similarly, using a suitable linear transformation, it is possible to find a density u^w such that Σ_{u^w} is undirectional and $|\Sigma_{u^w}| = |\Sigma_u|$. On top of these, the linear transformation leaves $\max(u^w) = \max(u)$. From Eqs (5.1) and (5.2), u^w must be spherical uniform. Furthermore, f^w is spherical uniform if and only if f is elliptical uniform.

Next, starting from undirectional densities f^w and u^w, we construct spherical densities f^s and u^s which satisfy item 2. Let f_j denote the resultant probability density function when a rotation operator $R_j \in \mathcal{SO}(d)$ is applied onto the support space of f^w. Then using the result from Corollary 5.1, we have

$$\Sigma_{f_j} = \Sigma_{f^w}\,.$$

Furthermore, from Eq (5.4) of Theorem 5.1, we have

$$\mu_{f_j} = R_j\, \mu_{f^w} = \mathbf{0} = \mu_{f^w}\,.$$

In other words, the mean and covariance of f^w are *invariant to rotation.* For any rotation operators $R_i, R_j \in \mathcal{SO}(d)$, we shall demonstrate below that any *mixture* of f_i and f_j will have the same mean and covariance structure again. Denoting the mixture by g, we have

$$g(\mathbf{x}) = \alpha f_i(\mathbf{x}) + (1-\alpha) f_j(\mathbf{x}), \tag{6.1}$$

where $0 < \alpha < 1$. The covariance of g is given by

$$\Sigma_g = \alpha \Sigma_{f_i} + (1-\alpha)\Sigma_{f_j} + \alpha(1-\alpha)(\mu_{f_i} - \mu_{f_j})(\mu_{f_i} - \mu_{f_j})^T = \Sigma_{f^w}. \tag{6.2}$$

In two dimensions, a rotation operator can be represented as

$$R^{\theta} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} .$$

From Theorem 5.6, it is possible to represent any rotation in d-dimensions as a product of D Given's rotations shown below.

$$R = R_1^{\theta_1} \cdots R_D^{\theta_D}$$

where $D = d\,(d-1)/2$. Some of the Given rotations may be equal to the identity matrix, *i.e.* there may exist j $(1 \le j \le D)$ such that

$$R_j^{\theta_j} = \mathbf{I}_d \,.$$

Now, let us define a density, f^s, as follows:

$$f^s(\mathbf{x}) = (\frac{1}{2\pi})^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} f^w(R_1^{\theta_1} \cdots R_D^{\theta_D} \mathbf{x})\, d\theta_1 \cdots d\theta_D \,. \tag{6.3}$$

f^s is therefore the uniform mixture of all possible rotations of the probability density function f in d-dimension. To show that $\Sigma_{f^s} = \Sigma_{f^w}$, note that

$$\begin{aligned} \Sigma_{f^s} &= \int \mathbf{x}\,\mathbf{x}^T f^s(\mathbf{x})\, d\mathbf{x} \\ &= (\frac{1}{2\pi})^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} \{ \int \mathbf{x}\,\mathbf{x}^T f^w(R_1^{\theta_1} \cdots R_D^{\theta_D} \mathbf{x})\, d\mathbf{x} \}\, d\theta_1 \cdots d\theta_D \,. \end{aligned}$$

The term in braces {} is simply the covariance matrix of the density after application of rotation operator $R_1^{\theta_1} \cdots R_D^{\theta_D}$. As Σ_{f^w} is invariant to rotation, the result remains as Σ_{f^w}. Therefore,

$$\Sigma_{f^s} = (\frac{1}{2\pi})^D \{ \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} d\theta_1 \cdots d\theta_D \}\, \Sigma_{f^w} = \Sigma_{f^w} \,.$$

Furthermore, f^s must be *spherical* by construction, as one can easily verify that $f^s(R\,\mathbf{x}) = f^s(\mathbf{x})$ for any $R \in \mathcal{SO}(d) \subset \mathcal{O}(d)$. On top of these, from Eq (6.3), we have

$$f^s(\mathbf{x}) \le (\frac{1}{2\pi})^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} \max(f^w)\, d\theta_1 \cdots d\theta_D = \max(f^w) = \max(f)\,. \qquad (6.4)$$

We have therefore constructed a spherical density f^s. It is apparent that $u^s = u^w$, which is spherical uniform. Now we are left with item 3, *i.e.* to prove that $|\Sigma_{f^s}| \ge |\Sigma_{u^s}|$ to complete the proof of the lemma.

It is also obvious that the densities f^s and u^s are undirectional. This can be seen from $\Sigma_{f^s} = \Sigma_{f^w}$ and $\Sigma_{u^s} = \Sigma_{u^w}$. Hence, it is possible to find $k_f > 0$ and $k_u > 0$ such that $\Sigma_{f^s} = k_f\,\mathbf{I}_d$ and $\Sigma_{u^s} = k_u\,\mathbf{I}_d$. Our goal will be accomplished if we can prove that $k_f \ge k_u$. From Eq (6.4), we have $f^s(\mathbf{x}) \le \max(f) = M_f$, and the followings are straightfoward:

1. $u^s(\mathbf{x}) \ge f^s(\mathbf{x})$ for $|\mathbf{x}| \le R$, where $u(\mathbf{x}) = M_f$ throughout.

2. $u^s(\mathbf{x}) \le f^s(\mathbf{x})$ for $|\mathbf{x}| > R$, where $u(\mathbf{x}) = 0$ throughout.

Here, R represents the radius of the spherical uniform u^s, and

$$R^d = \frac{\Gamma(\frac{d}{2}+1)}{M_f\,\pi^{\frac{d}{2}}}.$$

Moreover, as f^s and u^s are both spherical and have means $\mathbf{0}$, there exist functions $\tilde{f^s}$ and $\tilde{u^s}$ such that

$$f^s(\mathbf{x}) = \tilde{f^s}(|\mathbf{x}|) = \tilde{f^s}(r); \quad u^s(\mathbf{x}) = \tilde{u}^s(|\mathbf{x}|) = \tilde{u^s}(r)\,,$$

using Definition 5.1 and representation in the hyperspherical coordinates. Furthermore, we define $h(\mathbf{x}) \equiv f^s(\mathbf{x}) - u^s(\mathbf{x})$. Note that $h(\mathbf{x})$ is *not* a probability density function as $h(\mathbf{x})$ takes negative values and

$$\int h(\mathbf{x})\, d\mathbf{x} = 0. \qquad (6.5)$$

Using the hyperspherical coordinate representation, there must exist a function $\tilde{h}$ such that $h(|\mathbf{x}|) = \tilde{h}(r)$, and

$$\tilde{h}(r) \begin{cases} \leq 0 & \text{for } r \leq R; \\ \geq 0 & \text{for } r > R. \end{cases} \tag{6.6}$$

Note that $\tilde{h}$ is identically 0 if and only if $f^s = u^s$, or equivalently, f is elliptical uniform. Now,

$$\begin{aligned} k_f - k_u &= \mathbf{e_1}^T (\Sigma_{f^s} - \Sigma_{u^s}) \, \mathbf{e_1} \\ &= \int \mathbf{e_1}^T \mathbf{x}\mathbf{x}^T \mathbf{e_1} \{ f^s(\mathbf{x}) - u^s(\mathbf{x}) \} \, d\mathbf{x} \\ &= \int |\mathbf{e_1}^T \mathbf{x}|^2 \; h(\mathbf{x}) \; d\mathbf{x} \, . \end{aligned}$$

Here, $\mathbf{e}_1$ is the unit vector parallel to the first axis. Representation via spherical coordinates yields

$$\begin{aligned} k_f - k_u &= \int \cdots \int x_1^2 \, \tilde{h}(r) \, r^{d-1} \, \sin^{d-2}(\phi_1) \cdots \sin(\phi_{d-2}) \, dr \, d\phi_1 \cdots d\phi_{d-1} \\ &= \int_0^\infty r^{d+1} \, \tilde{h}(r) \, dr \times \Phi_1 \times \cdots \times \Phi_{d-1} \, , \end{aligned}$$

with

$$\Phi_1 = \int_0^\pi \cos^2(\phi_1) \, \sin^{d-2}(\phi_1) \, d\phi_1, \quad \Phi_{d-1} = 2\,\pi \, ,$$

and the rest of Φ_i's ($2 \leq i \leq d-2$) satisfying

$$\Phi_i = \int_0^\pi \sin^{d-i-1}(\phi_i) \, d\phi_i \, .$$

Apparently, all Φ_i's are strictly positive and we only need to prove that

$$\int_0^\infty r^{d+1} \, \tilde{h}(r) \, dr \geq 0 \tag{$*$}$$

to arrive at the conclusion that $k_f \geq k_u$. Representing equation (6.5) via hyperspherical coordinates, we have

$$\int_0^\infty r^{d-1} \, \tilde{h}(r) \, dr \times \int_0^\pi \sin^{d-2}(\phi_1) \, d\phi_1 \times \Phi_2 \times \cdots \times \Phi_{d-1} = 0$$

and therefore

$$\int_0^\infty r^{d-1}\,\tilde{h}(r)\,dr = 0\,. \tag{6.7}$$

To prove $(*)$, we break up the integral into as follows:

$$\begin{aligned}\int_0^\infty r^{d+1}\,\tilde{h}(r)\,dr &= \int_0^R r^{d-1}\,r^2\,\underbrace{\tilde{h}(r)}_{\leq 0}\,dr + \int_R^\infty r^{d-1}\,r^2\,\underbrace{\tilde{h}(r)}_{\geq 0}\,dr \\ &\geq \int_0^R r^{d-1}\,R^2\,\tilde{h}(r)\,dr + \int_R^\infty r^{d-1}\,R^2\,\tilde{h}(r)\,dr \\ &= R^2 \times \int_0^\infty r^{d-1}\,\tilde{h}(r)\,dr = 0\,.\end{aligned}$$

Equality holds if and only if $\tilde{h} = 0$ identically, in other words if and only if f is elliptical uniform. This proves $k_f \geq k_u$ and consequently completes the proof for the Lemma 6.1. ▮

We are ready to prove Theorem 6.1.

Proof of Theorem 6.1 Let u be an elliptical uniform density on $\mathbf{x} \in \mathcal{R}^d\,(d > 1)$. We need to prove that for any decomposition pair $\{v, w\}$ of u,

$$|\Sigma_v|^{\frac{1}{2}} + |\Sigma_w|^{\frac{1}{2}} > |\Sigma_u|^{\frac{1}{2}}\,.$$

(The case of $d = 1$ has already been proven in Theorem 2.1.) Without loss of generality, set $\max(u) = M$ and therefore

$$|\Sigma_u|^{\frac{1}{2}} = \frac{\Gamma(\frac{d}{2}+1)}{M\,[\pi(d+2)]^{\frac{d}{2}}}\,.$$

Since $u(\mathbf{x}) = \alpha\,v(\mathbf{x}) + (1-\alpha)\,w(\mathbf{x})$, we have

$$v(\mathbf{x}) \leq \frac{u(\mathbf{x})}{\alpha} \leq \frac{M}{\alpha}; \qquad w(\mathbf{x}) \leq \frac{u(\mathbf{x})}{1-\alpha} \leq \frac{M}{1-\alpha}. \tag{6.8}$$

Using Lemma 6.1, the determinants of covariances of v and w are evaluated as follows:

$$\begin{aligned}|\Sigma_v|^{\frac{1}{2}} &\geq \frac{\alpha\,\Gamma(\frac{d}{2}+1)}{M\,[\pi\,(d+2)]^{\frac{d}{2}}} = \alpha\,|\Sigma_u|^{\frac{1}{2}}; \\ |\Sigma_w|^{\frac{1}{2}} &\geq \frac{(1-\alpha)\,\Gamma(\frac{d}{2}+1)}{M\,[\pi\,(d+2)]^{\frac{d}{2}}} = (1-\alpha)\,|\Sigma_u|^{\frac{1}{2}}.\end{aligned} \tag{6.9}$$

with equalities holding only if the density in question is elliptical uniform. Now, for $d > 1$, we can have *at most one* but *never both* of v, w to be elliptical uniform satisfying equation (6.9). Therefore,

$$|\Sigma_v|^{\frac{1}{2}} + |\Sigma_w|^{\frac{1}{2}} > |\Sigma_u|^{\frac{1}{2}}. \quad \blacksquare$$

Chapter 7

Elliptical Unimodal Densities

7.1 Definition and Representation

In this section, we provide a definition for elliptical unimodal densities. Elliptical densities in d-dimension have been treated in detail by many researchers, see *e.g.* [*Fang et al.*, 1990] and references within. Symmetry in d-dimensions is depicted via ellipticity. Here, we adopt the ideas in [*Fang et al.*, 1990] and modify them to define elliptical unimodal densities in the d-dimensional space.

Definition 7.1 (Elliptical Unimodal Densities) *We say that f is* elliptical unimodal *if there exist $\mu \in \mathcal{R}^{d\times 1}$, $\Sigma \in \mathcal{R}^{d\times d}$ and non-decreasing function g on $\mathcal{R}^{+} \cup \{0\}$ such that*

$$\{\mathbf{x}|f(\mathbf{x}) = g(r)\} = \{\mathbf{x}|(\mathbf{x}-\mu)^T\,\Sigma^{-1}\,(\mathbf{x}-\mu) = r^2\}\,.$$

According to the definition above, elliptical unimodal densities are those whose cross-sections are elliptical, and with mean (μ) and covariance structure (Σ). The above definition encompasses most general densities including d-dimensional Gaussian, logistic, Laplace, Von Mises, beta(k,k), student-t and many other artifical densities.

Theorem 7.1 *Let f be an elliptical unimodal density with mean μ and covariance Σ. Then f can be represented as follows:*

$$f(\mathbf{x}) \propto k_1^d\, u_1(\mathbf{x}) + \cdots + k_n^d\, u_n(\mathbf{x})$$

where u_i is elliptical uniform such that

$$u_i(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mu)^T\,\Sigma^{-1}\,(\mathbf{x}-\mu)<k_i^2}\,. \tag{7.1}$$

Remark From the above representation, each elliptical uniform component is weighted proportionally to the hypervolume of its cross-section. The original elliptical unimodal density is "sliced longitudinally" into elliptical uniforms with a prefixed constant "thickness".

Proof Refer to Theorem 3.1. ▮

7.2 Main Results

Theorem 7.2 (Inequality on Elliptical Unimodal Densities) *Let f be an elliptical unimodal density with finite second moments. Then, for any decomposition pair $\{g, h\}$,*

$$|\Sigma_f|^{\frac{1}{2}} \leq |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}}\,.$$

Identity is possible only when f is uniform in one-dimension.

Proof From Theorem 7.1, we can express f as a finite mixture of elliptical uniform densities as

$$f(\mathbf{x}) = \sum_{i=1}^{n} a_i \cdot u_i(\mathbf{x}), \qquad a_i \propto k_i^d$$

where u_i's, as described in Eq (7.1), are uniform ellipsoidal densities sharing the same mean and aligned in the same direction. We can express g and h as follows:

$$g(\mathbf{x}) = \sum_{i=1}^{n} b_i \cdot v_i(\mathbf{x}), \qquad h(\mathbf{x}) = \sum_{i=1}^{n} c_i \cdot w_i(\mathbf{x}).$$

which satisfies $f(\mathbf{x}) = \alpha\, g(\mathbf{x}) + (1-\alpha)\, h(\mathbf{x})$, yielding

$$a_i \cdot u_i(\mathbf{x}) = \alpha\, b_i \cdot v_i(\mathbf{x}) + (1-\alpha)\, c_i \cdot w_i(\mathbf{x}),$$
$$a_i = \alpha\, b_i + (1-\alpha)\, c_i$$

for all $1 \leq i \leq n$. Following the argument presented in Theorem 6.1, we have

$$|\Sigma_{v_i}|^{\frac{1}{2}} \geq \frac{\alpha b_i}{a_i} |\Sigma_{u_i}|^{\frac{1}{2}}, \qquad |\Sigma_{w_i}|^{\frac{1}{2}} \geq \frac{(1-\alpha)c_i}{a_i} |\Sigma_{u_i}|^{\frac{1}{2}},$$

At the same time, we shall define the following two new densities, corresponding to g and h respectively:

$$\tilde{g}(\mathbf{x}) = \sum_{i=1}^{n} b_i \cdot \tilde{v}_i(\mathbf{x}), \qquad \tilde{h}(\mathbf{x}) = \sum_{i=1}^{n} c_i \cdot \tilde{w}_i(\mathbf{x}).$$

Here, all $\tilde{v}_i$'s and $\tilde{w}_i$'s are ellipsoidal uniforms. $\tilde{v}_i$'s have the same means and the same applies to $\tilde{w}_i$'s. Meanwhile, the determinant of covariance of g is given as follows:

$$\begin{aligned}
|\Sigma_g| &= |(b_1\Sigma_{v_1} + \cdots + b_n\Sigma_{v_n}) + (b_1\mu_{v_1}\mu_{v_1}^T + \cdots + b_n\mu_{v_n}\mu_{v_n}^T)| \\
&\geq |b_1\Sigma_{v_1} + \cdots + b_n\Sigma_{v_n}| \\
&\geq (b_1|\Sigma_{v_1}|^{\frac{1}{d}} + \cdots + b_n|\Sigma_{v_n}|^{\frac{1}{d}})^d \\
&\geq (b_1|\Sigma_{\tilde{v}_1}|^{\frac{1}{d}} + \cdots + b_n|\Sigma_{\tilde{v}_n}|^{\frac{1}{d}})^d \\
&= |b_1\Sigma_{\tilde{v}_1} + \cdots + b_n\Sigma_{\tilde{v}_n}| \\
&= |\Sigma_{\tilde{g}}|.
\end{aligned}$$

The first two inequalities are the direct result of Theorem 5.4 and Theorem 5.5 given in [*Cover and Thomas*, 1988]. The third inequality holds as we must have

$$|\Sigma_{v_i}| \geq |\Sigma_{\tilde{v}_i}|\,,$$

as a direct result of Theorem 6.1. The equality that follows the third inequality is again a result of Theorem 5.5, as all $\Sigma_{\tilde{v}_i}$'s are proportional. We obtain $|\Sigma_{\tilde{g}}|$ as

$$|\Sigma_{\tilde{g}}| = \frac{1}{(d+2)^d} \cdot [\frac{b_1^{1+\frac{2}{d}} + \cdots + b_n^{1+\frac{2}{d}}}{b_1 + \cdots + b_n}]^d.$$

Similarly, we must have

$$|\Sigma_{\tilde{h}}| = \frac{1}{(d+2)^d} \cdot [\frac{c_1^{1+\frac{2}{d}} + \cdots + c_n^{1+\frac{2}{d}}}{c_1 + \cdots + c_n}]^d,$$

and

$$|\Sigma_f| = \frac{1}{(d+2)^d} \cdot [\frac{a_1^{1+\frac{2}{d}} + \cdots + a_n^{1+\frac{2}{d}}}{a_1 + \cdots + a_n}]^d$$

where $a_i = b_i + c_i$ for all i.

$$\begin{aligned} & |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}} \\ \geq & |\Sigma_{\tilde{g}}|^{\frac{1}{2}} + |\Sigma_{\tilde{h}}|^{\frac{1}{2}} \\ = & \frac{1}{(d+2)^{\frac{d}{2}}} \cdot ([\frac{b_1^{1+\frac{2}{d}} + \cdots + b_n^{1+\frac{2}{d}}}{b_1 + \cdots + b_n}]^{\frac{d}{2}} + [\frac{c_1^{1+\frac{2}{d}} + \cdots + c_n^{1+\frac{2}{d}}}{c_1 + \cdots + c_n}]^{\frac{d}{2}}), \end{aligned}$$

and

$$|\Sigma_f|^{\frac{1}{2}} = \frac{1}{(d+2)^{\frac{d}{2}}} \cdot [\frac{a_1^{1+\frac{2}{d}} + \cdots + a_n^{1+\frac{2}{d}}}{a_1 + \cdots + a_n}]^{\frac{d}{2}}.$$

Therefore, Lemma 7.1 below provides a sufficient condition for Theorem 7.2.

Lemma 7.1 *Let a_i, b_i, c_i be sequences of non-negative real numbers such that for all i, $a_i = b_i + c_i$ and $a_i > 0$. Then the following inequality holds for any positive integers d and n.*

$$[\frac{a_1^{1+\frac{2}{d}} + \cdots + a_n^{1+\frac{2}{d}}}{a_1 + \cdots + a_n}]^{\frac{d}{2}} \leq [\frac{b_1^{1+\frac{2}{d}} + \cdots + b_n^{1+\frac{2}{d}}}{b_1 + \cdots + b_n}]^{\frac{d}{2}} + [\frac{c_1^{1+\frac{2}{d}} + \cdots + c_n^{1+\frac{2}{d}}}{c_1 + \cdots + c_n}]^{\frac{d}{2}}.$$

Equality holds if and only if the sequences a_i, b_i and c_i are linearly dependent.

Proof The proof is similar to that of Lemma 4.1, with the only difference being in d. Set $\mathbf{x} \equiv [x_1, \cdots, x_n]^T$, $\mathbf{y} \equiv [y_1, \cdots, y_n]^T$ and $\mathbf{z} \equiv [z_1, \cdots, z_n]^T$ and similarly for $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Let $\mathbf{x} = t\,\mathbf{y} + (1-t)\,\mathbf{z}$, *i.e.* $x_i = t\,y_i + (1-t)\,z_i$ for all i. Furthermore, define the function f as follows:

$$f(\mathbf{x}) = [\frac{x_1^{1+\frac{2}{d}} + \cdots + x_n^{1+\frac{2}{d}}}{x_1 + \cdots + x_n}]^{\frac{d}{2}}$$

and set $\phi(t) = f(t\,\mathbf{y} + (1-t)\,\mathbf{z}) \equiv f(\mathbf{x})$ where $0 \le t \le 1$. It suffices to prove that $\phi''(t) \ge 0$ for $0 \le t \le 1$. This is an immediate consequence of Jensen's inequality as $\phi''(t) \ge 0$ implies $\phi(t) \le t\,\phi(0) + (1-t)\,\phi(1)$. Setting $t = \frac{1}{2}$, we have $f(\frac{\mathbf{y}}{2} + \frac{\mathbf{z}}{2}) \le \frac{1}{2}\,f(\mathbf{y}) + \frac{1}{2}\,f(\mathbf{z})$. Denoting by $\mathbf{y} = \mathbf{b}$, $\mathbf{z} = \mathbf{c}$, this becomes $2\,f(\frac{\mathbf{a}}{2}) \le f(\mathbf{b}) + f(\mathbf{c})$. However, from the definition of f, we must have

$$f(\frac{\mathbf{a}}{2}) = [\frac{1}{2}]^{(1+\frac{2}{d}-1)\cdot\frac{d}{2}}\,f(\mathbf{a}) = \frac{1}{2}f(a)\,.$$

Therefore $\phi''(t) \ge 0$ implies $f(\mathbf{a}) \le f(\mathbf{b}) + f(\mathbf{c})$ as required. Equality holds if and only if $\phi''(t) = 0$.

We shall begin from the definition of ϕ as follows:

$$\phi(t) = f(\mathbf{x}) = [\Sigma\,x_i^{1+\frac{2}{d}}]^{\frac{d}{2}}\,[\Sigma\,x_j]^{-\frac{d}{2}}. \tag{7.2}$$

Differentiating once with respect to t, bearing in mind $x_i = t\,y_i + (1-t)\,z_i$, we have

$$\begin{aligned}\phi'(t) &= (\frac{d}{2})\cdot[\Sigma\,x_i^{1+\frac{2}{d}}]^{\frac{d}{2}-1}\cdot(1+\frac{2}{d})\cdot[\Sigma\,x_k^{\frac{2}{d}}\,(y_k - z_k)]\cdot[\Sigma\,x_j]^{-\frac{d}{2}}\\ &+ [\Sigma\,x_i^{1+\frac{2}{d}}]^{\frac{d}{2}}\cdot(-\frac{d}{2})\cdot[\Sigma x_j]^{-\frac{d}{2}-1}\cdot[\Sigma(y_k - z_k)]\\ &= (\frac{d+2}{2})\cdot\phi(t)\cdot[\Sigma\,x_i^{1+\frac{2}{d}}]^{-1}\cdot[\Sigma\,x_k^{\frac{2}{d}}\,(y_k - z_k)]\\ &- (\frac{d}{2})\cdot\phi(t)\cdot[\Sigma x_j]^{-1}\cdot[\Sigma(y_k - z_k)]\end{aligned} \tag{7.3}$$

Differentiating again with respect to t, we have

$$\begin{aligned}\phi''(t) &= (\frac{d+2}{2})\cdot\phi'(t)\cdot[\Sigma\,x_i^{1+\frac{2}{d}}]^{-1}\cdot[\Sigma\,x_k^{\frac{2}{d}}\,(y_k - z_k)]\\ &+ (\frac{d+2}{2})\cdot\phi(t)\cdot(-1)\cdot[\Sigma\,x_i^{1+\frac{2}{d}}]^{-2}\cdot(\frac{d+2}{d})\cdot[\Sigma\,x_k^{\frac{2}{d}}\,(y_k - z_k)]^2\\ &+ (\frac{d+2}{2})\cdot\phi(t)\cdot[\Sigma\,x_i^{1+\frac{2}{d}}]^{-1}\cdot(\frac{2}{d})\cdot[\Sigma\,x_k^{\frac{2}{d}-1}\,(y_k - z_k)^2]\\ &- (\frac{d}{2})\cdot\phi'(t)\cdot[\Sigma x_j]^{-1}\cdot[\Sigma(y_k - z_k)]\\ &- (\frac{d}{2})\cdot\phi(t)\cdot(-1)\cdot[\Sigma x_j]^{-2}\cdot[\Sigma(y_k - z_k)]^2\end{aligned} \tag{7.4}$$

Replacing the $\phi'(t)$ terms in equation (7.4) with equation (7.3) and rearranging, we have

$$\begin{aligned}
\frac{\phi''(t)}{\phi(t)} &= \frac{d\,(d+2)}{4} \cdot [\Sigma x_j]^{-2} \cdot [\Sigma(y_k - z_k)]^2 \\
&- \frac{d\,(d+2)}{2} \cdot [\Sigma x_i^{1+\frac{2}{d}}]^{-1} \cdot [\Sigma x_j]^{-1} \cdot [\Sigma x_k^{\frac{2}{d}}\,(y_k - z_k)] \cdot [\Sigma(y_l - z_l)] \\
&+ (d+2)^2 \cdot (\frac{1}{4} - \frac{1}{2\,d}) \cdot [\Sigma x_i^{1+\frac{2}{d}}]^{-2} \cdot [\Sigma x_k^{\frac{2}{d}}\,(y_k - z_k)]^2 \\
&+ (\frac{d+2}{d}) \cdot [\Sigma x_i^{1+\frac{2}{d}}]^{-1} \cdot [\Sigma x_k^{\frac{2}{d}-1}\,(y_k - z_k)^2] \qquad (7.5)\\
&= \frac{d\,(d+2)}{4} \cdot \underbrace{\{\,[\Sigma x_j]^{-1} \cdot [\Sigma(y_k - z_k)] - [\Sigma x_i^{1+\frac{2}{d}}]^{-1} \cdot [\Sigma x_k^{\frac{2}{d}}\,(y_k - z_k)]\,\}^2}_{A} \\
&+ (\frac{d+2}{d}) \cdot [\Sigma x_i^{1+\frac{2}{d}}]^{-2} \cdot \underbrace{\{\,[\Sigma x_i^{1+\frac{2}{d}}] \cdot [\Sigma x_j^{\frac{2}{d}-1}(y_j - z_j)^2] - [\Sigma x_k^{\frac{2}{d}}\,(y_k - z_k)]^2\,\}}_{B}
\end{aligned}$$

The term A is expressible as a square and therefore greater or equal to 0. Evaluating B, setting $p_i^2 = x_i^{1+\frac{2}{d}}$ and $q_j^2 = x_j^{\frac{2}{d}-1}(y_j - z_j)^2$, we have

$$\begin{aligned}
B &= [\Sigma x_i^{1+\frac{2}{d}}] \cdot [\Sigma x_j^{\frac{2}{d}-1}(y_j - z_j)^2] - [\Sigma x_k^{\frac{2}{d}}\,(y_k - z_k)]^2 \\
&= [\Sigma p_i^2] \cdot [\Sigma q_j^2] - [\Sigma p_k\, q_k]^2 \qquad (7.6)\\
&\geq 0.
\end{aligned}$$

Inequality holds via Cauchy-Schwarz's inequality. Therefore we must have

$$\phi''(t) \geq 0$$

due to the non-negativeness of x_i, y_i and z_i. Hence Lemma 7.1, and consequently, theorem 7.2 is proved. ▮

Chapter 8

Applications

The main result of the theoretical aspects of $\mathcal{M}$-decomposability in the preceding chapters is the demonstration of $\mathcal{M}$-undecomposability of one large class of densities: the class of elliptical unimodals densities with finite second moments. In this chapter, we present Theorem 8.1 and demonstrate the potential applications of $\mathcal{M}$-decomposability.

Theorem 8.1 ($\mathcal{M}$-Decomposability and Kullback-Leibler Divergence) *Let f be probability density functions defined on $\mathbf{x} \in \mathcal{R}^d$. Let $\{g, h\}$ be a decomposition pair of f such that $f(\mathbf{x}) = \alpha\, g(\mathbf{x}) + (1-\alpha)\, h(\mathbf{x})$. Then the following result applies:*

$$\begin{aligned} & |\Sigma_f|^{\frac{1}{2}} > |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}} \\ \Rightarrow \quad & KL[f \,\|\, \tilde{f}] > KL[f \,\|\, \alpha\, \tilde{g} + (1-\alpha)\, \tilde{h}] \,. \end{aligned}$$

Here, $KL[\,p \,\|\, q\,]$ denotes the Kullback-Leibler divergence given as

$$KL[\,p \,\|\, q\,] = \int p(x) \log \frac{p(x)}{q(x)}\, dx \,,$$

$\tilde{f}$ denotes the Gaussian density with $\mu_{\tilde{f}} = \mu_f$, $\Sigma_{\tilde{f}} = \Sigma_f$; while $\tilde{g}$ and $\tilde{h}$ are similarly defined.

Proof We shall prove that

$$\int f(\mathbf{x}) \, \log \tilde{f}(\mathbf{x}) \, d\mathbf{x} < \int f(\mathbf{x}) \, \log\{\alpha \, \tilde{g}(\mathbf{x}) + (1-\alpha) \, \tilde{h}(\mathbf{x})\} \, d\mathbf{x} \qquad (*)$$

which is an equivalent statement to $KL[\, f \,\|\, \tilde{f}\,] > KL[\, f \,\|\, \alpha \, \tilde{g} + (1-\alpha) \, \tilde{h}\,]$. We have the followings:

$$\begin{aligned}
\text{RHS of } (*) &= \int [\alpha \, g(\mathbf{x}) + (1-\alpha) \, h(\mathbf{x})] \;\; \log[\alpha \, \tilde{g}(\mathbf{x}) + (1-\alpha) \, \tilde{h}(\mathbf{x})] \, d\mathbf{x} \\
&> \alpha \int g(\mathbf{x}) \; \log[\alpha \, \tilde{g}(\mathbf{x})] \, d\mathbf{x} \\
&\quad + (1-\alpha) \int h(\mathbf{x}) \; \log[(1-\alpha) \, \tilde{h}(\mathbf{x})] \, d\mathbf{x} \\
&= \alpha \, [\log \alpha + \int g(\mathbf{x}) \log \tilde{g}(\mathbf{x}) \, d\mathbf{x}\,] \\
&\quad + (1-\alpha) \, [\log(1-\alpha) + \int h(\mathbf{x}) \log \tilde{h}(\mathbf{x}) \, d\mathbf{x}\,].
\end{aligned}$$

From definitions, the probabilitiy density function of $\tilde{g}(\mathbf{x})$ is given as follows

$$\tilde{g}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \, |\Sigma_g|^{-\frac{1}{2}} \exp[-\frac{1}{2} \, (\mathbf{x} - \mu_{\mathbf{g}})^T \, \Sigma_g^{-1} (\mathbf{x} - \mu_{\mathbf{g}})];$$

from which we obtain the followings

$$\begin{aligned}
\log \tilde{g}(\mathbf{x}) &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \, |\Sigma_g| - \frac{1}{2} \, (\mathbf{x} - \mu_{\mathbf{g}})^T \, \Sigma_g^{-1} (\mathbf{x} - \mu_{\mathbf{g}}); \\
\int g(\mathbf{x}) \; \log \tilde{g}(\mathbf{x}) \, d\mathbf{x} &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \, |\Sigma_g| - \frac{1}{2} \int (\mathbf{x} - \mu_{\mathbf{g}})^T \, \Sigma_g^{-1} (\mathbf{x} - \mu_{\mathbf{g}}) \, g(\mathbf{x}) \, d\mathbf{x} \\
&= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \, |\Sigma_g| - \frac{d}{2},
\end{aligned}$$

and similarly applies for $\int f(\mathbf{x}) \log \tilde{f}(\mathbf{x}) \, d\mathbf{x}$ and $\int h(\mathbf{x}) \log \tilde{h}(\mathbf{x}) \, d\mathbf{x}$. We can therefore say that

$$\begin{aligned}
\text{RHS of } (*) &> \alpha \, [\log \alpha + \int g(\mathbf{x}) \log \tilde{g}(\mathbf{x}) \, d\mathbf{x}\,] \\
&\quad + (1-\alpha) \, [\log(1-\alpha) + \int h(\mathbf{x}) \log \tilde{h}(\mathbf{x}) \, d\mathbf{x}\,] \\
&= \alpha \, [\log \alpha - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_g| - \frac{d}{2}] \\
&\quad + (1-\alpha) \, [\log(1-\alpha) - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_h| - \frac{d}{2}] \\
&= \alpha \, [\, \log \alpha - \frac{1}{2} \log |\Sigma_g|\,] + (1-\alpha) \, [\, \log(1-\alpha) - \frac{1}{2} \log |\Sigma_h|\,] \\
&\quad - \frac{d}{2} \log(2\pi) - \frac{d}{2}.
\end{aligned}$$

Meanwhile, applying similarly to f, we have

$$\text{LHS of } (*) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\,|\Sigma_f| - \frac{d}{2}.$$

To complete the prove of the theorem, it suffices to demonstrate that

$$\alpha\,[\,\log\alpha - \frac{1}{2}\log|\Sigma_g|\,] + (1-\alpha)[\,\log(1-\alpha) - \frac{1}{2}\log|\Sigma_h|\,] > -\frac{1}{2}\log|\Sigma_f|,$$

or equivalently,

$$\alpha\,[\log\frac{|\Sigma_g|^{\frac{1}{2}}}{\alpha}] + (1-\alpha)\,[\log\frac{|\Sigma_h|^{\frac{1}{2}}}{(1-\alpha)}] < \log|\Sigma_f|^{\frac{1}{2}}. \qquad (**)$$

Using Jensen's inequality, we have

$$\begin{aligned} \text{LHS of } (**) &\leq \log\{\alpha\frac{|\Sigma_g|^{\frac{1}{2}}}{\alpha} + (1-\alpha)\frac{|\Sigma_h|^{\frac{1}{2}}}{(1-\alpha)}\} \\ &= \log(|\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}}) \\ &\leq \log|\Sigma_f|^{\frac{1}{2}} = \text{RHS of } (**). \quad \blacksquare \end{aligned}$$

Hence, the proof of Theorem 8.1 is complete.

Theorem 8.1 says that if one finds a pair of mixture components $\{g, h\}$ of f such that the sum of pseudo-volumes of the mixture components is less than the pseudo-volume of the original density then in Kullback-Leibler sense, the Gaussian mixture obtained by $\tilde{g}$ and $\tilde{h}$ makes a better estimate of the original f than the Gaussian $\tilde{f}$. In practice, it is possible to further relax the conditions to adapt to applicational needs. In the following sections, we demonstrate the use of Theorems 7.2 and 8.1 to statistical applications, namely clustering and density estimation.

8.1 Clustering via $\mathcal{M}$-Decomposability

Clustering is the subject of active research for more than fifty years. It has been applied in several fields, such as statistic, pattern recognition and machine learning. Many clustering techniques and algorithms have been developed. The survey paper [*Berkhin*, 2002] provides a detail reference of most of the popular techniques and algorithms in use today.

In this thesis, we show that it is straightfoward to apply $\mathcal{M}$-decomposability to cluster analysis. Here, we demonstrate a clustering strategy using both Theorem 7.2 and Theorem 8.1. Our strategy is non-parametric and hence it is possible to locate clusters without prior knowledge of the function structure of clusters. Furthermore, the number of clusters does not have to be known beforehand.

Given a sample of size n, $\{X_1, \cdots, X_n\}$. The task in cluster analysis is to subdivide the original sample into distinct groups such that members of each group are close to each other, and distant from members of different groups. There exist many approaches to cluster analysis. One popular approach is to assume that each underlying cluster is drawn from a Gaussian, or some other known parametric distribution. In other words, the sample is assumed to generated a mixture distribution of known functional form. The problem becomes one of parameter estimation, and the unknown parameters are approximated via maximum likelihood. Here, the EM algorithm is extensively used. This approach via finite mixture is described in detail in [*McLachlan and Peel*, 2000]. Besides the necessity for prior knowledge of distribution of each cluster, one additional difficulty in this approach is that the total number of clusters has to be set beforehand. The unknown number of clusters can be evaluated via a Bayesian approach, or independently via AIC.

Our approach assumes that each cluster is approximately symmetric unimodal. As such, there is no need to know the functional structure of the underlying distribution beforehand. From the given sample $F = \{X_1, \cdots, X_n\}$, we are interested to know if the original sample can be decomposed into two groups, such that the sum of pseudo-volumes of the groups is less than that of the original sample. We denote any two subgroups as G, H such that

$$G = \{Y_1, \cdots, Y_m\}, \quad H = \{Y_{m+1}, \cdots, Y_n\}$$

and $G \cup H = F$, with Y's being a regrouping of X. We further denote the sample covariance matrix of F, G, H as $\mathcal{S}_F$, $\mathcal{S}_G$ and $\mathcal{S}_H$. Our task is to find the optimal grouping (or approximation of decomposition pair) $\{G, H\}$ such that

$$|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}$$

is minimized and test this value against $|\mathcal{S}_F|^{\frac{1}{2}}$. If

$$|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}} < |\mathcal{S}_F|^{\frac{1}{2}},$$

then, from Theorem 7.2, we conclude that F is not symmetric unimodal and from Theorem 8.1, a Gaussian mixture of G and H provides a better estimation of the underlying density that F. We conclude that it is better to break F up into G and H. However, if

$$|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}} \geq |\mathcal{S}_F|^{\frac{1}{2}},$$

then there is no reason to decompose F as F is $\mathcal{M}$-undecomposable.

When one arrives at the conclusion that F is $\mathcal{M}$-undecomposable, it is possible to stop the cluster analysis process with one cluster. However, if F is found to be $\mathcal{M}$-decomposable with subgroups (decomposition pair) $\{G, H\}$, one may repeat the decomposition process with G and H, until all subgroups are $\mathcal{M}$-undecomposable. When that happens, the "splitting" process of our strategy ends.

To prevent overclustering, our strategy also includes "merging" of clusters. At the point when all splitted subclusters are $\mathcal{M}$-undecomposable, we take two subclusters at a time and perform the following test. Now, let Q, R denote the two chosen subclusters and P be the union of the two subclusters, *i.e.* $P = Q \cup R$. We then check the sum of the pseudo-volumes of Q and R and compare against that of P. If

$$|\mathcal{S}_Q|^{\frac{1}{2}} + |\mathcal{S}_R|^{\frac{1}{2}} \geq |\mathcal{S}_P|^{\frac{1}{2}} ,$$

then we conclude that Q and R should be merged to form a larger cluster P. This process is repeated until there are no more mergeable subclusters left.

We have described the concept of using $\mathcal{M}$-decomposability to perform cluster analysis. As for the algorithm, the crucial point is to find the optimal decomposition $\{G, H\}$ such that $|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}$ is minimized. There are many possible approaches to this task. To perform this task rigorously to find the global minimum is computationally difficult and may be equivalent to a NP-hard class of problem. Here, we propose a computationally simpler approach. At each spitting stage, starting from F, we fit a two-mixture Gaussian and use the EM algorithm to obtain the decomposition pair $\{G, H\}$. We present an example of cluster analysis using this more feasible alternative.

8.1.1 Simulation Example

The example provided here is a sample F drawn from a three-mixture of logistic distribution as shown in Fig 8.1. Neither the number of clusters nor the functional form of the clusters are know beforehand. At the first splitting, we assume that F is generated from a two-Gassian mixture, and perform EM to obtain the approximate decomposition pair of $\{G, H\}$, which is shown in Fig 8.2. It turns out that with this decomposition pair,

$$|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}} < |\mathcal{S}_F|^{\frac{1}{2}}$$

and therefore F is splitted into G and H. The splitting process is repeated for G and H and the results are shown in Figs 8.3 and 8.4. At this point, all four subclusters are $\mathcal{M}$-undecomposable.

Finally, we begin the merging process and find that two clusters Q and R (shown in blue and green in Fig 8.5) satisfy

$$|\mathcal{S}_Q|^{\frac{1}{2}} + |\mathcal{S}_R|^{\frac{1}{2}} \geq |\mathcal{S}_P|^{\frac{1}{2}} ,$$

where $P = Q \cup R$. The two clusters are then merged and we are left with three clusters shown in Fig 8.6.

8.1.2 Analysis of Iris Dataset via $\mathcal{M}$-Decomposability

Next, we attempt to analyze the famous Iris dataset using $\mathcal{M}$-decomposability. The dataset is first provided by [*Anderson*, 1935] and has been extensively used over the years by many researchers, including [*Fisher*, 1936]. It is also available electronically at the University of California at Irvine (UCI) machine learning group [*Newman et al.*, 1998]. The dataset consists of 150 four-dimensional data. The four attribute information given are sepal length, sepal width, petal length and petal width, all given in centimetres. There are altogether three classes, namely "Setosa", "Versicolor" and "Virginica", given in the proportion of 50 : 50 : 50.

Starting from the 150 four-dimensional data and assuming that the number of underlying classes are unknown, we perform cluster analysis via $\mathcal{M}$-decomposability. We are able to confirm that there are altogether three underlying clusters, in the proportion of 50 : 45 : 55. The first 50 data coincide with "Setosa" (0 misspecification). For "Versicolor" and "Virginica", there are altogether five misspecifications. (Five "Versicolor" are mislabeled as "Virginica"). Using $\mathcal{M}$-decomposability, we achieve 0% misspecification for the "Setosa" class and 5% misspecification for the more challenging "Versicolor" and "Virginica" classes. The original data, true class, and the class estimated via $\mathcal{M}$-decomposability are given in Table 8.1. The data is also depicted graphically in Fig 8.7 (true class) and Fig 8.8 (estimated class).

Despite the fact that our analysis results in five cases of misspecifications, it should be noted that given the four attribute information, our allocation of "Versicolor" and "Virginica" achieves a tighter pseudo-volume than the "true underlying". Denoting the classes of "Versicolor" and "Virginica" by v_1 and v_2 respectively, our estimation yields

$$|\hat{\Sigma}_{v_1}|^{\frac{1}{2}} + |\hat{\Sigma}_{v_2}|^{\frac{1}{2}} \approx 0.01563\,,$$

as compared to

$$|\Sigma_{v_1}|^{\frac{1}{2}} + |\Sigma_{v_2}|^{\frac{1}{2}} \approx 0.01587$$

of the true underlying. Furthermore, refering to Fig 8.7 and 8.8, we see that the five misspecified data lie in the vicinity of the "boundary" between "Versicolor" and "Virginica".

Table 8.1: Iris Data

sepal length	sepal width	petal length	petal width	true class	estimated class
5.1	3.5	1.4	0.2	Setosa	Setosa
4.9	3.0	1.4	0.2	Setosa	Setosa
4.7	3.2	1.3	0.2	Setosa	Setosa
4.6	3.1	1.5	0.2	Setosa	Setosa
5.0	3.6	1.4	0.2	Setosa	Setosa
5.4	3.9	1.7	0.4	Setosa	Setosa
4.6	3.4	1.4	0.3	Setosa	Setosa
5.0	3.4	1.5	0.2	Setosa	Setosa
4.4	2.9	1.4	0.2	Setosa	Setosa
4.9	3.1	1.5	0.1	Setosa	Setosa
5.4	3.7	1.5	0.2	Setosa	Setosa
4.8	3.4	1.6	0.2	Setosa	Setosa
4.8	3.0	1.4	0.1	Setosa	Setosa
4.3	3.0	1.1	0.1	Setosa	Setosa
5.8	4.0	1.2	0.2	Setosa	Setosa
5.7	4.4	1.5	0.4	Setosa	Setosa
5.4	3.9	1.3	0.4	Setosa	Setosa
5.1	3.5	1.4	0.3	Setosa	Setosa
5.7	3.8	1.7	0.3	Setosa	Setosa
5.1	3.8	1.5	0.3	Setosa	Setosa
5.4	3.4	1.7	0.2	Setosa	Setosa
5.1	3.7	1.5	0.4	Setosa	Setosa
4.6	3.6	1.0	0.2	Setosa	Setosa

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
5.1	3.3	1.7	0.5	Setosa	Setosa
4.8	3.4	1.9	0.2	Setosa	Setosa
5.0	3.0	1.6	0.2	Setosa	Setosa
5.0	3.4	1.6	0.4	Setosa	Setosa
5.2	3.5	1.5	0.2	Setosa	Setosa
5.2	3.4	1.4	0.2	Setosa	Setosa
4.7	3.2	1.6	0.2	Setosa	Setosa
4.8	3.1	1.6	0.2	Setosa	Setosa
5.4	3.4	1.5	0.4	Setosa	Setosa
5.2	4.1	1.5	0.1	Setosa	Setosa
5.5	4.2	1.4	0.2	Setosa	Setosa
4.9	3.1	1.5	0.1	Setosa	Setosa
5.0	3.2	1.2	0.2	Setosa	Setosa
5.5	3.5	1.3	0.2	Setosa	Setosa
4.9	3.1	1.5	0.1	Setosa	Setosa
4.4	3.0	1.3	0.2	Setosa	Setosa
5.1	3.4	1.5	0.2	Setosa	Setosa
5.0	3.5	1.3	0.3	Setosa	Setosa
4.5	2.3	1.3	0.3	Setosa	Setosa
4.4	3.2	1.3	0.2	Setosa	Setosa
5.0	3.5	1.6	0.6	Setosa	Setosa
5.1	3.8	1.9	0.4	Setosa	Setosa
4.8	3.0	1.4	0.3	Setosa	Setosa
5.1	3.8	1.6	0.2	Setosa	Setosa

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
4.6	3.2	1.4	0.2	Setosa	Setosa
5.3	3.7	1.5	0.2	Setosa	Setosa
5.0	3.3	1.4	0.2	Setosa	Setosa
7.0	3.2	4.7	1.4	Versicolor	Versicolor
6.4	3.2	4.5	1.5	Versicolor	Versicolor
6.9	3.1	4.9	1.5	Versicolor	Versicolor
5.5	2.3	4.0	1.3	Versicolor	Versicolor
6.5	2.8	4.6	1.5	Versicolor	Versicolor
5.7	2.8	4.5	1.3	Versicolor	Versicolor
6.3	3.3	4.7	1.6	Versicolor	Versicolor
4.9	2.4	3.3	1.0	Versicolor	Versicolor
6.6	2.9	4.6	1.3	Versicolor	Versicolor
5.2	2.7	3.9	1.4	Versicolor	Versicolor
5.0	2.0	3.5	1.0	Versicolor	Versicolor
5.9	3.0	4.2	1.5	Versicolor	Versicolor
6.0	2.2	4.0	1.0	Versicolor	Versicolor
6.1	2.9	4.7	1.4	Versicolor	Versicolor
5.6	2.9	3.6	1.3	Versicolor	Versicolor
6.7	3.1	4.4	1.4	Versicolor	Versicolor
5.6	3.0	4.5	1.5	Versicolor	Versicolor
5.8	2.7	4.1	1.0	Versicolor	Versicolor
6.2	2.2	4.5	1.5	Versicolor	*Virginica*
5.6	2.5	3.9	1.1	Versicolor	Versicolor
5.9	3.2	4.8	1.8	Versicolor	*Virginica*

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
6.1	2.8	4.0	1.3	Versicolor	Versicolor
6.3	2.5	4.9	1.5	Versicolor	*Virginica*
6.1	2.8	4.7	1.2	Versicolor	Versicolor
6.4	2.9	4.3	1.3	Versicolor	Versicolor
6.6	3.0	4.4	1.4	Versicolor	Versicolor
6.8	2.8	4.8	1.4	Versicolor	Versicolor
6.7	3.0	5.0	1.7	Versicolor	*Virginica*
6.0	2.9	4.5	1.5	Versicolor	Versicolor
5.7	2.6	3.5	1.0	Versicolor	Versicolor
5.5	2.4	3.8	1.1	Versicolor	Versicolor
5.5	2.4	3.7	1.0	Versicolor	Versicolor
5.8	2.7	3.9	1.2	Versicolor	Versicolor
6.0	2.7	5.1	1.6	Versicolor	*Virginica*
5.4	3.0	4.5	1.5	Versicolor	Versicolor
6.0	3.4	4.5	1.6	Versicolor	Versicolor
6.7	3.1	4.7	1.5	Versicolor	Versicolor
6.3	2.3	4.4	1.3	Versicolor	Versicolor
5.6	3.0	4.1	1.3	Versicolor	Versicolor
5.5	2.5	4.0	1.3	Versicolor	Versicolor
5.5	2.6	4.4	1.2	Versicolor	Versicolor
6.1	3.0	4.6	1.4	Versicolor	Versicolor
5.8	2.6	4.0	1.2	Versicolor	Versicolor
5.0	2.3	3.3	1.0	Versicolor	Versicolor
5.6	2.7	4.2	1.3	Versicolor	Versicolor

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
5.7	3.0	4.2	1.2	Versicolor	Versicolor
5.7	2.9	4.2	1.3	Versicolor	Versicolor
6.2	2.9	4.3	1.3	Versicolor	Versicolor
5.1	2.5	3.0	1.1	Versicolor	Versicolor
5.7	2.8	4.1	1.3	Versicolor	Versicolor
6.3	3.3	6.0	2.5	Virginica	Virginica
5.8	2.7	5.1	1.9	Virginica	Virginica
7.1	3.0	5.9	2.1	Virginica	Virginica
6.3	2.9	5.6	1.8	Virginica	Virginica
6.5	3.0	5.8	2.2	Virginica	Virginica
7.6	3.0	6.6	2.1	Virginica	Virginica
4.9	2.5	4.5	1.7	Virginica	Virginica
7.3	2.9	6.3	1.8	Virginica	Virginica
6.7	2.5	5.8	1.8	Virginica	Virginica
7.2	3.6	6.1	2.5	Virginica	Virginica
6.5	3.2	5.1	2.0	Virginica	Virginica
6.4	2.7	5.3	1.9	Virginica	Virginica
6.8	3.0	5.5	2.1	Virginica	Virginica
5.7	2.5	5.0	2.0	Virginica	Virginica
5.8	2.8	5.1	2.4	Virginica	Virginica
6.4	3.2	5.3	2.3	Virginica	Virginica
6.5	3.0	5.5	1.8	Virginica	Virginica
7.7	3.8	6.7	2.2	Virginica	Virginica
7.7	2.6	6.9	2.3	Virginica	Virginica

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
6.0	2.2	5.0	1.5	Virginica	Virginica
6.9	3.2	5.7	2.3	Virginica	Virginica
5.6	2.8	4.9	2.0	Virginica	Virginica
7.7	2.8	6.7	2.0	Virginica	Virginica
6.3	2.7	4.9	1.8	Virginica	Virginica
6.7	3.3	5.7	2.1	Virginica	Virginica
7.2	3.2	6.0	1.8	Virginica	Virginica
6.2	2.8	4.8	1.8	Virginica	Virginica
6.1	3.0	4.9	1.8	Virginica	Virginica
6.4	2.8	5.6	2.1	Virginica	Virginica
7.2	3.0	5.8	1.6	Virginica	Virginica
7.4	2.8	6.1	1.9	Virginica	Virginica
7.9	3.8	6.4	2.0	Virginica	Virginica
6.4	2.8	5.6	2.2	Virginica	Virginica
6.3	2.8	5.1	1.5	Virginica	Virginica
6.1	2.6	5.6	1.4	Virginica	Virginica
7.7	3.0	6.1	2.3	Virginica	Virginica
6.3	3.4	5.6	2.4	Virginica	Virginica
6.4	3.1	5.5	1.8	Virginica	Virginica
6.0	3.0	4.8	1.8	Virginica	Virginica
6.9	3.1	5.4	2.1	Virginica	Virginica
6.7	3.1	5.6	2.4	Virginica	Virginica
6.9	3.1	5.1	2.3	Virginica	Virginica
5.8	2.7	5.1	1.9	Virginica	Virginica

continued on next page

continued from previous page

sepal length	sepal width	petal length	petal width	true class	estimated class
6.8	3.2	5.9	2.3	Virginica	Virginica
6.7	3.3	5.7	2.5	Virginica	Virginica
6.7	3.0	5.2	2.3	Virginica	Virginica
6.3	2.5	5.0	1.9	Virginica	Virginica
6.5	3.0	5.2	2.0	Virginica	Virginica
6.2	3.4	5.4	2.3	Virginica	Virginica
5.9	3.0	5.1	1.8	Virginica	Virginica

Table 8.1: True Iris data

8.2 Density Estimation

Density estimation is an important statistical tool that has seen wide applications in many scientific and engineering fields. Given raw measurements or data, the task is to reconstruct or estimate the underlying density from which the original data is presumed to be generated. The problem statement is as follows. Given $\{X_1, \cdots X_n\}$ which is assumed to be generated from an unknown distribution with density f, the task is to estimate f. For simplicity, we consider only univariate density estimation.

In density estimation, one of the greatest difficulty occurs when the underlying density is rich in modal structures. Theorem 8.1, in its own right, can be utilized as a strategy for parametric density estimation using Gaussian mixtures. Besides density estimation via Gaussian mixtures, a popular approach is via the kernel density estimator, and is treated in detail in [*Scott*, 1992], [*Silverman*, 1986], [*Wand and Jones*, 1995]. The formula for the kernel density estimator, given data $\{X_1, \cdots X_n\}$ is

$$\hat{f}(x; b) = (nb)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/b\}\,, \tag{8.1}$$

taken directly from [*Wand and Jones*, 1995]. Usually K is chosen to be a unimodal density that is symmetric about zero, and is called the *kernel*. The positive number b is called the *bandwidth*. Such a formulation ensures that $\hat{f}(x; b)$ is also a density. One property of the kernel density estimator is that the choice bandwidth is more important than the choice of the kernel itself. The optimal choice of the bandwidth ensures that the density estimate becomes optimally smoothed. One popular choice of the bandwidth is

$$b = n^{-\frac{1}{5}} \hat{\sigma}\,, \tag{8.2}$$

where $\hat{\sigma}$ is the sample standard deviation of the given data and n denotes the sample size. One known problem of the bandwidth given in Eq (8.2) is that it works well for densities that are approximately symmetric unimodal. For multimodal densities, the bandwidth tends produce an oversmoothed density.

Here, we propose an algorithm using $\mathcal{M}$-decomposability to improve the kernel density estimator via the bandwidth given in Eq (8.2). As we are only dealing with the univariate case, we consider just the sorted data $F = \{X_{[1]}, \cdots X_{[n]}\}$. Similar to Section 8.1, we perform clustering of F via splitting and merging. In one-dimension, the splitting process becomes much simpler as we just have to find m $(2 < m < n-1)$ such that $(\sigma_G + \sigma_H)$ is minimized.

In general, the original data may be divided into many subclusters. For clarity of explanation, we assume that the original data F has two subclusters $G = \{X_{[1]}, \cdots X_{[m]}\}$ and $H = \{X_{[m+1]}, \cdots X_{[n]}\}$ after cluster analysis. By relaxing the strict condition of Gaussianity on Theorem 8.1 and applying a weaker condition of approximate symmetric unimodality instead, it is possible to use to improve the density estimation of F. We can expect the density estimation via the mixture of the decomposition pair to be better than that of the original data set, since $\sigma_G + \sigma_H < \sigma_F$. Therefore, one may propose an mixture kernel density estimator $\hat{f}_1$ of F given as follows:

$$\hat{f}_1(x) = \frac{m}{n}\hat{g}(x; b_g) + \frac{1-m}{n}\hat{h}(x; b_h)$$

where

$$b_g = m^{-\frac{1}{5}}\hat{\sigma}_G\,, \quad b_h = (n-m)^{-\frac{1}{5}}\hat{\sigma}_H\,,$$

and

$$\hat{g}(x; b_g) = (mb_g)^{-1}\sum_{i=1}^{m} K\{(x - X_{[i]})/b_g\}\,,$$

$$\hat{h}(x; b_h) = [(n-m)b_h]^{-1}\sum_{i=m+1}^{n} K\{(x - X_{[i]})/b_h\}\,.$$

The original kernel density estimator $\hat{f}$ of F is given in Eq (8.1).

As an experiment, we generate a sample of size 1000 from a bimodal density shown in black in Fig 8.9. By simply computing one single bandwith b on the whole sample set, we obtain a kernel density estimator (computed using $\hat{f}$). The result is shown in blue. By using $\mathcal{M}$-decomposability and splitting the data into two clusters, we obtain a mixture kernel density estimator (computed using $\hat{f}_1$). The result is shown in red. From Fig 8.9, it is clear that the kernel density estimator which is incorporated using $\mathcal{M}$-decomposability is nearer to the truth. As mentioned earlier, the effect of oversmoothing (blue line) is apparent in this example for the plain kernel density estimator with a single bandwidth. This is because the original density is bimodal with modes well separated. The undesirable effect of oversmoothing is alleviated by implementing $\mathcal{M}$-decomposability.

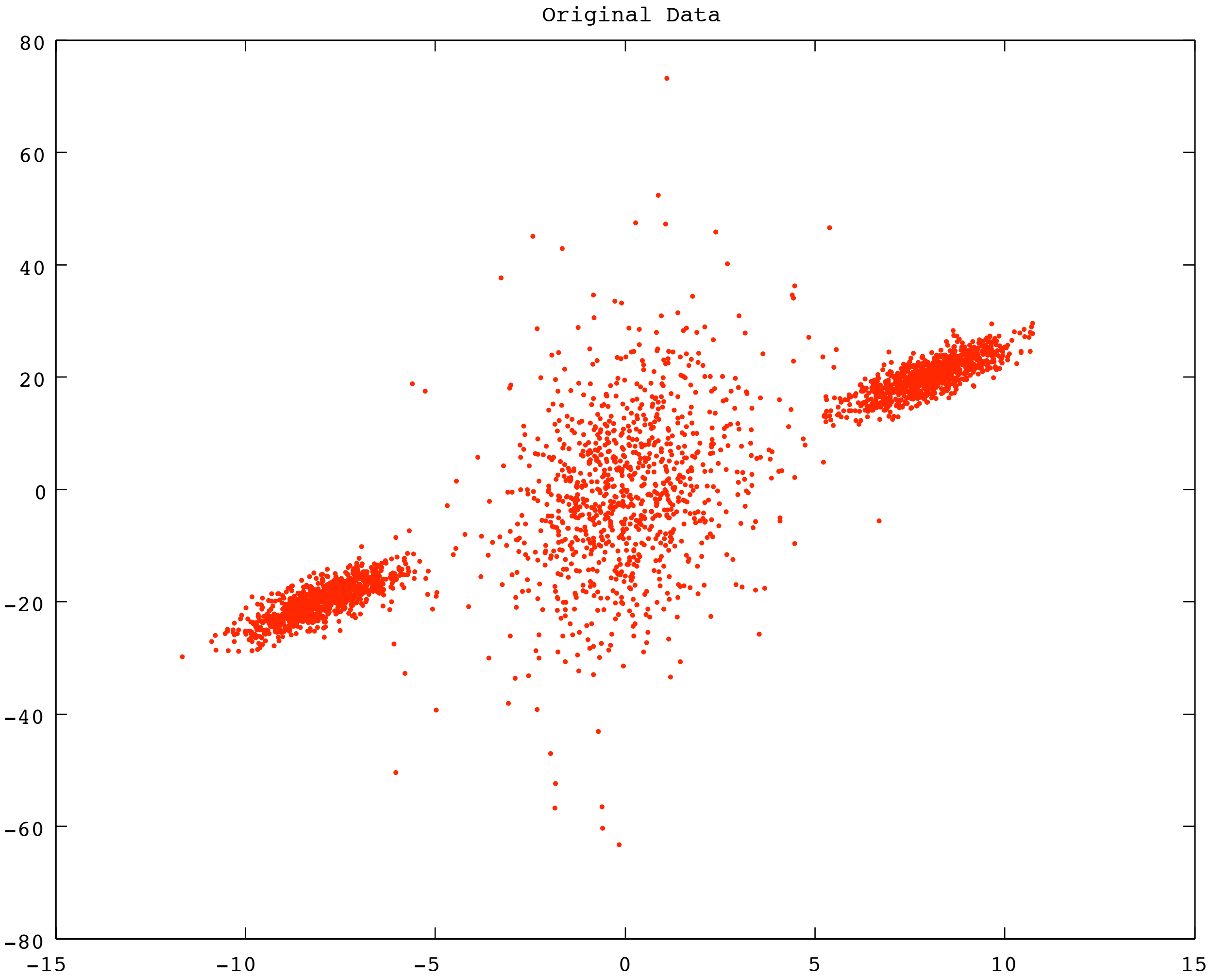

Figure 8.1. Original data from multimodal density.

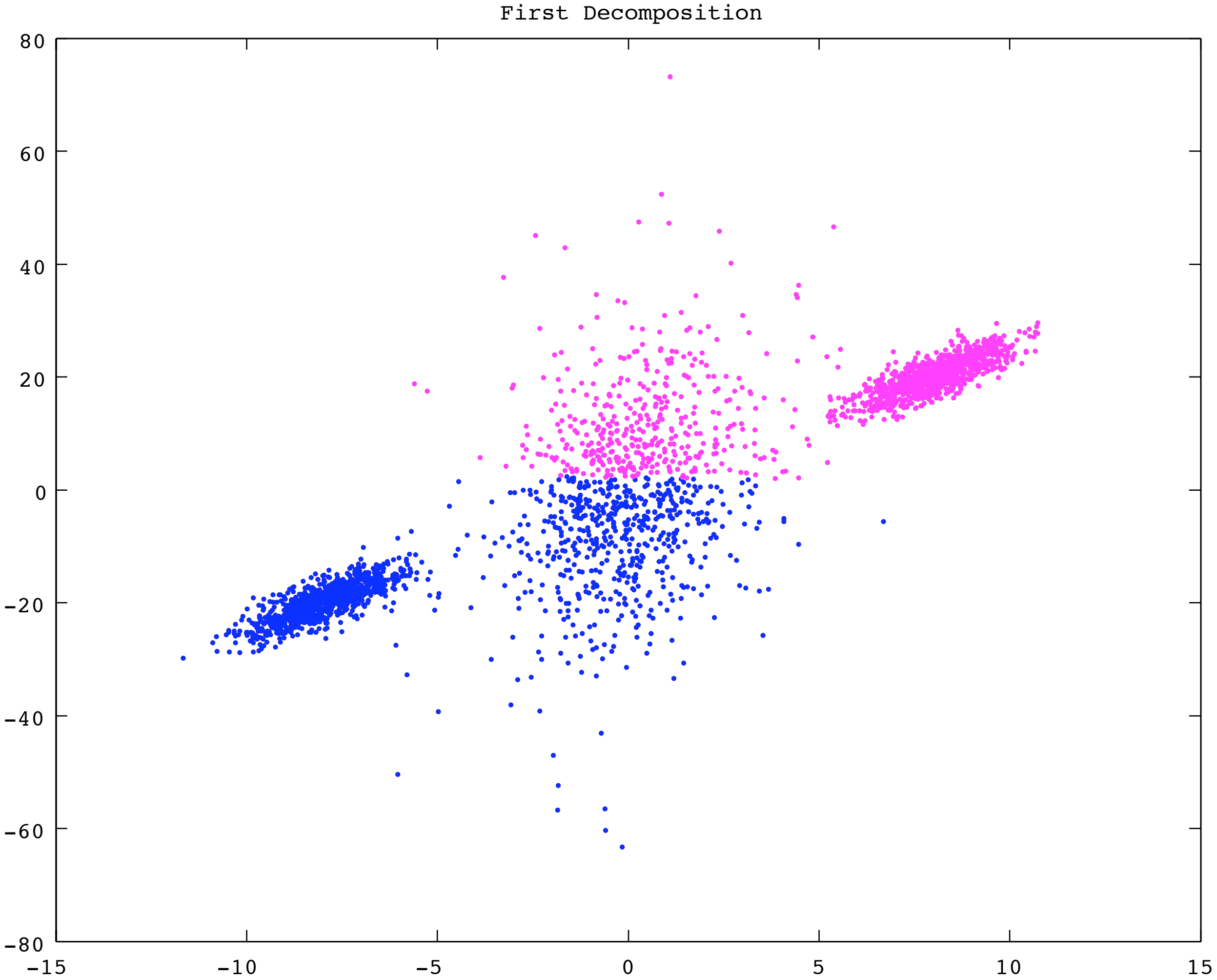

Figure 8.2. Decomposition pair of original density.

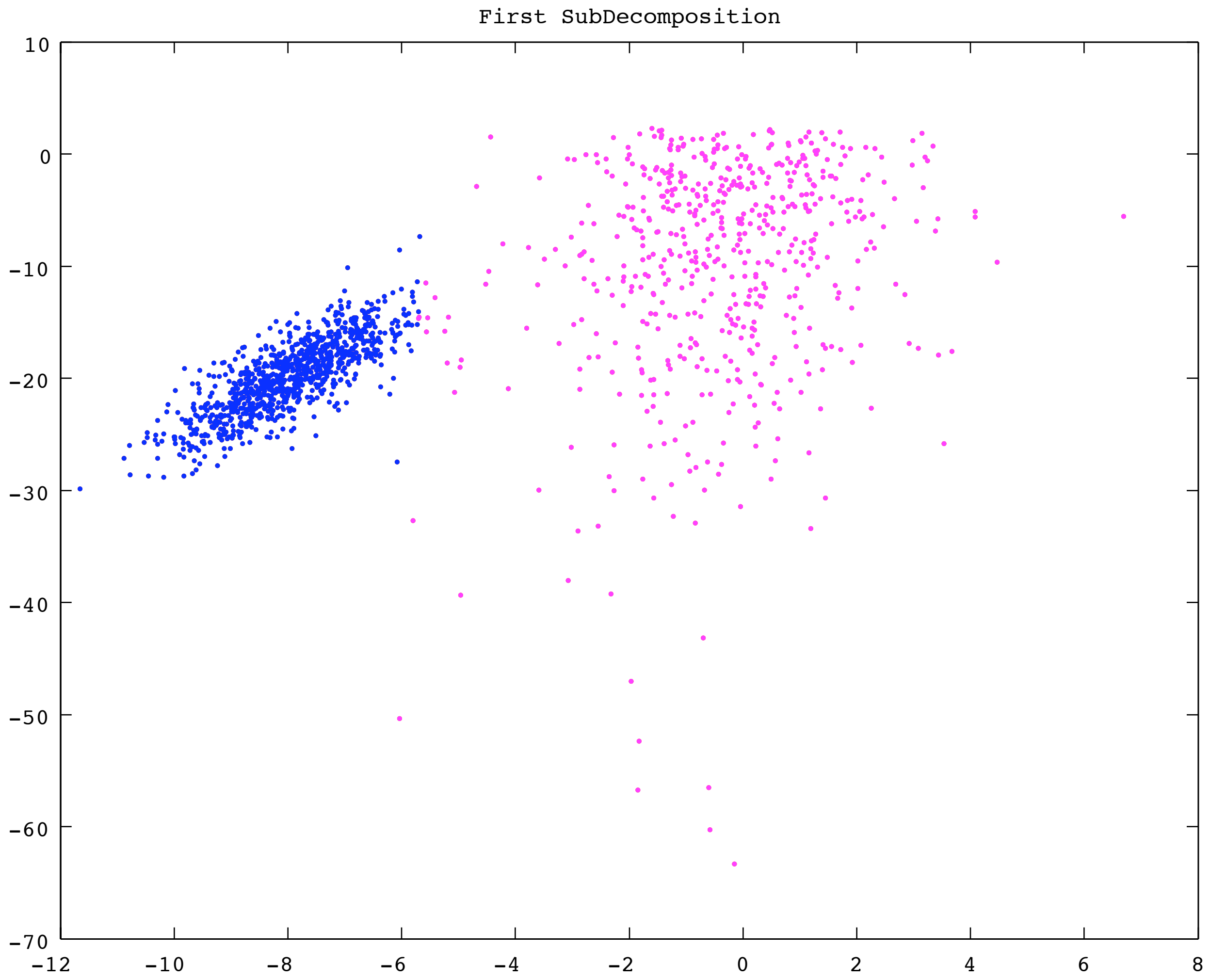

Figure 8.3. Decomposition pair of first mixture component from Fig 8.2.

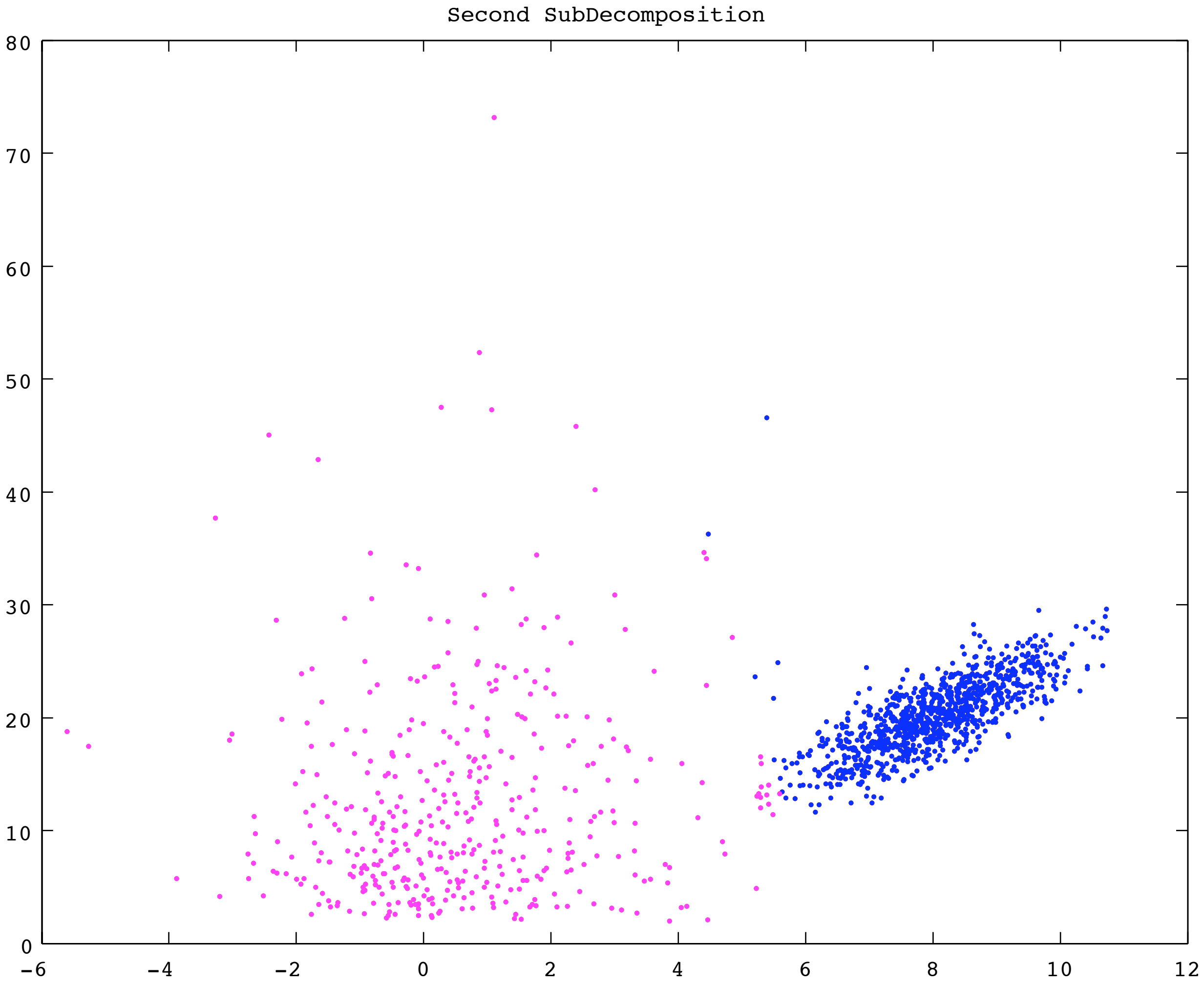

Figure 8.4. Decomposition pair of second mixture component from Fig 8.2.

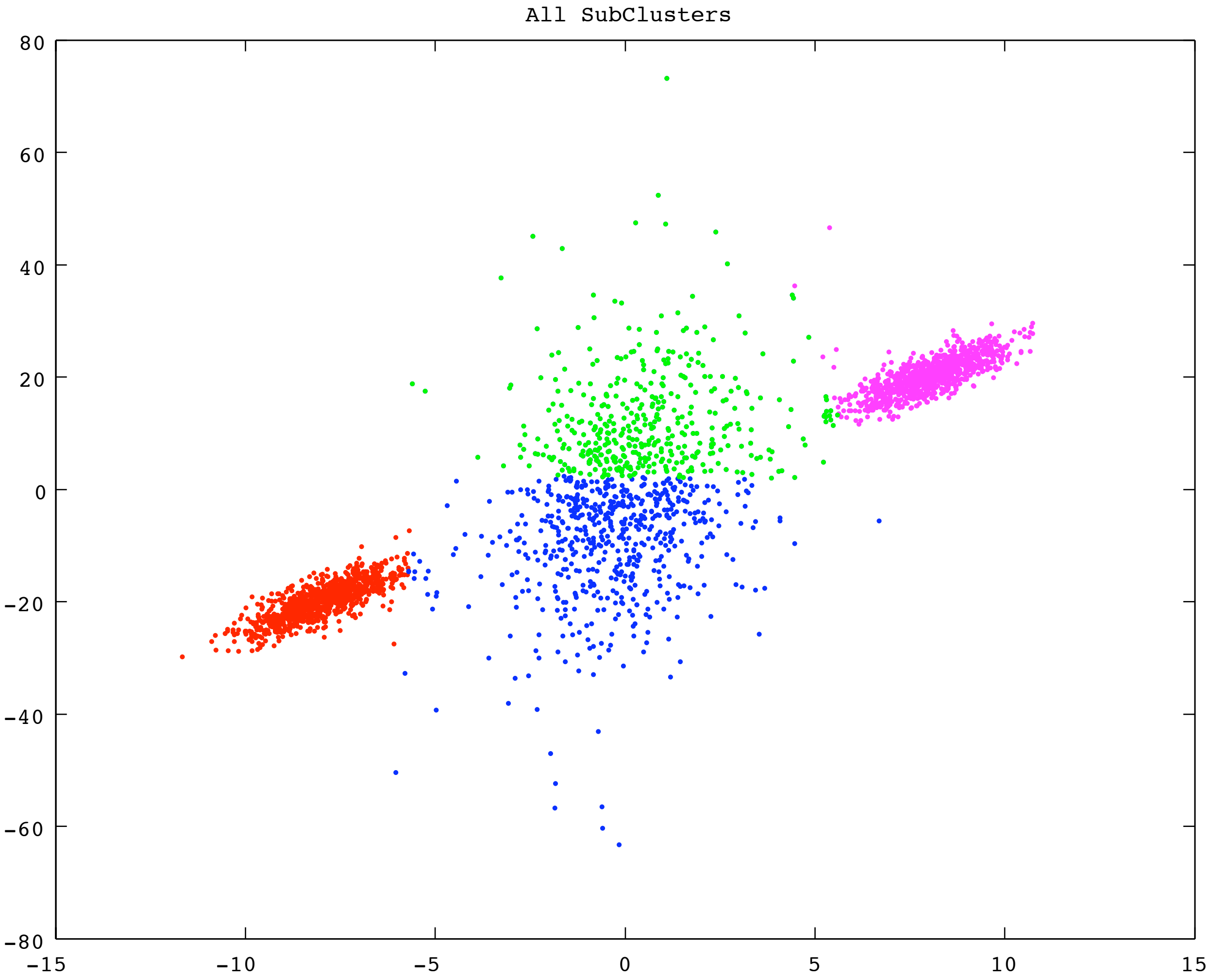

Figure 8.5. All $\mathcal{M}$-undecomposable subclusters.

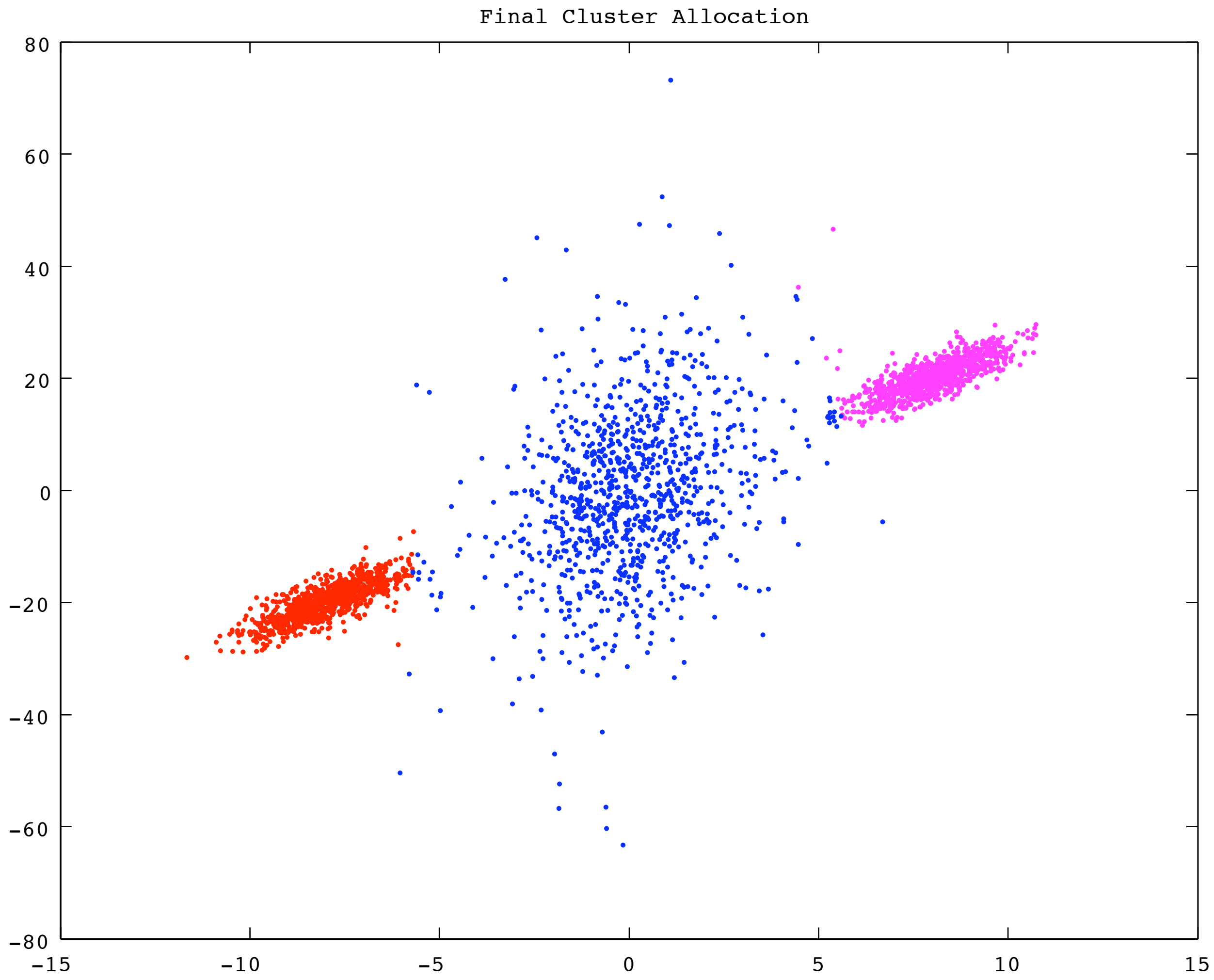

Figure 8.6. Final cluster allocation.

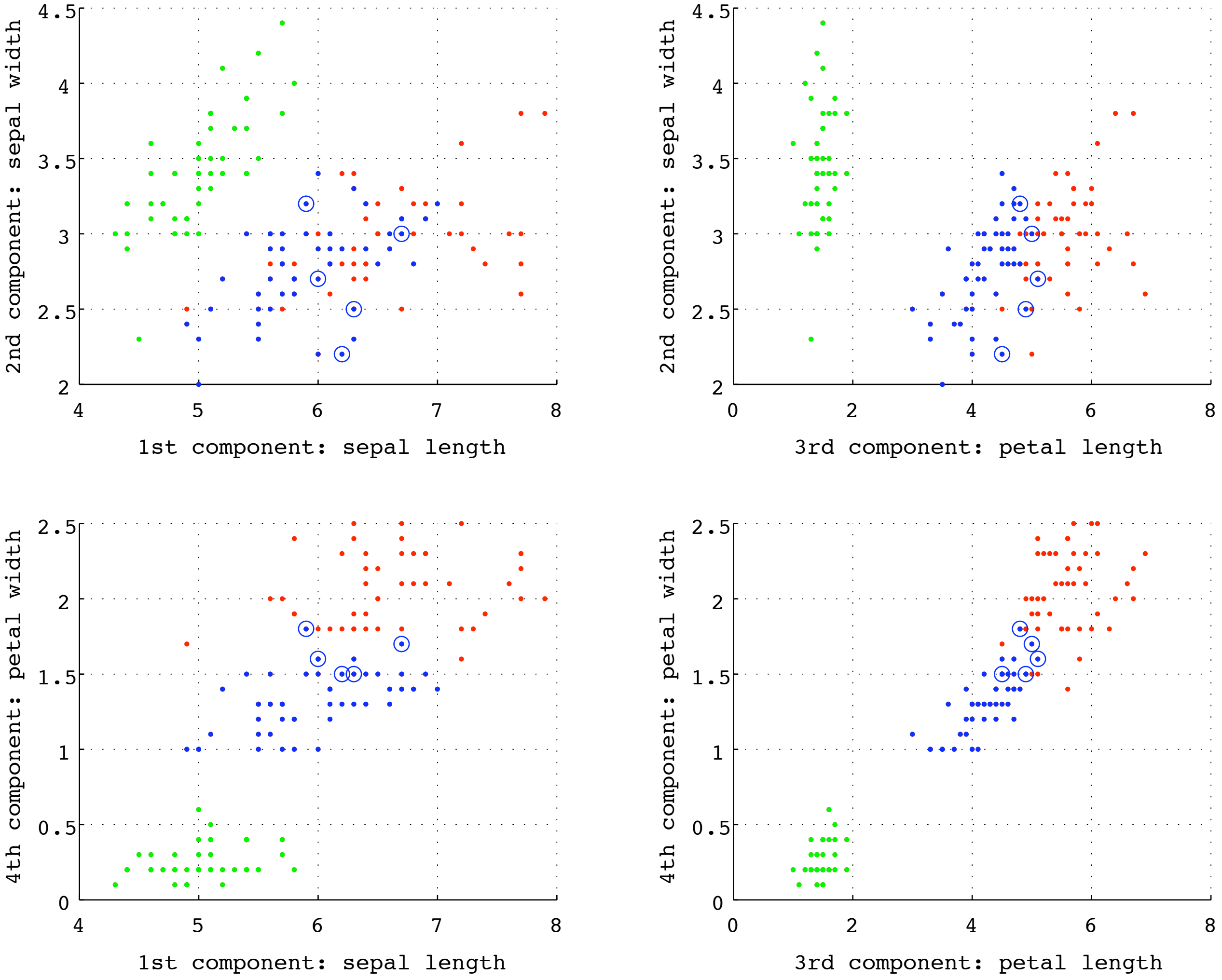

Figure 8.7. True Iris data: setosa(green), versicolor(blue), virginica(red). The circles denote data that are misspecified by $\mathcal{M}$-decomposability.

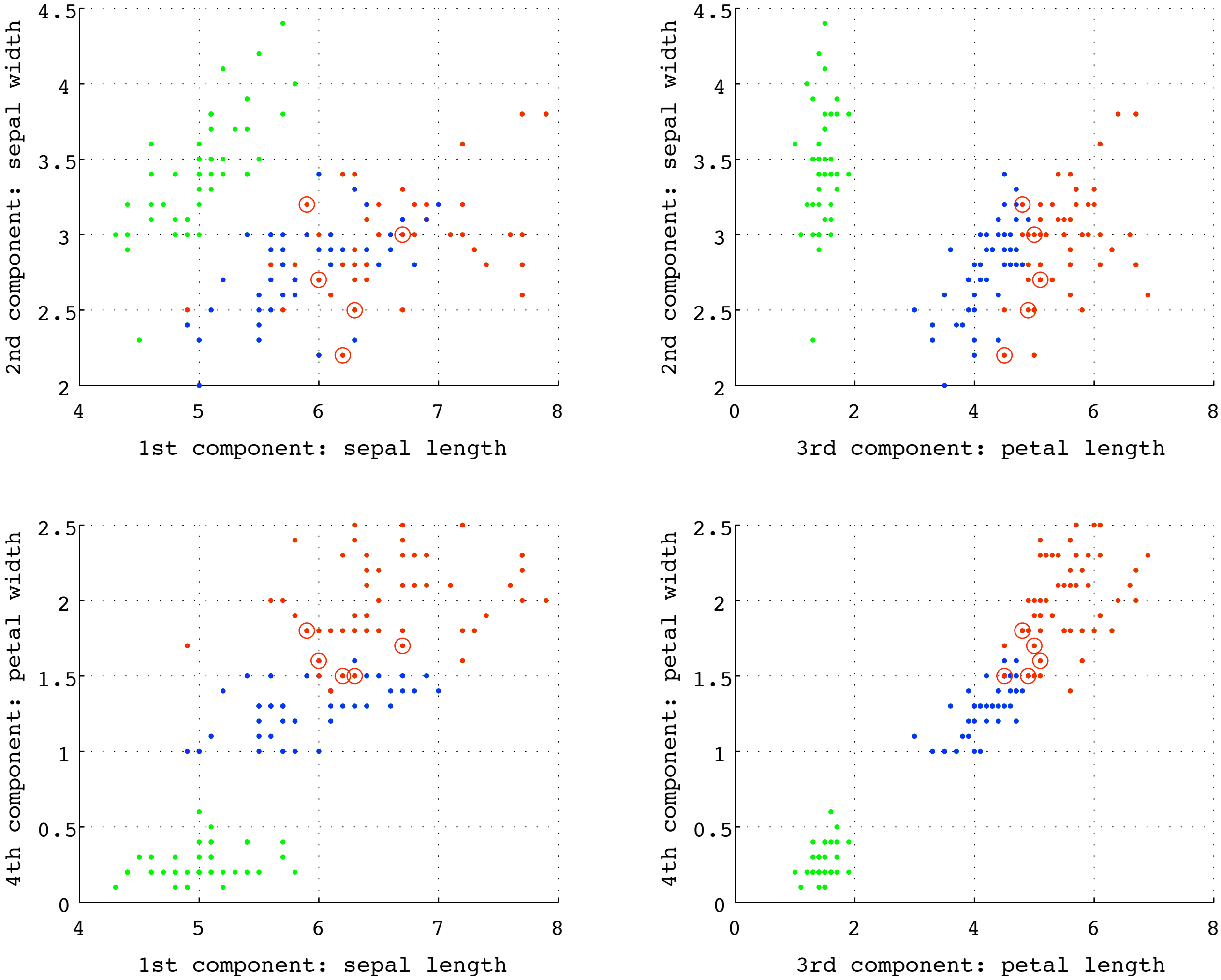

Figure 8.8. Iris data recovered via $\mathcal{M}$-decomposability: setosa(green), versicolor(blue), virginica(red). The circles denote data that are misspecified by $\mathcal{M}$-decomposability.

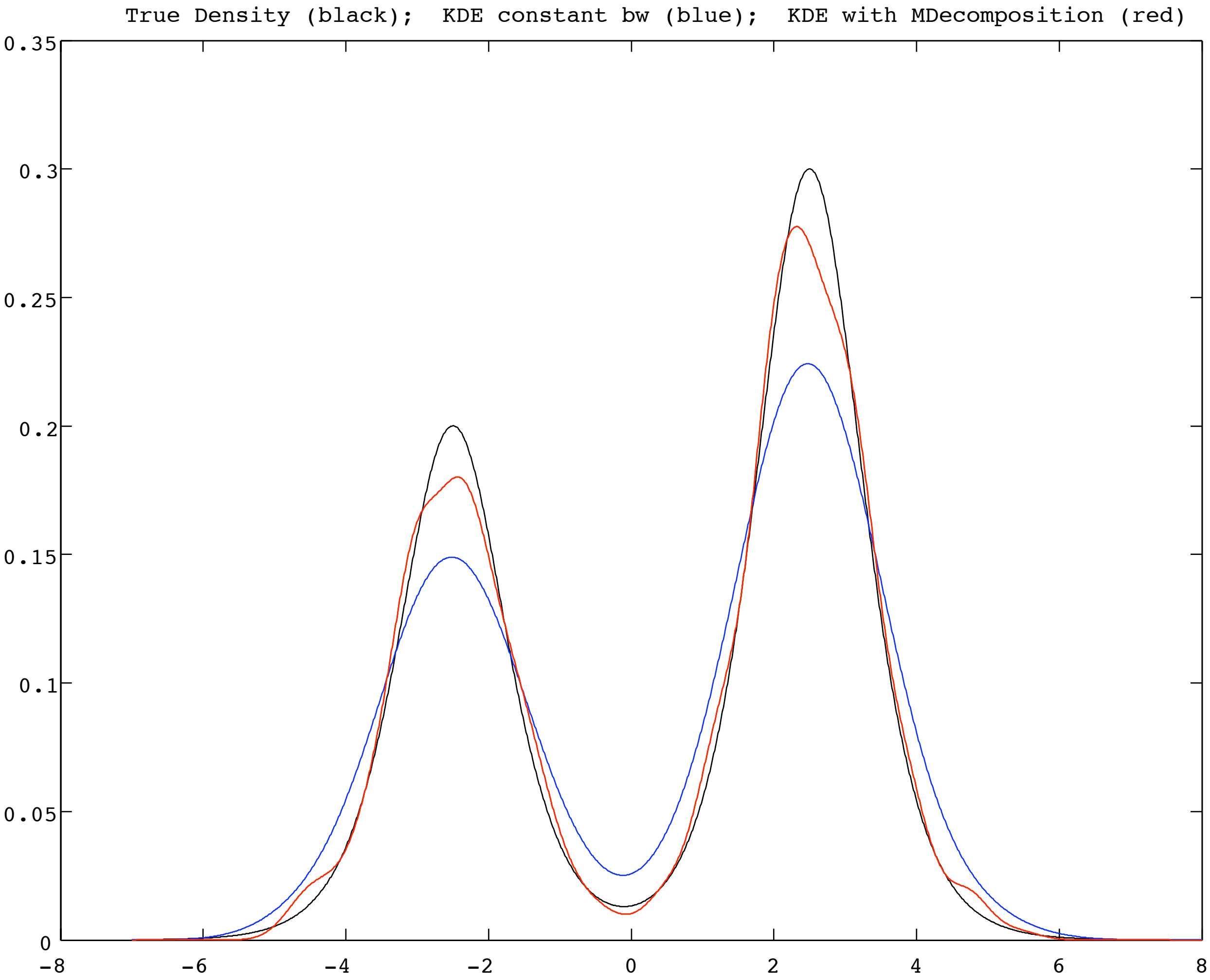

Figure 8.9. True density in black; kernel estimate with one bandwith in blue; kernel estimate with $\mathcal{M}$-decomposbility in red.

Chapter 9

Concluding Remarks

In this thesis, there are two main contributions. First of all, we introduce the idea of $\mathcal{M}$-decomposability, a novel concept that applies to probability density functions in *any* dimension. We build the theoretical foundations of $\mathcal{M}$-decomposability from ground up, from the very definition to the derivation of two theorems pertaining to $\mathcal{M}$-decomposability. As the concept of $\mathcal{M}$-decomposability is closely linked to the modality of probability density functions, the theoretical results attained from this thesis should be of interest to theoreticians and practitioners alike.

Secondly, we demonstrate the possible practical applications of the theoretical results derived from this thesis to statistical data analysis. One example of using the idea of $\mathcal{M}$-decomposability in practice is non-parametric clustering, whereby the determination of the number of underlying clusters is automatic. Another example is density estimation. We demonstrate a scheme for the refinement of kernel bandwidth to ameliorate kernel density estimation.

9.1 $\mathcal{M}$-Decomposability as a Criterion for Classification

In the thesis, we adopt a divide-and-conquer approach to analysis of probability density functions. Let f, g, h be probability density functions defined on $\mathbf{x} \in \mathcal{R}^d$. We say that $\{g, h\}$ is a decomposition pair of f if there exists $\alpha \in (0, 1)$ such that

$$f(\mathbf{x}) = \alpha\, g(\mathbf{x}) + (1 - \alpha)\, h(\mathbf{x})\,.$$

The following definition and results are derived from the thesis:

1. *According to Definition 6.2, a density f is said to be* $\mathcal{M}$**-decomposable** *if there exists a decomposition pair $\{g, h\}$ such that*

   $$|\Sigma_f|^{\frac{1}{2}} > |\Sigma_g|^{\frac{1}{2}} + |\Sigma_h|^{\frac{1}{2}}\,. \tag{9.1}$$

   Otherwise, we say that f is $\mathcal{M}$**-undecomposable***.*

   Note that there are no restrictions imposed on the functional forms of the mixture components g and h, as compared to parametric methods like finite mixture models. $\mathcal{M}$-decomposability can therefore be used as a flexible, non-parametric criterion for clustering or discriminant analysis. We then provide further evidence of the validity of $\mathcal{M}$-decomposability as a criterion for classification via the Theorem 7.2.

2. *All elliptical unimodal densities with finite second moments defined on $\mathbf{x} \in \mathcal{R}^d$ are M-undecomposable. (Theorem 7.2)*

   Theorem 7.2 applies to a wide class of commom probability density functions. These include and are not limited to Gaussian, Laplace, logistic, Student's t (with degrees of freedom greater than 2 to ensure the finiteness of second moments), Von Mises, elliptical uniform and many others, natural or artificial, mixture or otherwise. The class of elliptical unimodal densities (or *symmetric unimodal densities* in one-dimension) is much larger and more flexible than any parametric class and hence Theorem 7.2 can potentially have a far reaching impact both theoretically and practically. Theoretically, Theorem 7.2 implies a certain inherent fundamental optimality in elliptical unimodal densities, as opposed to densities which are $\mathcal{M}$-decomposable. In practical terms, if a density f is found to be $\mathcal{M}$-decomposable, one immediately concludes that f cannot belong to the class of elliptical unimodal densities. It is then probable that f is multimodal and should be subdivided into structurally simpler mixture components g and h. One natural, non-parametric criterion for cluster analysis or discrimination is to decompose a *multimodal* density into a mixture of elliptical unimodal densities. In this respect, $\mathcal{M}$-decomposability provides an intuitive solution to clustering.

The third result shown below, which relates $\mathcal{M}$-decomposable to Kullback-Leibler divergence, provides a stronger justification for $\mathcal{M}$-decomposability.

3. *Given any probability density function f, if there exists a decomposition pair $\{g, h\}$ such that equation (*9.1*) holds, then*

$$KLD[f\|\tilde{f}] > KLD[f\|\alpha\,\tilde{g} + (1-\alpha)\,\tilde{h}]\,.$$

Here, $KLD[\cdot\|\cdot]$ denotes the Kullback-Leibler divergence between the respective pdf's; $\tilde{f}$ denotes the Gaussian distribution which has the same mean and covariance structure as pdf f, the same applies to g and h. (Theorem 8.1)

Theorem 8.1 basically says that if an unknown pdf f is $\mathcal{M}$-decomposable, the mixture components g and h (which are apparently analytically unknown as well) can be used to improve estimation of f via Gaussian approximation. The mixture of the Gaussian approximates of the components, expressed in terms of $\alpha\,\tilde{g} + (1-\alpha)\,\tilde{h}$, provides a better estimation of the original density f than the

Gaussian approximate of f, in Kullback-Leibler sense. With Theorem 8.1, f no longer needs to be multimodal to benefit from the ideas of $\mathcal{M}$-decomposition. As long as f is $\mathcal{M}$-decomposable, it is guaranteed, in Kullback-Leibler sense, that one cannot go wrong with decomposition into the mixture components. Apart from cluster analysis, another straightforward application of Theorem 8.1 is *density estimation*. In Chapter 8, we demonstrate a strategy to improve kernel density estimation via $\mathcal{M}$-decomposability.

9.2 Future Work

The theoretical groundwork in this thesis has been shown to be directly applicable to cluster analysis as well as density estimation. As cluster analysis is also related to statistical learning, it is forseeable that further scientific and statistical applications of $\mathcal{M}$-decomposability may include principal component analysis, independent component analysis, machine learning, *etc.* Furthermore, as $\mathcal{M}$-decomposability has been demonstrated to improve density estimation, a direct application in this direction may be the improvement of particle filtering and MCMC methodologies.

Bibliography

Anderson, E., "The irises of the Gaspé peninsula", *Bulletin of the American Iris Society*, *59*, 2–5, 1935.

Anderson, T., "The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities", *Proc. Amer. Math. Soc*, *6*, 170–176, 1955.

Berkhin, P., "Survey Of clustering data mining techniques", *electronic version available at http://citeseer.nj.nec.com/berkhin02survey.html*, 2002.

Bernstein, D., *Matrix Mathematics*, Princeton University Press, Princeton, Oxford, 2005.

Cover, T., and J. Thomas, "Determinant inequalities via information theory", *SIAM J. Matrix Anal. Appl.*, *9*(3), July 1988, 384–392, 1988.

Dharmadhikari, S., and K. Joag-Dev, *Unimodality, Convexity and Applications*, Academic Press, New York, 1987.

Fang, K., S. Kotz, and K. Ng, *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London, 1990.

Farebrother, R., and I. Wrobel, "Regular and reflected rotation matrices", *IMAGE*, *29*, 24–25, 2002.

Fisher, R., "The use of multiple measurements in taxonomic problems", *Annual Eugenics*, *7*(2), 179–188, 1936.

Hardy, G., J. Littlewood, and G. Pòlya, *Inequalities*, second ed., Cambridge University Press, Cambridge, 1988.

Ibragimov, I., "On the composition of unimodal distributions", *Theor. Probability Appl.*, *1*, 255–260, 1956.

Kotz, S., C. Read, N. Balakrishnan, and B. Vidakovic, *Encyclopedia of Statistical Sciences*, 16 volume set, second ed., John Wiley and Sons, 2005.

McLachlan, G., and K. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

McLachlan, G., and D. Peel, *Finite Mixture Models*, Wiley Interscience, New York, 2000.

Newman, D., S. Hettich, C. Blake, and C. Merz, "UCI Repository of machine learning databases", *http://www.ics.uci.edu/∼mlearn/MLRepository.html*, 1998.

Pòlya, G., and G. Szegö, *Problems and Theorems in Analysis I, (English Edition)*, Springer-Verlag, Berlin, 1972.

Scott, D., *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley, New York, 1992.

Silverman, B., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

Wand, M., and M. Jones, *Kernel Smoothing*, Chapman and Hall, London, 1995.