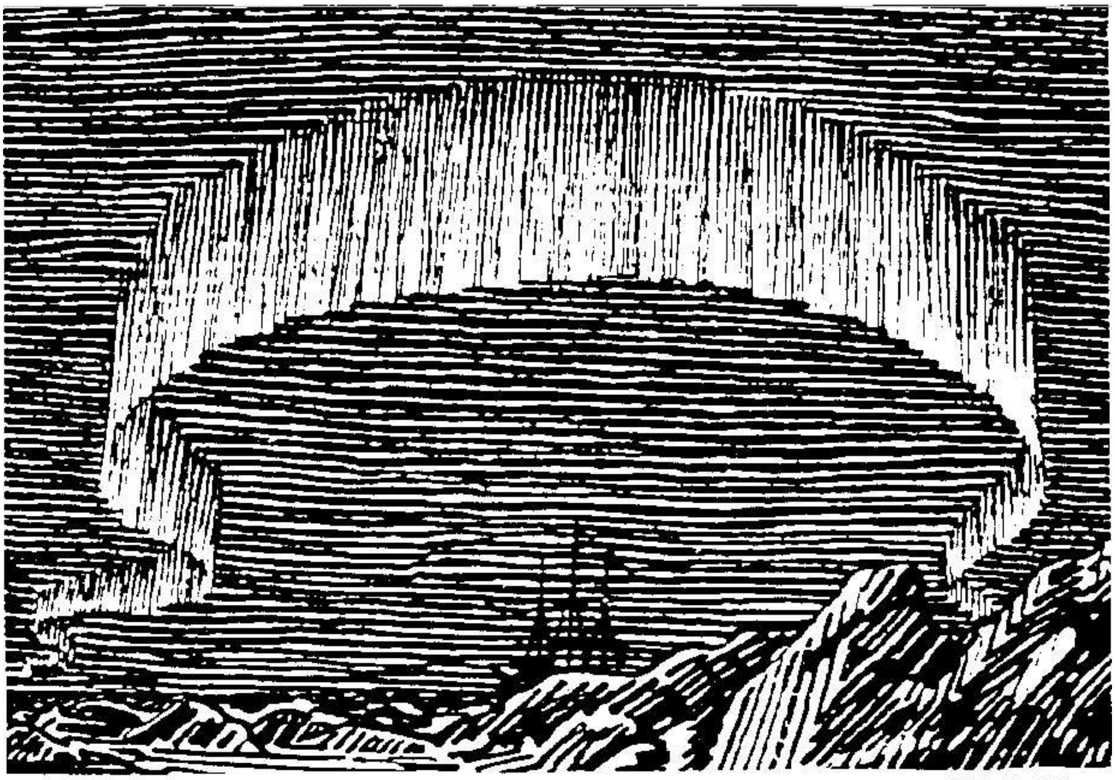
Master Thesis

Statistical Modelling of Polarimetric SAR Data

Anthony P. Doulgeris

Tromsø, May 2006



Original by F. Nansen

Faculty of Science
Department of Physics
University of Tromsø, N-9037 Tromsø, Norway, telephone: +47 77 64 51 50, fax no: +47 77 64 55 80

Abstract

This thesis discusses a statistical modelling technique to analyse polarimetric synthetic aperture radar (PolSAR) data. Polarimetric SAR data consists of four complex scattering coefficients at each image location, and must therefore be treated with multivariate modelling techniques. The work focuses on a simple class of multivariate distributions, the scale mixture of Gaussians, and in particular three parametric model families of that class (the multivariate Laplacian, K and normal inverse Gaussian distributions). They are chosen because closed form expressions have been derived for these models and because their general characteristics, i.e., sparse symmetric non-Gaussian distributions, have been observed in experimental SAR data. The commonly used multivariate Gaussian distribution is included as a reference. The primary aim is to investigate whether any particular model has advantages over the others, or the Gaussian, with a secondary aim of using the produced parametric features for image classification.

The models are characterised and parameter estimation methods, including optional constraints, are discussed and tested. Each of the models, plus the multivariate Gaussian, are inter-compared and assessed in terms of their goodness-of-fit to both simulated distributions and real PolSAR data. Goodness-of-fit methods are reviewed, and the Cauchy-Schwarz and Log-likelihood methods are tested in detail. The Log-likelihood method is found to be the preferred choice for this work. The modelling assessment concludes that a single flexible parametric model is sufficient to characterise the majority of the PolSAR data's statistical distributions. The multivariate normal inverse Gaussian distribution is the most representative of the studied models, closely followed by the multivariate K distribution.

Analysis of the PolSAR data with the single statistical model produces a feature set of two scalar parameters, a mean vector parameter and a covariance structure matrix parameter. The theoretically most relevant terms are extracted and combined to form a 5 dimensional feature set to be used for clustering and unsupervised segmentation. Feature space investigations indicate functional relations between the features and transforms are determined (experimentally) to simplify the data set into a more linearly spread feature space. The transformed data is then suitable to utilise very simple clustering techniques, of which the k-means and discrete mixture of Gaussians methods are applied here. The transformations effectively convert the two scalar model parameters into the model independent statistical measures of width and non-Gaussianity. The interpretation of the new feature space indicates that it is essentially a multivariate Gaussian analysis with the one additional feature of non-Gaussianity.

Segmented image maps, with and without certain options are inter-compared, however, their effectiveness cannot be rigourously determined due to inadequate ground truth data. Initial indications show that the method produces a realistically segmented image, and the main class features are visually consistent with a coarse land cover map of the same area.

Acknowledgements

I would like to acknowledge Professor Torbjørn Eltoft as my supervisor and guide in producing this work. My starting point was Torbjørn's theoretical work on the three models studied, and I have used (or perhaps abused) and built upon his project ideas and methods throughout this research. Torbjørn emphasises a collaborative style, and I appreciate the time that I was given to present and discuss my results and views.

I am grateful all members of the research group, for their patience, encouragement and advice at our informal meetings, which were a good forum for sounding out ideas and learning new techniques. Stian Anfinsen, Robert Jenssen and Tom Rune Lauknes have all contributed to my understanding of this field and how to do and present research. Thanks also for the great deal of time spent proof reading and commenting on my work, as that is the best way for me to really learn how to write.

Thanks go to the team at Norut IT, Tromsø, that supplied the original PolSAR data and land cover maps. Kjell Arild Høgda, Bernt Johansen, Harald Johnsen and Inge Lauknes have all helped me at some stage of this work and their time and effort is appreciated.

I would also like to thank all the students and teachers from the classes that I attended as all have made my comeback to university life a most enjoyable experience. And most of all, I can not thank Corine Davids enough for her encouragement, interest, understanding and support throughout this lengthy endeavour.

Thanks to you all.

Anthony Doulgeris
Tromsø, Norway, May 2006.

Contents

Chapter 1

Introduction

1.1 Introduction

Synthetic aperture radar (SAR) is widely used for monitoring and imaging the Earth's surface. Radar wavelengths have good atmospheric and cloud penetrating characteristics, making it a useful all weather observation system, and the synthetic aperture technique allows for useful high resolution imaging even from satellite-borne systems. Recent fully polarimetric SAR (PolSAR) systems provide a more complete description of the backscatter behaviour of the target surface, with the potential to improve the discriminating power for remote sensing purposes.

The distributed and often rough textured nature of remotely sensed targets, necessitates the use of a statistical approach and interest lies in the statistical properties of collections of target scattering measurements. The multivariate Gaussian distribution has often been utilised in modelling polarimetric data, and agrees reasonably well with observations of homogeneous regions with coarse spatial resolution. However, for fine spatial resolutions or heterogeneous backscattering media, the observations are distinctly non-Gaussian in nature [1]. Several distributions have been investigated and the K-distribution [1] has proved particularly useful in characterising the amplitude distribution of radar backscatter signals from diverse natural surfaces.

The K-distribution model has since been generalised to treat polarimetric SAR signatures [2], accounting for the covariance of the polarisation channels. The multivariate K distribution can be interpreted as a product model of a gamma distributed scalar variable and a multivariate Gaussian variable. This product model has also been generalised with differing scaling distributions to derive alternative non-Gaussian distributions [3].

This study investigates a simple class of multivariate scale mixtures of Gaussians models as non-Gaussian models for characterising polarimetric SAR data. The motivation is to implement the three scale mixture models derived in [3], and investigate whether any particular model has advantages over the others. To facilitate this comparison, a method of determining a multivariate goodness-of-fit must be established, and this is investigated in some detail. Each model's goodness-of-fit are compared to one another for a test PolSAR data set, and interpretations are discussed regarding both the particular models and the significance for the PolSAR data set.

Once the parametric modelling technique is implemented, it must be tested to see whether the feature set thus produced is meaningful and/or beneficial. The emphasis here is for land cover classification, and some basic unsupervised segmentation methods are applied to the modelled data set. Initial results are tested by visual comparison to an existing land cover map derived from Landsat 5 TM data.

The modelling is investigated with the objective of practical image segmentation and the goodness-of-fit testing is aiming simply to fit real empirical data. Scattering theory and the theoretical foundations of each model are not debated here, and potential theoretical implications of the modelling results are not investigated. The entire field of polarimetric decomposition theorems is likewise beyond the scope of this study, although incorporating simple decompositions (e.g. [4]) into the modelling could be a worthwhile future exercise. The focus here is this one simple class of models and whether they can be useful for analysing PolSAR data.

Although this study works solely with a PolSAR data set, the modelling methods are not specifically linked to this data type. The scale mixture of Gaussians class could equally well be applied to any multivariate data set, of any dimensionality, where distributions with correlations are involved. The model class studied here was the symmetric case of the generalised mean and variance scale mixture of Gaussians class [5], and perhaps these methods could be extended to the general asymmetric case.

1.2 Structure of thesis

Following this introduction is a brief overview of polarimetric synthetic aperture radar and scattering theory. The main thesis is divided into two parts. The first part deals solely with the modelling and the second part deals with the image segmentation of the modelled PolSAR data.

In Part I, Chapter 3 introduces the scale mixture of Gaussians class, the particular parametric families to be studied, and discusses their basic derivation and properties. Chapter 4 covers the different methods of parameter estimation, from the initial iterative implementation to the faster moment based implementation. The different methods are tested for accuracy and variance with simulated data sets, including the application of certain constraints. Chapter 5 briefly discusses the concept of mixtures of several models. Chapter 6 reviews and investigates goodness-of-fit testing and decides on the best approach for the required task. Chapter 7 discusses the fitting of the models against real PolSAR data, and Chapter 8 sums up all the concepts from Part I, relating to modelling, best-fitting and the PolSAR data set.

Part II, begins by introducing the comparison ground truth data in Chapter 9 and classification methods in Chapter 10. Chapter 11 investigates the modelled parameter feature space, with some interesting observations, and Chapter 12 shows preliminary segmentation results, with key observations summarised in Chapter 13.

Chapter 14 concludes the current study, stating the main observations and the many paths yet to be investigated.

Chapter 2

Polarimetric SAR Data

A detailed introduction to synthetic aperture radar and polarimetry can be found in text books such as [6], [7] and [8]. A few relevant concepts are described in this chapter to give an introduction to the PolSAR data set used in this study.

2.1 RADAR, SAR and PolSAR

Radar imaging systems are active electromagnetic systems operating at microwave wavelengths, and for earth observation are usually implemented as air-borne or more recently as satellite-borne systems. They transmit a short microwave pulse down to the target surface and measure the signal that is scattered back to the antenna, using signal processing to build up an image of the illuminated area as the antenna moves along. The systems can operate in many different frequency bands, for example this study encounters the C-band, at 5.6 GHz (0.053 m wavelength), and the L-band, at 2.4 GHz (0.125 m).

Side looking air-borne radar (SLAR) systems are real-aperture radar (RAR) systems that achieve a useful high spatial resolution in the direction perpendicular to the flight path (range) by sending a very short pulse length, and in the flight direction (azimuth) by using as large an antenna as possible. The extra distance to target of satellite systems would require impracticably large antennae to achieve the same resolution as airborne systems. Synthetic aperture radar (SAR) systems achieve finer azimuth resolution, for their given real antenna size, by advanced signal processing. The process coherently combines a sequence of signals from pulses transmitted as the antenna moves along its flight path, thereby synthesising a much larger aperture (antenna). The processing may produce a single combined image with the highest azimuth resolution (single-look data) or be processed into several images of lesser azimuth resolution (multi-look data).

Antennae are designed to produce a polarised electromagnetic wave orientated with the antenna, and early radar systems used a single antenna in a fixed orientation for both send and receive. They measured the signal strength of only one polarisation aspect of the target reflectance, for example horizontal send-horizontal receive (HH). More recent polarimetric systems have two antennae and can send and receive both horizontally and vertically, measuring all combinations of orientations, HH, HV, VH, and VV, thereby defining the response of any arbitrary polarisation. Modern electronic systems are also capable of measuring the magnitude and phase of the returned signal, rather than just its amplitude, and therefore are able to fully record the scattering matrix as four complex valued components. The term PolSAR refers to fully polarimetric synthetic aperture radar systems, and in this study the PolSAR data set is the 4 complex scattering matrix components from a single-look image.

SAR systems require preservation of phase between pulses and accurate position monitoring for correct compensations. Polarimetric measurements may also suffer from cross-talk and channel imbalance distortions. It is assumed that all of the system calibrations and corrections have been made for the data sets used in this study.

2.2 Wave Polarisation

The polarisation of a plane electro-magnetic wave describes the orientation of the electric field as a function of time. In the general case, the locus of the E-field in a plane perpendicular to the direction of propagation is an ellipse, with special cases for linear and circular polarisation. Any E-field can be completely described by two orthogonal components, usually referred to as vertical and horizontal, with their proportions and relative phase determining all variations of polarisation. There are many different representations and co-ordinate systems for describing waves. The following descriptions use the vertical-horizontal basis and the backscatter alignment (BSA) convention, which is most relevant to bistatic systems where the send and receive antennae are at the same location.

When a transmitted plane wave ($\mathbf{E}^t$) interacts with materials it may be reflected in different proportions horizontally and vertically and therefore change the polarisation properties of the wave. This scattering process can be described by means of a scattering matrix ($\mathbf{S}$) transformation. That is, the transmitted wave is defined as

$$\mathbf{E}^t = \begin{bmatrix} E_v^t \\ E_h^t \end{bmatrix} e^{ik_0 r}, \tag{2.1}$$

where E_v^t and E_h^t are complex valued wave representations of magnitude and phase of the form $E = a\ e^{-i\phi}$, r is the distance between the scatterer and the receiving antenna and k_0 is the wavenumber of the illuminating wave. The returned wave, transformed by the scattering process, is

$$\mathbf{E}^r = \frac{e^{ik_0 r}}{r}\,\mathbf{S}\,\mathbf{E}^t, \tag{2.2}$$

which in expanded form becomes

$$\begin{bmatrix} E_v^r \\ E_h^r \end{bmatrix} = \frac{e^{ik_0 r}}{r} \begin{bmatrix} S_{vv} & S_{vh} \\ S_{hv} & S_{hh} \end{bmatrix} \begin{bmatrix} E_v^t \\ E_h^t \end{bmatrix}. \tag{2.3}$$

The elements of the scattering matrix, **S**, are known as the complex scattering amplitudes and describe how the scatterer transforms the polarisation of the incident wave. Each scattering amplitude may be a function of frequency, the illuminating angle, and the orientation of the scatterer relative to the co-ordinate system.

2.3 Reciprocity

The theorem of reciprocity states that the two cross-polar terms are equal, that is

$$S_{vh} = S_{hv} \equiv S_x, \tag{2.4}$$

for targets whose internal state is unaltered by the polarisation of the probing wave. This is expected to be the case for most naturally occurring scatterers. However, real data may not always obey the reciprocity theorem exactly due to statistical fluctuations and measurement errors.

The cross-polar term, S_x, is often taken as the average of S_{vh} and S_{hv}, multiplied by $\sqrt{2}$ to conserve the total power in the vector, and is performed to reduce the statistical non-equality found in real data. For statistical modelling there are two comments against this approach. Firstly, the averaging may affect the statistical distribution in more ways than just the width (intensity). A simple example is that combining two uniform pulse distributions produces a triangular pulse of twice the width. The width is corrected for with the $\sqrt{2}$ term, however the shape has been changed by the averaging. Secondly, although both cross-polar terms hold the same information, they are two statistical samples of that distribution. The statistical estimation methods will achieve better results by using both samples to estimate their properties. If processing time is an issue with the extra dimensions, then use one term only.

2.4 Scattering Theory

The mathematical description discussed above refers to individual scattered waves and point targets, however the signal measured by the radar system will be the resulting signal of many individual scattered waves over a distributed target area or volume. A natural target will not be a single pure scatterer and the target area may have a significant surface roughness or volume scattering component. The measured scattering matrix will therefore represent a statistical property of the target location.

This study considers only the case of incoherent scattering, which is usually the case for the natural environment. Incoherent scattering assumes that the signal for a given target location is comprised of returns from an unknown number of individual waves, whose magnitude and phase are determined by the many individual point conditions, such as distance to target, surface angle, material type and polarising orientation. The vector sum of all such waves is the resulting signal measured by the antenna, and can be modelled as a two dimensional random walk in the (v, h) basis. It is assumed that the target area is sufficiently large and textured, relative to the illuminating wavelength, so that the individual returns can be considered independent and the phase of the vector sum can be considered uniformly random. The central limit theorem implies that when the number of individual scattering points per resolution cell is very large and the scattering medium is homogeneous, then the scattering process would be Gaussian distributed. If the scattering medium is not spatially homogeneous, or the resolution cell is not sufficiently large that the central limit theorem applies, then the distribution may be non-Gaussian [9].

In practice, the amplitude will be a measure of the average target reflectance, but the random phase means that this value is evenly shared between the real and imaginary returned signal. That is, the magnitude and phase will have equal average intensity, and the indivdual values will be uncorrelated. Additionally, being magnitude and phase of the complex vector E-field signal, implies that they should both measure plus or minus values centred around a mean of zero.

2.5 Statistical Sampling

In order to apply statistical analysis methods the data must be grouped into a sample set that is assumed to represent the target area. Statistical sampling may be achieved by using either many multi-look processed images, where the sampling at every image location comes from several images, or by sampling a local neighbourhood of pixels from a single-look processed image.

Both techniques result in some loss of resolution relative to the maximum resolution of the single-look processed image. Multi-look processing produced several images with a lower azimuthal resolution for each pixel, and neighbourhood averaging results in a bi-directional smoothing over the single highest resolution processed image. This study looks only at the neighbourhood sampling method because the original data was supplied as a single-look complex image.

2.6 Data format

The single-look complex image data represents the complex scattering matrix at each pixel location. These four complex numbers can be written as a 4-D complex vector or, by splitting the real and imaginary parts, an 8-D real vector,

$$\begin{pmatrix} S_{vv} & S_{vh} \\ S_{hv} & S_{hh} \end{pmatrix} \Rightarrow \begin{pmatrix} S_{vv} \\ S_{vh} \\ S_{hv} \\ S_{hh} \end{pmatrix} \Rightarrow \begin{pmatrix} \Re\{S_{vv}\} \\ \Im\{S_{vv}\} \\ \Re\{S_{vh}\} \\ \Im\{S_{vh}\} \\ \Re\{S_{hv}\} \\ \Im\{S_{hv}\} \\ \Re\{S_{hh}\} \\ \Im\{S_{hh}\} \end{pmatrix} \equiv \mathbf{S}_R. \tag{2.5}$$

Each image location is then statistically represented by a neighbourhood collection of such point vectors such that the image point (x, y) has the statistical sample set $\mathbf{D}_{x,y}$ representing it, and

$$\mathbf{D}_{x,y} = \{\mathbf{S}_{R,i}\} \qquad \text{for samples } i = 1, \ldots, N, \tag{2.6}$$

where the samples are taken from the image vectors from the region $(x \pm \Delta, y \pm \Delta)$, and Δ is the neighbourhood size. N is therefore equal to $(2\Delta + 1)^2$, and there will be some overlap in the set representations from one pixel to the next. In practise, the outer border of width Δ cannot be sampled, so the analysed image will be 2Δ smaller, in each direction, than the orignal image.

The statistical distribution of the multi-dimensional data set at each point, $\mathbf{D}_{x,y}$, is then analysed using multivariate modelling.

Part I

Parametric Modelling

Chapter 3

Parametric Statistical Models

3.1 The Models

A parametric statistical model is an explicit formula, with one or more parameters, for a probability density function. Different values of the parameters result in different curves, thus a given model function describes a family of curves. Sometimes the parameters relate to some simple aspect such as width or shape, but often they are inter-related in complex ways. There are often constraints upon the allowable values of the parameters, such as a requirement to be nonnegative or real valued, and there are also the overall constraints of probability density functions, i.e., that the function is always nonnegative and integrates to 1. If the model function's domain is a multi-dimensional space then it is known as multivariate.

Gaussian statistics have frequently been assumed for radar return signals, however abundant experimental evidence indicates that the signal is often not Gaussian [1, 2]. This study investigates four multivariate statistical models: the Gaussian, Laplacian, K-distribution and normal inverse Gaussian. The multivariate Gaussian is the simplest and most commonly used model in statistics. The other non-Gaussian models studied are all derived from the multivariate extension of the scale mixtures of Gaussians method [10] and are described in [3]. Figure 3.1 of Section 3.1.5 shows 1 dimensional examples of each distribution as an aid for interpretation. They are all obtained as a product of a random variable scale factor and a multivariate Gaussian distribution. By separating the internal covariance structure from the Gaussian variable we can obtain a general form of the multivariate scale mixture of Gaussians distributed variable $\mathbf{Y}$, as

$$\text{scale mixture:} \qquad \mathbf{Y} = \boldsymbol{\mu} + \sqrt{Z}\,\boldsymbol{\Gamma}^{\frac{1}{2}}\,\mathbf{X}, \tag{3.1}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Gamma}$ is the covariance structure matrix with determinant 1, Z is the scale random variable, and $\mathbf{X}$ is a standard multivariate normal variable with zero mean and unit (identity) variance. The scale Z must be positive only and real, such that $\sqrt{Z}$ is a positive scalar factor. Note that the formulation used here is a simplification of the more general normal variance-mean mixture model, proposed in [5], given by

$$\text{normal variance-mean mixture:} \quad \mathbf{Y} = \boldsymbol{\mu} + Z\boldsymbol{\beta} + \sqrt{Z}\,\boldsymbol{\Gamma}^{\frac{1}{2}}\,\mathbf{X}. \tag{3.2}$$

The simplification that $\boldsymbol{\beta} = \mathbf{0}$, restricts the general form to the special case of semi-symmetric models, i.e., each dimension is independently symmetric around its mean value yet may include correlation between dimensions.

Clearly, the conditional probability density function (pdf) of $\mathbf{Y}$ given Z must be a multivariate Gaussian and can be expressed as

$$f_{\mathbf{Y}|Z}(\mathbf{y}|z) = \frac{1}{(2\pi z)^{\frac{d}{2}}}\ \exp\left[-\frac{1}{2z}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right], \tag{3.3}$$

where d is the data dimension.

Given a pdf for Z, $f_Z(z)$, and the mixture method (3.1), then the marginal pdf of $\mathbf{Y}$ is obtained by integrating the conditional pdf (3.3), over the prior distribution of Z. That is

$$f_{\mathbf{Y}}(\mathbf{y}) = \int f_{\mathbf{Y}|Z}(\mathbf{y}|z)\ f_Z(z)\ dz. \tag{3.4}$$

The different models are obtained by taking Z from different random distributions. The observed non-Gaussian behaviour of SAR data tends towards the super Gaussians or sparse distributions, i.e., heavier tailed than the Gaussian. This observation together with theoretical considerations, such as Brownian motion time of passage, motivated the choice of scale distributions from Exponential, Gamma and inverse Gaussian distributions [11, 12].

In each case the marginal integral is solved by recognising that it has the form of a moment of an inverse Gaussian (IG) integral and substituting terms into the known solution for inverse Gaussian moments [11, 5].

For simplicity, $q(\mathbf{y})$ is defined as

$$q(\mathbf{y}) = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.5}$$

in the following sections.

3.1.1 Multivariate Gaussian

The well known multivariate Gaussian density distribution is given by the formula

$$\text{MG:} \qquad f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}}\ \exp\biggl(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\biggr), \tag{3.6}$$

where the parameters are the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. d is the space dimension, $|\cdot|$ represents the matrix determinant operator, and $(\cdot)^T$ denotes transposition. Note that as a probability density function, $f_X(\mathbf{x})$ is always nonnegative, and it integrates to 1. The normal Gaussian distribution is sometimes written as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

When the mean is zero and the covariance is the identity matrix ($\mathcal{I}$) then

$$\mathcal{N}(\mathbf{0}, \mathcal{I}) = \frac{1}{\sqrt{(2\pi)^d}} \exp\Big(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\Big), \tag{3.7}$$

and is symmetric around zero, and equally scaled in each dimension.

Note that the standard multivariate Gaussian distribution fits into the scale mixture of Gaussian generation method by considering the scale parameter's pdf as a Dirac delta distribution, that is, Z has a constant value.

3.1.2 Multivariate Laplacian

The one dimensional Laplacian distribution is also known as the double exponential distribution because of its characteristic shape (see Figure 3.1). The multivariate Laplacian distribution is obtained from the multivariate extension of the scale mixture of Gaussians model (3.1), by taking the scale random variable Z from an exponential distribution. That is, given the exponential distrubution for Z

$$\text{Exponential dist.:} \qquad f_Z(z) = \frac{1}{\lambda}\exp(-\frac{z}{\lambda}) \qquad : z > 0,\ \lambda > 0, \tag{3.8}$$

and that the conditional pdf of $\mathbf{Y}|Z$ is multivariate Gaussian (3.3), then the marginal pdf of Y is found by the integral (3.4) to be

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \int f_{\mathbf{Y}|Z}(\mathbf{y}|z)\ f_Z(z)\ dz \\ &= \int_0^\infty \frac{1}{(2\pi z)^{\frac{d}{2}}} \exp\left[-\frac{1}{2z}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \frac{1}{\lambda}\exp(-\frac{z}{\lambda})\ dz. \end{aligned} \tag{3.9}$$

Substituting $\delta^2 = q(\mathbf{y})$ and $\gamma^2 = \frac{2}{\lambda}$ gives

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{d}{2}}}\frac{\gamma^2}{2}\frac{\sqrt{2\pi}e^{-\delta\gamma}}{\delta}\int_0^\infty z^{-(\frac{d}{2}-\frac{3}{2})}\frac{\delta}{\sqrt{2\pi}}e^{\delta\gamma}z^{-\frac{3}{2}}\exp[-\frac{1}{2}(\frac{\delta^2}{z}+\gamma^2 z)]\ dz, \tag{3.10}$$

which is recognised as the $-(\frac{d}{2}-\frac{3}{2})$-order moment of an inverse Gaussian distribution (3.15). From [5], the k-order moment of an IG distribution is given as

$$\mathcal{E}\{z^k\}_{IG} = \left(\frac{\delta}{\gamma}\right)^k \frac{K_{-\frac{1}{2}+k}(\delta\gamma)}{K_{-\frac{1}{2}}(\delta\gamma)}, \tag{3.11}$$

where $K_m(x)$ is a modified Bessel function of the second kind with order m.

Using (3.11), basic properties of Bessel K functions and substituting back into (3.10) yields the closed form expression for the marginal pdf of the multivariate Laplacian distribution

$$\text{ML:} \qquad f_Y(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \, \frac{2}{\lambda} \, \frac{K_{\frac{d}{2}-1}\big(\sqrt{\frac{2}{\lambda}q(\mathbf{y})}\big)}{\left(\sqrt{\frac{\lambda}{2}q(\mathbf{y})}\right)^{\frac{d}{2}-1}}. \tag{3.12}$$

The multivariate Laplacian has parameters λ, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ and can be denoted as $ML(\lambda, \boldsymbol{\mu}, \boldsymbol{\Gamma})$. Note particularly that for dimension 2 or higher, the Laplacian actually goes to infinity at the peak centre.

3.1.3 Multivariate K-distribution

The multivariate K distribution mixture model is obtained with the scale variable Z from a Gamma distribution

$$\text{Gamma dist.:} \qquad f_Z(z) = \frac{\lambda^{\alpha+1} z^{\alpha}}{\Gamma(\alpha+1)} \exp(-\lambda z) \quad : \; z > 0, \; \lambda > 0, \; \alpha > -1, \tag{3.13}$$

and the general pdf of the K-distribution, by integrating (3.4) similarly to the multivariate Laplacian case, is found to be

$$\text{MK:} \qquad f_Y(\mathbf{y}) = \frac{2}{(2\pi)^{\frac{d}{2}}} \, \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} \left(\sqrt{\frac{q(\mathbf{y})}{2\lambda}}\right)^{\alpha+1-\frac{d}{2}} K_{\alpha+1-\frac{d}{2}}\big(\sqrt{2\lambda q(\mathbf{y})}\big), \tag{3.14}$$

with $q(\mathbf{y})$ and $K_m(x)$ defined above.

The K-distribution has two parameters, α and λ, in addition to the mean, $\boldsymbol{\mu}$, and covariance, $\boldsymbol{\Gamma}$, and can be denoted $MK(\alpha, \lambda, \boldsymbol{\mu}, \boldsymbol{\Gamma})$. The (scalar) α is known as the shape parameter and (scalar) λ as the width parameter, from the way that they affect the distribution. When $\alpha = 0$, the multivariate K-distribution simplifies to the multivariate Laplacian (3.12). The MK distribution may also reach infinity at the peak centre whenever $\alpha \leq \frac{d}{2} - 1$.

Note that there are several different versions of the K-distribution referred to in the literature, for example, the Homodyne K distribution [13] and the Generalised K distribution [14]. They are generally not multivariate and refer to the distribution of the envelope or intensity. They are all generated from a Gamma distributed scale paramter, and this is one reason that the multivariate K-distribution, used here and described in [3], is also named with a K.

3.1.4 Multivariate Normal Inverse Gaussian

The multivariate normal inverse Gaussian (MNIG) is a scale mixture with Z from an inverse Gaussian distribution

$$\text{IG dist.:} \qquad f_Z(z) = \frac{\delta}{\sqrt{2\pi}} e^{\delta\gamma}\, z^{-\frac{3}{2}}\ \exp\big(-\frac{1}{2}(\frac{\delta^2}{z} + \gamma^2 z)\big) \quad : \ z > 0,\ \delta > 0,\ \gamma > 0, \tag{3.15}$$

and the marginal MNIG pdf is given by

$$\text{MNIG:} \qquad f_Y(\mathbf{y}) = 2\delta e^{\delta\gamma} \left(\frac{\gamma}{2\pi\sqrt{\delta^2 + q(\mathbf{y})}} \right)^{\frac{d+1}{2}} K_{\frac{d+1}{2}}\big(\gamma\sqrt{\delta^2 + q(\mathbf{y})}\big), \tag{3.16}$$

again with $q(\mathbf{y})$ and $K_m(x)$ as defined previously.

The MNIG has two new parameters, δ and γ, plus the mean and covariance, and is denoted $MNIG(\delta, \gamma, \boldsymbol{\mu}, \mathbf{\Gamma})$. The scalar δ is considered a shape parameter and γ a width parameter. Note that in contrast to the ML and MK models, the MNIG does not go to infinity at the peak centre.

3.1.5 Properties

All models are symmetric about the mean and although each dimension may have different relative widths, distributed by the covariance matrix $\mathbf{\Gamma}$, they will each have a similar shape governed by the global scalar parameters. The MG and ML distributions have fixed shapes, the rounded Gaussian and the pointed Laplacian, and only vary in width. The MK's and MNIG's two scalar parameters lead to a range of shapes as well as overall widths. The shapes range from more pointed than Laplacian, through to the rounded Gaussian (see Figure 3.1) and are usually determined directly from one scalar parameter. The shape parameter does not vary linearly in value with most of the variation occuring near the lowest value and converging asymptotically towards the Gaussian curve for large values. The width parameter is slightly misnamed as the distribution width is usually an expression combining both the width and shape parameters, however given that the shape is solely determined by the first parameter, then the other, although indirectly, is referred to as the width parameter. Also note that both the ML and MK distribution's pdfs can go to infinity at the mean value, whereas the MG and MNIG always have a finite peak.

It is meaningful to interpret the scalar parameters as global, or average, shape and scale parameters, since they act upon all dimensions equally. The main polarimetric information is contained within the covariance structure matrix. That is, the specific relative scales of one dimension to the other, which relate back to the target scattering matrix components for the given HV polarisation basis.

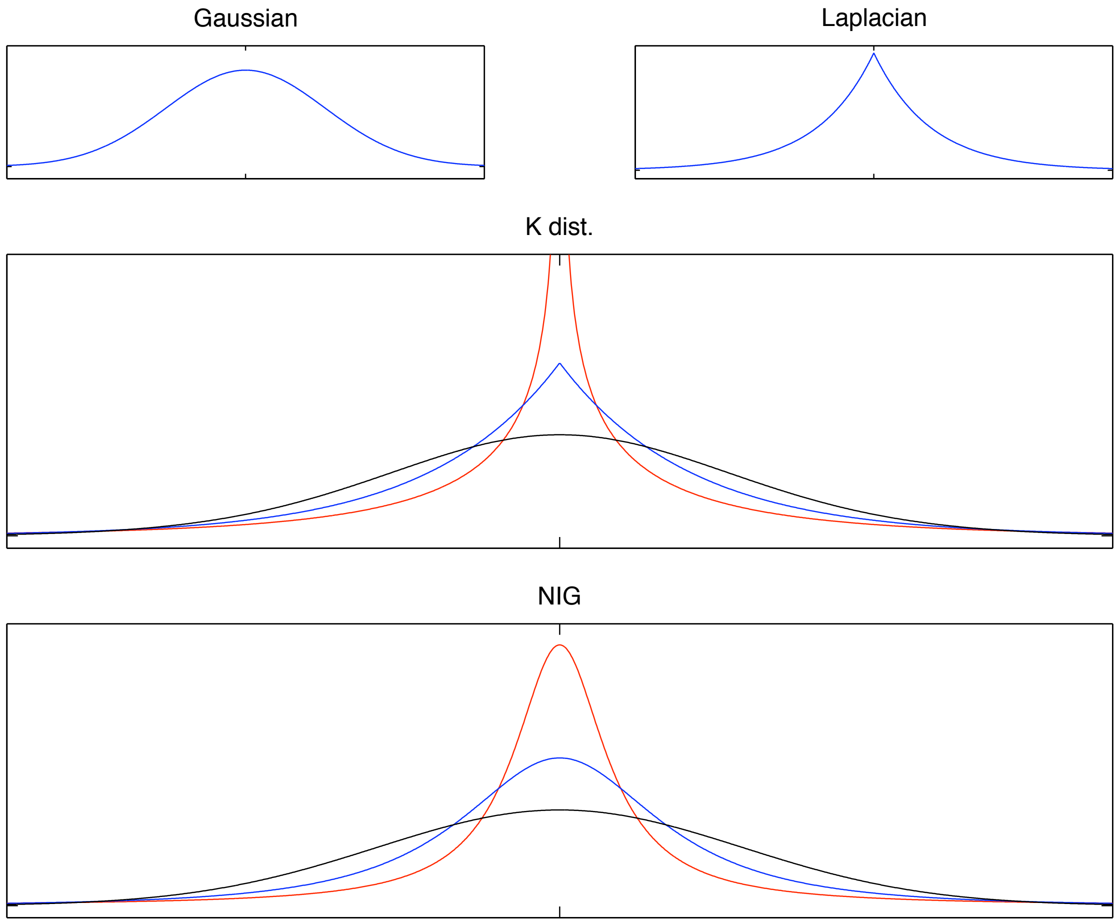


Figure 3.1: Example shapes of each model with fixed width. Note that the MG and ML have a fixed shape and only a width parameter, whereas the MK and MNIG, with two parameters each, have a range of shapes.

Chapter 4

Parameter Estimation

Parameter estimation is used to find the particular parameters, for a given parametric model, that best describes the data set. This is essentially *best fitting* the model to the data and finds the parameter values for the particular curve of the model's family of curves that is closest to the data. Obviously, each of the four models can be best fitted in this way to find the estimated parameter set for each one.

This study started by implimenting the iterative decomposition method described in [3] and [15] for each of the models (see Section 4.1). This is based upon the Expectation Maximisation (EM) algorithm [16] and was stable but very slow. Two minor iterative improvements were investigated. Firstly, an alternative to the linear regressions (Section 4.2), and secondly a version needing only half of the time consuming Bessel function calculations (Section 4.3).

The slow processing time of the iterative method led to investigations of empirical moment based methods, the basis of which is described as Method I in [11]. Two variations are studied here (Section 4.4), both using second and fourth order moments in the (assumed scale mixture of Gaussian) sample distribution, $\mathbf{Y}$, to estimate first and second order moments in the scale distribution, Z, and thereby the Z distribution's parameters. The first version uses a simple one dimensional kurtosis expression (Section 4.4.1) and the second uses the multidimensional Mardia's kurtosis [17].

4.1 Iterative estimation and the EM algorithm

Maximum likelihood estimation (MLE) of model parameters is often used when the form of the distribution is known. MLE finds the optimum parameters that maximise the likelihood or log-likelihood function (see Section 6.4). The MLE estimate is the most likely set of parameter values given the observed data. However, when a closed form solution is unobtainable (from the derivative of the likelihood function), then an iterative numerical method must be used to find the maximum.

The multivariate Gaussian distribution can be solved analytically from the log-likelihood function (one reason Gaussians are so often used in modelling) and results in estimates for the parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, as

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i \tag{4.1}$$

$$\hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^T. \tag{4.2}$$

The other distributions formed by the mixture model involve the latent variable Z, as well as modified Bessel functions, and are not analytically solvable. The iterative EM algorithm [16] can be used, which maximises the expectation of the log-likelihood function of the conditional probability density function $Z|\mathbf{Y}$, conditioned on the observed samples, the current iteration estimates of z_i and any other parameters. The algorithm has two steps, the expectation or E-step, which computes the expected value of the conditional log-likelihood function, given the current estimates of the parameters, and the maximisation or M-step, which computes the next estimates of the parameters by a maximisation, often via partial derivatives, of the estimated log-likelihood function from the E step. The process is repeated until the parameters converge to their optimum values.

Implementing the iterative algorithm required testing for convergence of each of the parameters rather than just one, because they may converge at different rates. Considering that the goodness-of-fit testing would subsequently be testing based upon the log-likelihood score (see Chapter 6), it seemed appropriate to use the log-likelihood score as the convergence test also, thus combining the convergence of several parameters into one measure.

4.1.1 Estimating ML λ, MK α, λ and MNIG δ, γ

Each of the mixture models have a common *a posteriori* probability density function, $f_{Z|\mathbf{Y}}(z|y)$, that is a Generalised Inverse Gaussian (GIG) distribution [5]

$$\text{GIG:} \qquad f_Z(z;\theta,\delta,\gamma) = (\frac{\gamma}{\delta})^\theta \, \frac{1}{2K_\theta(\delta\gamma)} z^{\theta-1} \, \exp\Big(-\frac{1}{2}(\frac{\delta^2}{z} + \gamma^2 z)\Big), \tag{4.3}$$

and its k^{th}-order moments are

$$\boldsymbol{\mu}_Z^k = E\{Z^k\} = \big(\frac{\delta}{\gamma}\big)^k \, \frac{K_{\theta+k}(\delta\gamma)}{K_\theta(\delta\gamma)}. \tag{4.4}$$

However, each model has different substitutions for the GIG parameters. Note also that the GIG δ and γ are not the same as those described in the multivariate NIG distribution.

The model parameters can each be estimated from one or two of the moments $E\{Z^k\}$ using an EM style algorithm. The E-step involves updating the moments of $Z|Y$, and the M-step updates the parameters using a combination of moment estimators and maximum likelihood estimators.

Laplacian λ

The *a posteriori* conditional pdf, $f_{Z|\mathbf{Y}}(z|y)$, associated with the ML model (3.12) can be shown to be $GIG\{-\frac{d}{2}+1, \sqrt{q(\mathbf{y})}, \sqrt{\frac{2}{\lambda}}\}$. Hence, for a given observation y_i, estimating the first moment of $Z|Y$, from (4.4) with $k=1$, gives

$$\eta_i = \mathcal{E}\{Z|\mathbf{y}_i\} = \sqrt{\frac{\lambda q(\mathbf{y}_i)}{2}}\,\frac{K_{-\frac{d}{2}+2}\big(\sqrt{\frac{2}{\lambda}q(\mathbf{y}_i)}\big)}{K_{-\frac{d}{2}+1}\big(\sqrt{\frac{2}{\lambda}q(\mathbf{y}_i)}\big)}. \tag{4.5}$$

The first inverse moment, for $k=-1$, is given by

$$\xi_i = \mathcal{E}\{\frac{1}{Z}|\mathbf{y}_i\} = \sqrt{\frac{2}{\lambda q(\mathbf{y}_i)}}\,\frac{K_{-\frac{d}{2}}\big(\sqrt{\frac{2}{\lambda}q(\mathbf{y}_i)}\big)}{K_{-\frac{d}{2}+1}\big(\sqrt{\frac{2}{\lambda}q(\mathbf{y}_i)}\big)}. \tag{4.6}$$

When Z is exponential, given by (3.8), then it is easily shown that $\mathcal{E}\{Z\}=\lambda$ and hence λ may be estimated as the average of $\mathcal{E}\{Z|Y\}$ over the data set,

$$\hat{\lambda} = \mathcal{E}\{Z\} \approx \bar{\eta} = \frac{1}{N}\sum_{i=1}^{N}\eta_i. \tag{4.7}$$

K-distribution α and λ

The *a posteriori* conditional pdf of $Z|\mathbf{Y}$ associated with the MK model (3.14) is a $GIG\{-\frac{d}{2}+\alpha+1, \sqrt{q(\mathbf{y})}, \sqrt{2\lambda}\}$. To solve for two parameters requires two moment estimates. Estimating the moments with $k=1$ and $k=-1$ of $Z|Y$ gives

$$\eta_i = \mathcal{E}\{Z|\mathbf{y}_i\} = \sqrt{\frac{q(\mathbf{y}_i)}{2\lambda}}\,\frac{K_{\alpha-\frac{d}{2}+2}\big(\sqrt{2\lambda q(\mathbf{y}_i)}\big)}{K_{\alpha-\frac{d}{2}+1}\big(\sqrt{2\lambda q(\mathbf{y}_i)}\big)} \tag{4.8}$$

and

$$\xi_i = \mathcal{E}\{\frac{1}{Z}|\mathbf{y}_i\} = \sqrt{\frac{2\lambda}{q(\mathbf{y}_i)}}\,\frac{K_{\alpha-\frac{d}{2}}\big(\sqrt{2\lambda q(\mathbf{y}_i)}\big)}{K_{\alpha-\frac{d}{2}+1}\big(\sqrt{2\lambda q(\mathbf{y}_i)}\big)}. \tag{4.9}$$

When Z is Gamma distributed (3.13), then

$$\mathcal{E}\{Z\} = \frac{(\alpha+1)}{\lambda} \tag{4.10}$$

$$\mathcal{E}\{Z^{-1}\} = \frac{\lambda}{\alpha}, \tag{4.11}$$

and therefore α and λ may be estimated as

$$\hat{\alpha} = \frac{1}{\bar{\eta}\bar{\xi} - 1} \tag{4.12}$$

$$\hat{\lambda} = \hat{\alpha}\bar{\xi}, \tag{4.13}$$

where $\bar{\eta}$ and $\bar{\xi}$ are the simple averages as in (4.7).

NIG δ and γ

The MNIG model (3.16) leads to an *a posteriori* conditional pdf of $Z|\mathbf{Y}$ as a $GIG\{-\frac{d+1}{2}, \sqrt{\delta^2 + q(\mathbf{y})}, \gamma\}$. Estimating moments for $k = 1$ and $k = -1$ of $Z|Y$ gives

$$\eta_i = \mathcal{E}\{Z|\mathbf{y}_i\} = \frac{\sqrt{\delta^2 + q(\mathbf{y}_i)}}{\gamma} \frac{K_{-\frac{d-1}{2}}\left(\gamma\sqrt{\delta^2 + q(\mathbf{y}_i)}\right)}{K_{-\frac{d+1}{2}}\left(\gamma\sqrt{\delta^2 + q(\mathbf{y}_i)}\right)} \tag{4.14}$$

and

$$\xi_i = \mathcal{E}\{\frac{1}{Z}|\mathbf{y}_i\} = \frac{\gamma}{\sqrt{\delta^2 + q(\mathbf{y}_i)}} \frac{K_{-\frac{d+3}{2}}\left(\gamma\sqrt{\delta^2 + q(\mathbf{y}_i)}\right)}{K_{-\frac{d+1}{2}}\left(\gamma\sqrt{\delta^2 + q(\mathbf{y}_i)}\right)}. \tag{4.15}$$

When Z is inverse Gaussian distributed (3.15),

$$\mathcal{E}\{Z\} = \frac{\delta}{\gamma} \tag{4.16}$$

$$\mathcal{E}\{Z^{-1}\} = \frac{(1+\delta\gamma)}{\delta^2}, \tag{4.17}$$

and therefore δ and γ may be estimated as

$$\hat{\delta} = \frac{1}{\sqrt{\bar{\xi} - \frac{1}{\bar{\eta}}}} \tag{4.18}$$

$$\hat{\gamma} = \frac{\hat{\delta}}{\bar{\eta}}. \tag{4.19}$$

4.1.2 Estimation of mean $\boldsymbol{\mu}$ and covariance structure matrix $\boldsymbol{\Gamma}$

All the mixture models have a common EM method for finding $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, since they all originate from the same mixture process. Their conditional distributions, $Y|Z$, are always a multivariate Gaussian for which the log-likelihood function can be maximised. The EM algorithm therefore estimates the weights $\frac{1}{z_i}$, from the conditional pdf of $Z|\mathbf{Y}$, given current estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ and then maximises the log-likelihood function of $Y|Z$ to obtain new estimates, iterating both steps until convergence.

Mean $\boldsymbol{\mu}$ from weighted average

Since $\mathbf{Y} = \boldsymbol{\mu} + \sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X}$, then

$$f_{\mathbf{Y}|Z}(\mathbf{y}|Z=z) = \frac{1}{(2\pi\ z)^{\frac{d}{2}}}\ \exp\left[-\frac{1}{2z}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right], \tag{4.20}$$

and the log-likelihood function

$$\begin{aligned} LL(\boldsymbol{\mu},\boldsymbol{\Gamma}) \propto &-\frac{N}{2}\ \log(\det\boldsymbol{\Gamma}) \\ &-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{z_i}(\mathbf{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}). \end{aligned} \tag{4.21}$$

The second term of (4.21) can be expanded as

$$-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{z_i}\left[\mathbf{y}_i^T\boldsymbol{\Gamma}^{-1}\mathbf{y}_i - \mathbf{y}_i^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Gamma}^{-1}\mathbf{y}_i + \boldsymbol{\mu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}\right]. \tag{4.22}$$

Maximising with respect to $\boldsymbol{\mu}$, is achieved by finding the solution for the partial derivative equal to zero. Thus

$$\begin{aligned} \frac{\partial LL(\boldsymbol{\mu},\boldsymbol{\Gamma})}{\partial\boldsymbol{\mu}} &= 0 - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{z_i}\left[0 - \boldsymbol{\Gamma}^{-1}\mathbf{y}_i - \boldsymbol{\Gamma}^{-1}\mathbf{y}_i + 2\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}\right] \\ &= \sum_{i=1}^{N}\frac{1}{z_i}\left[\boldsymbol{\Gamma}^{-1}\mathbf{y}_i - \boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}\right] \end{aligned}$$

$$\frac{\partial LL}{\partial\boldsymbol{\mu}} = 0 \implies \boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}\left(\sum_{i=1}^{N}\frac{1}{z_i}\right) = \boldsymbol{\Gamma}^{-1}\sum_{i=1}^{N}\frac{1}{z_i}\mathbf{y}_i, \tag{4.23}$$

and $\boldsymbol{\Gamma}$ may be cancelled since it is positive semi-definite. The remaining equation gives the estimate for $\boldsymbol{\mu}$ as a weighted sample mean

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{N}\frac{1}{z_i}\mathbf{y}_i}{\sum_{i=1}^{N}\frac{1}{z_i}}, \tag{4.24}$$

where the weights $\frac{1}{z_i}$ are substituted with the specific ξ_i functions for each model, being updated each time through the iteration.

Covariance structure matrix $\mathbf{\Gamma}$ from linear regresssions

The method used by [3] and [18] decomposes the inverse $\mathbf{\Gamma}$ matrix into upper triangular and diagonal matrices, maximising for each part. It results in solving (dimension-1) linear regressions from a weighted covariance style matrix, and some $d \times d$ matrix manipulation to obtain the maximised estimate for $\mathbf{\Gamma}$.

The covariance structure matrix, $\mathbf{\Gamma}$, is a positive definite matrix, and therefore so is $\mathbf{\Gamma}^{-1}$. Hence $\mathbf{\Gamma}^{-1}$ decomposes as $\mathbf{U}^T\mathbf{D}\mathbf{U}$ where $\mathbf{U}$ is an unit upper triangular matrix with ones on the diagonal, and $\mathbf{D}$ is a positive diagonal matrix. Also, $\mathbf{B} = \mathbf{I} - \mathbf{U}$ is an upper triangular matrix with zeros along the diagonal.

Considering this decomposition, the log-likelihood function for $Y|Z$ becomes

$$\begin{aligned}
&LL(\boldsymbol{\mu}, \mathbf{D}, \mathbf{U}) \\
&\qquad \propto -\frac{N}{2}\log\left(\frac{1}{\det \mathbf{U}^T \det \mathbf{D} \det \mathbf{U}}\right) - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{z_i}(\mathbf{y}_i - \boldsymbol{\mu})^T\mathbf{U}^T\mathbf{D}\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu}) \\
&\qquad =\frac{N}{2}\ \log(\det \mathbf{D}) - \frac{1}{2}\ \sum_{i=1}^{N}\frac{1}{z_i}\left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right]^T \mathbf{D}\ \ \left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right]. \qquad (4.25)
\end{aligned}$$

Optimising the log-likelihood function via partial derivatives with respect to $\mathbf{D}$ gives

$$\begin{aligned}
\frac{\partial LL(\boldsymbol{\mu}, \mathbf{D}, \mathbf{U})}{\partial \mathbf{D}} &= \frac{N}{2}\ \mathbf{D}^{-T} - \frac{1}{2}\ \sum_{i=1}^{N}\frac{1}{z_i}\left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right]\left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right]^T \\
\frac{\partial LL}{\partial \mathbf{D}} = 0 &\implies \hat{\mathbf{D}}^{-1} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})][\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})]^T. \qquad (4.26)
\end{aligned}$$

$\mathbf{D}$ is a diagonal matrix by definition, so only the diagonal terms are required. Furthermore, since det $(\mathbf{\Gamma}) = 1$ and det $(\mathbf{U}) = 1$, it is required that det $\mathbf{D}$ be normalised to 1. The diagonal terms are

$$\begin{aligned}
\hat{\mathbf{D}}_{kk}^{-1} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}\left[(\mathbf{U}(\mathbf{y}_i-\boldsymbol{\mu}))_k\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}\left[((\mathbf{I}-\mathbf{B})(\mathbf{y}_i-\boldsymbol{\mu}))_k\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}\left[((\mathbf{y}_i-\boldsymbol{\mu})-\mathbf{B}(\mathbf{y}_i-\boldsymbol{\mu}))_k\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}\left[((\mathbf{y}_{ik}-\boldsymbol{\mu}_k)-\sum_{j=k+1}^{d}(\mathbf{y}_{ij}-\boldsymbol{\mu}_j)B_{kj})\right]^2, \qquad (4.27)
\end{aligned}$$

since $\mathbf{B}$ is upper triangular with zeros along the diagonals. This estimate is then normalised by dividing by its determinant.

Equation (4.27) is exactly the form of a least squared error formulation for B_{kj} that would minimise each $\mathbf{D}^{-1}$ term, thus maximising each $\mathbf{D}$ term. Therefore the maximisation for $\mathbf{D}$ indicates that $\mathbf{B}$ is expressed as a set of linear regressions of the form

$$\hat{\mathbf{b}}_{\mathbf{k}} = \arg\min_{B_{kj}}\left\{\sum_{i=1}^{N}\frac{1}{z_i}\left[(y_{ik}-\boldsymbol{\mu}_k)-\sum_{j=k+1}^{d}(y_{ij}-\boldsymbol{\mu}_j)B_{kj}\right]^2\right\}, \qquad (4.28)$$

where $\hat{\mathbf{b}}_{\mathbf{k}}$ denotes an estimate of the k^{th} row of $\mathbf{B}$, and y_{ij} refers to component j of observation vector $\mathbf{y}_i$.

Solving linear regressions using matrix methods involves calculating all combinations of sums of the form $\frac{1}{z_i}(\mathbf{y}_{il}-\mu_l)(\mathbf{y}_{im}-\mu_m)$ to build a $d \times d$ combination matrix. Then for each dimension, k, up to $d-1$, take the lower right minor square matrices (terms from $k+1$ to d), invert the minor matrix and multiply by the $k+1$ to d terms of the k^{th} row.

The linear regressions also come about by looking at the partial derivative with respect to **U** (equivalent to **B**).

$$\begin{aligned}
\frac{\partial LL(\boldsymbol{\mu}, \mathbf{D}, \mathbf{U})}{\partial \mathbf{U}} &= \frac{\partial}{\partial \mathbf{U}} \left\{ -\frac{1}{2} \sum_{i=1}^{N} \frac{1}{z_i} \left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right]^T \mathbf{D} \left[\mathbf{U}(\mathbf{y}_i - \boldsymbol{\mu})\right] \right\} \\
&= -\frac{1}{2} \sum_{i=1}^{N} \frac{1}{z_i} \left[2\mathbf{DU}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right] \\
&= -\mathbf{DU} \sum_{i=1}^{N} \frac{1}{z_i} \left[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right] \\
&= -\mathbf{D}(\mathbf{I} - \mathbf{B}) \sum_{i=1}^{N} \frac{1}{z_i} \left[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right] \\
\frac{\partial LL}{\partial \mathbf{U}} = 0 &\implies \mathbf{B} \sum_{i=1}^{N} \frac{1}{z_i} \left[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right] = \sum_{i=1}^{N} \frac{1}{z_i} \left[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right].
\end{aligned} \tag{4.29}$$

And since **B** is singular, being upper diagonal with zero diagonals, expanding individual terms leads to a set of equations relating the elements of each row $\mathbf{b}_k$.

Defining the weighted sample covariance matrix **C** as,

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{z_i} (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T, \tag{4.30}$$

then (4.29) now states that $\mathbf{BC} = \mathbf{C}$ (the factor of $\frac{1}{N}$ being included on both sides of equation (4.29)). Remembering that the partial derivative $\frac{\partial}{\partial \mathbf{B}}$ is a matrix notation for all of the separate partials $\frac{\partial}{\partial B_{k,l}}$, then the expansion is only valid for the terms in the matrix where $B_{k,l}$ exists, i.e., $k+1$ to d in each row k. Expanding each row gives

$$\text{for row } k \qquad \sum_{j=k+1}^{d} B_{k,j} C_{k,l} = C_{k,l} \qquad \forall\ l = (k+1), \ldots, d.$$

Therefore each row gives an exactly determined set of linear regressions whose terms come from the matrix **C** and can be solved using matrix methods. As an example, in the 4 dimensional case for row 1 this would look like

$$\begin{pmatrix} B_{1,2} & B_{1,3} & B_{1,4} \end{pmatrix} \begin{pmatrix} C_{2,2} & C_{2,3} & C_{2,4} \\ C_{3,2} & C_{3,3} & C_{3,4} \\ C_{4,2} & C_{4,3} & C_{4,4} \end{pmatrix} = \begin{pmatrix} C_{2,1} & C_{3,1} & C_{4,1} \end{pmatrix}.$$

Lets call the regression minor matrix $\mathbf{A}$, and the solution vector $\mathbf{v}$, therefore $\mathbf{b}_k \cdot \mathbf{A} = \mathbf{v}$. Since $\mathbf{A}$ comes from a sample covariance matrix, it is likely to be full rank and invertible. Therefore the solution is $\mathbf{b}_k = \mathbf{v} \cdot \mathbf{A}^{-1}$.

The $\frac{1}{z_i}$ weighted sample covariance martix $\mathbf{C}$ is exactly the same combination matrix discussed after (4.28). Therefore the solution for each row $\mathbf{b}_k$ is the same least squares estimate and the same set of linear regressions as before.

In summary, solving the regressions gives $\hat{\mathbf{B}}$, which leads to $\hat{\mathbf{U}} = \mathbf{I} - \hat{\mathbf{B}}$, and $\hat{\mathbf{D}}^{-1}$ by (4.26). Since $\hat{\mathbf{\Gamma}}^{-1} = \hat{\mathbf{U}}^T\hat{\mathbf{D}}\hat{\mathbf{U}}$ and $\mathbf{U}$ being upper diagonal with identity diagonals, it follows that $\hat{\mathbf{\Gamma}} = \hat{\mathbf{U}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{U}}^T$. This is repeated with the updated $\hat{\boldsymbol{\mu}}$ and weights $\frac{1}{z_i} \approx \xi_i$ each time through the iteration until convergence.

4.2 Iterative estimation: alternative $\mathbf{\Gamma}$

As an alternative to the linear regressions, $\mathbf{\Gamma}$ may be estimated directly from the weighted sample covariance estimate, centred on the current iterative estimate for $\hat{\boldsymbol{\mu}}$, plus the constraint that $\det(\mathbf{\Gamma}) = 1$. This is conceptually simpler, makes the process up to 15% faster, and should converge to the same maximum as the linear regression method.

Covariance structure matrix $\mathbf{\Gamma}$ from weighted covariance matrix

Starting from the expression for the log-likelihood function of $\boldsymbol{\mu}$ and $\mathbf{\Gamma}$ (4.21). The partial derivative with respect to $\mathbf{\Gamma}$ is

$$\begin{aligned}
\frac{\partial LL(\boldsymbol{\mu},\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= -\frac{N}{2}\mathbf{\Gamma}^{-T} - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{z_i}\left[-\mathbf{\Gamma}^{-T}(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^T\mathbf{\Gamma}^{-T}\right] \\
&= -\frac{N}{2}\mathbf{\Gamma}^{-T} + \frac{1}{2}\,\mathbf{\Gamma}^{-T}\left[\sum_{i=1}^{N}\frac{1}{z_i}\left[(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^T\right]\right]\mathbf{\Gamma}^{-T} \\
\frac{\partial LL}{\partial \mathbf{\Gamma}} = 0 \implies \frac{N}{2}\mathbf{\Gamma}^{-T} &= \frac{1}{2}\,\mathbf{\Gamma}^{-T}\left[\sum_{i=1}^{N}\frac{1}{z_i}\left[(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^T\right]\right]\mathbf{\Gamma}^{-T}. \qquad (4.31)
\end{aligned}$$

$\mathbf{\Gamma}$ is nonsingular and symmetric, so pre and post multiplying (4.31) by $\mathbf{\Gamma}$ leads directly to the optimised estimate for $\mathbf{\Gamma}$ as a weighted average,

$$\hat{\mathbf{\Gamma}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z_i}(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^T. \qquad (4.32)$$

This is indeed expected from the scale mixture of Gaussians generation method, since

$$\begin{aligned}
& \mathbf{y}_i = \boldsymbol{\mu} + \sqrt{z_i}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x}_i \qquad \text{for some Gaussian distributed } \mathbf{x}_i \\
\Rightarrow\quad & \frac{1}{\sqrt{z_i}}(\mathbf{y}_i - \boldsymbol{\mu}) = \boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x}_i \\
\Rightarrow\quad & E\left\{\left[\frac{1}{\sqrt{z_i}}(\mathbf{y}_i - \boldsymbol{\mu})\right]\left[\frac{1}{\sqrt{z_i}}(\mathbf{y}_i - \boldsymbol{\mu})\right]^T\right\} = E\left\{\left[\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x}_i\right]\left[\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x}_i\right]^T\right\} \\
\Rightarrow\quad & \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{z_i}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right] \approx \boldsymbol{\Gamma}^{\frac{1}{2}}E\left\{\mathbf{x}_i\mathbf{x}_i^T\right\}\boldsymbol{\Gamma}^{\frac{1}{2}}. \qquad (4.33)
\end{aligned}$$

By definition, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$, and so $E\{\mathbf{x}_i\mathbf{x}_i^T\} = \mathcal{I}$ and the right hand side of (4.33) becomes simply $\boldsymbol{\Gamma}$. Clearly, the de-scaled sample covariance is expected to be the covariance structure matrix of the scale mixture generation method.

Therefore, given estimates for $\frac{1}{z_i}$ and $\boldsymbol{\mu}$, the optimum estimate for the covariance structure matrix $\boldsymbol{\Gamma}$ is the weighted sample covariance matrix. Again, since the function ξ_i requires values for both $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, the optimisation is evaluated by iterations until convergence.

Equivalence of both methods

Both the linear regressions via decomposition method and the direct weighted sample covariance method should be equivalent since both are derived theoretically from optimising the same log-likelihood function. In the decomposition method, the regression solutions are simple combinations of terms of the weighted covariance matrix, and the subsequent matrix manipulation must recombine them in such a way that the mixed terms cancel away again. This can be shown explicitly in the simple 2-D case, but the regression solutions and determinants become too complex to show in the general case. Testing both methods on generated sample data showed that they obtain the same numerical solution at every stage of the iterative process as can be seen in the following results (Table 4.1) and convergence plot (Figure 4.1). The red line and black dashed line overlap at every stage of the iteration, showing that both methods produced the same numerical result. Note also that the $\frac{1}{z_i}$ weighted sum covariance matrix automatically converged to a normalised $\boldsymbol{\Gamma}$ matrix with determinant 1.

4.3 Iterative estimation: alternative $\bar{Z}$

The EM algorithm estimates the conditional pdf of $Z|\mathbf{Y}$ and then finds optimum parameters to maximise this function. Considering that the scale parameter Z of

Table 4.1: Numerical equivalence of linear regressions and weighted covariance matrix.

```
Generated Laplacian data with N = 5000, lambda = 0.05, Mu = zero and
Gamma =     2      0      1
            0      1      0
            1      0      1

Fitted data using linear regressions:
M.Laplacian log-likelihood score = 2828.5895.
#     lambda       mu              Gamma
14    0.05026      0.00160         2.00480      0.00363      0.97363
                   0.00097         0.00363      1.00933      0.00676
                   0.00286         0.97363      0.00676      0.96707

Fitted data using direct weighted sample covariance:
M.Laplacian log-likelihood score = 2828.5895.
#     lambda       mu              Gamma
14    0.05026      0.00160         2.00480      0.00363      0.97363
                   0.00097         0.00363      1.00933      0.00676
                   0.00286         0.97363      0.00676      0.96707
```

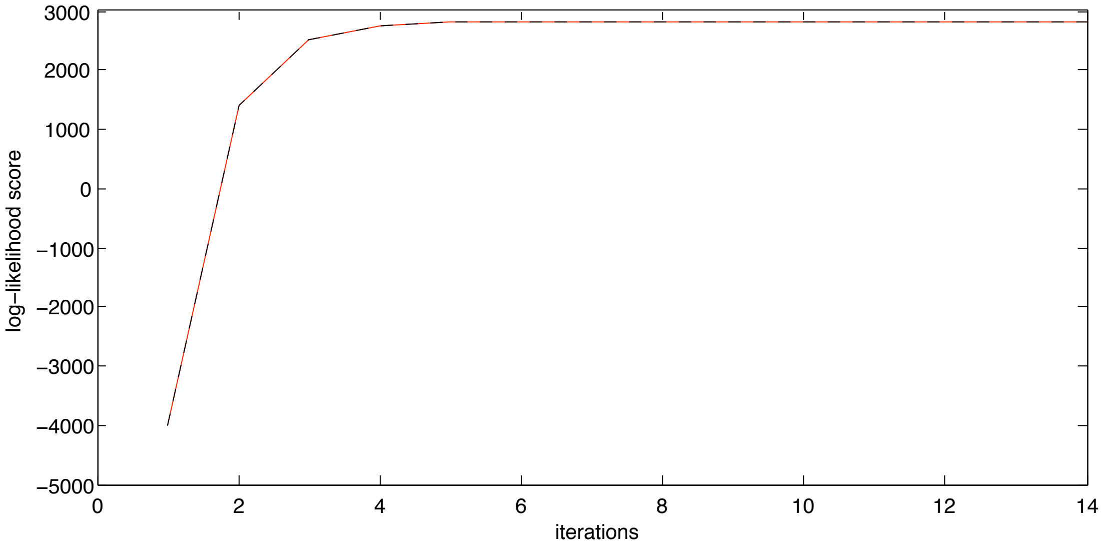


Figure 4.1: Convergence plot for linear regressions method in red and direct weighted average covariance method in dashed black. They are equivalent at every stage of the iteration.

the scale mixture of Gaussians method acts essentially the same for each dimension, after accounting for the covariance structure, it can be shown that the estimate of z_i given $\mathbf{y}_i$ can be estimated from the dimensional data for each sample. That is, given that $\mathbf{Y} = \boldsymbol{\mu} + \sqrt{Z}\ \boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X}$ then

$$\begin{aligned}(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) &= (\sqrt{z_i}\ \boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x}_i)^T\boldsymbol{\Gamma}^{-1}(\sqrt{z_i}\ \boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x_i}) \\ &= z_i\ \mathbf{x_i}^T\boldsymbol{\Gamma}^{\frac{T}{2}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{x_i} \\ &= z_i\ \mathbf{x_i}^T\mathbf{x_i} \\ &= z_i\ \sum_{k=1}^{d} x_{i,k}^2.\end{aligned}$$

The left hand side is simply $q(\mathbf{y_i})$, and since $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ then $\mathcal{E}\{x_{i,k}^2\} = 1$ and $\sum_{k=1}^{d} x_{i,k}^2 \approx d$. The best estimate for z_i, given only $\mathbf{y}_i$, can be considered to be $q(\mathbf{y}_i)/d$. Hence

$$\mathcal{E}\{Z\} = \bar{\eta} = \frac{\sum_{i=1}^{N} q(\mathbf{y}_i)}{Nd}. \tag{4.34}$$

This produces a good average estimate, $\bar{\eta}$, however it has great variance on individual estimates for z_i, due to estimating over only a few dimensions. Since the weighted sums require individual estimates for $\frac{1}{z_i}$, using this method for both η_i and ξ_i gave poor results. By using it purely for the η_i terms, which are only required in averaged form, at least avoided half of the Bessel function calculations, thus gaining significant ($\approx 40\%$) speed improvements.

4.4 Moment estimation methods

The fundamental assumptions of the multivariate scale mixture of Gaussians models are

$$\begin{aligned}&\mathbf{Y} = \ \boldsymbol{\mu} + \sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X} \\ &\det(\boldsymbol{\Gamma}) = \ 1 \\ &\mathbf{X} \sim \ \mathcal{N}(\mathbf{0}, \mathcal{I}) \\ &\mathbf{X}, Z \text{ are independent random variables} \\ &\boldsymbol{\mu} \text{ and } \boldsymbol{\Gamma} \text{ are constant parameters.}\end{aligned} \tag{4.35}$$

The sample mean leads directly to the estimate for $\boldsymbol{\mu}$ since $\mathcal{E}\{\mathbf{X}\} = \mathbf{0}$ and $\mathbf{X}$ and Z are independent. That is

$$\begin{aligned}\mathcal{E}\{\mathbf{Y}\} &= \mathcal{E}\{\boldsymbol{\mu} + \sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X}\} \\ &= \boldsymbol{\mu} + \mathcal{E}\{\sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X}\} \\ &= \boldsymbol{\mu} + \mathcal{E}\{\sqrt{Z}\}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathcal{E}\{\mathbf{X}\} \\ &= \boldsymbol{\mu}, \end{aligned} \tag{4.36}$$

and therefore

$$\hat{\boldsymbol{\mu}} = \mathcal{E}\{\mathbf{Y}\} = \text{mean}(\mathbf{Y}) \tag{4.37}$$

The sample covariance becomes

$$\begin{aligned}\mathcal{E}\{(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T\} &= \mathcal{E}\{(\sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X})(\sqrt{Z}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X})^T\} \\ &= \mathcal{E}\{Z\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{X}\mathbf{X}^T\boldsymbol{\Gamma}^{\frac{T}{2}}\} \\ &= \mathcal{E}\{Z\}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathcal{E}\{\mathbf{X}\mathbf{X}^T\}\boldsymbol{\Gamma}^{\frac{1}{2}}, \end{aligned} \tag{4.38}$$

and $\mathcal{E}\{\mathbf{X}\mathbf{X}^T\} = \mathcal{I}$. Hence,

$$\text{cov}(\mathbf{Y}) = \mathcal{E}\{(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T\} = \ \mathcal{E}\{Z\}\boldsymbol{\Gamma} \ , \tag{4.39}$$

and by definition, $\det(\boldsymbol{\Gamma}) = \ 1$, so $\boldsymbol{\Gamma}$ can be estimated from the covariance as

$$\hat{\boldsymbol{\Gamma}} = \ \frac{\text{cov}(\mathbf{Y})}{(\det(\text{cov}(\mathbf{Y})))^{\frac{1}{d}}} \ . \tag{4.40}$$

The first moment of Z is also found directly from the determinant of the sample covariance, such that

$$\bar{Z}_1 = \mathcal{E}\{Z\} = \ (\det(\text{cov}(\mathbf{Y})))^{\frac{1}{d}}. \tag{4.41}$$

In a similar fashion to the iterative method, the scalar parameters are obtained from one or two moment estimates of Z, in this case found by the separability of $\mathbf{X}$ and Z rather than from $Z|\mathbf{Y}$. One moment comes from the sample covariance, and another from a kurtosis measure, of which two forms were implemented. The first simple kurtosis expression treated each dimension as independent and used only the diagonal terms of the covariance matrix. It was found to have quite large variance and a noted bias. The improved second method derives from a proper multivariate kurtosis, accounting for the complete $\boldsymbol{\Gamma}$ matrix cross terms, and halved the variance and bias of the simple method.

The moments of Z used here, $\mathcal{E}\{Z\}$ and $\mathcal{E}\{Z^2\}$, are different from those used in the iterative method, $\mathcal{E}\{Z\}$ and $\mathcal{E}\{Z^{-1}\}$, and therefore require different solutions to find the parameters. Table 4.2 lists the new estimation expressions to solve for the parameters given the measured estimates $\bar{Z}_1$ and $\bar{Z}_2$ corresponding to theoretical expectations $\mathcal{E}\{Z\}$ and $\mathcal{E}\{Z^2\}$ respectively.

The terms $\bar{Z}_1$ and $\bar{Z}_2$ will be referred to in Part II and it is useful to understand that they are actual sample distribution measures that correspond to the first and second moments of the latent variable Z. $\bar{Z}_1$ can be considered an average sample *width* measure and $\bar{Z}_2$ a relative sample *kurtosis* measure.

Table 4.2: Moment expressions for the different models, theoretical expectations $\mathcal{E}\{Z\}$ and $\mathcal{E}\{Z^2\}$ are solved for the model parameters using measured estimates $\bar{Z}_1$ and $\bar{Z}_2$.

Model	Z distribution	$\bar{Z}_1 = \mathcal{E}\{Z\}$	$\bar{Z}_2 = \mathcal{E}\{Z^2\}$	Solution
ML	Exponential(λ)	λ	not needed	$\hat{\lambda} = \bar{Z}_1$
MK	Gamma(α, λ)	$\frac{(\alpha+1)}{\lambda}$	$\frac{(\alpha+1)(\alpha+2)}{\lambda^2}$	$\hat{\alpha} = \frac{1}{\frac{\bar{Z}_2}{(\bar{Z}_1)^2}-1} - 1$ $\hat{\lambda} = \frac{(\hat{\alpha}+1)}{\bar{Z}_1}$
MNIG	Inverse Gaussian(δ, γ)	$\frac{\delta}{\gamma}$	$(1 + \frac{1}{\delta\gamma})\frac{\delta^2}{\gamma^2}$	$\hat{\delta} = \sqrt{\frac{\bar{Z}_1}{(\frac{\bar{Z}_2}{(\bar{Z}_1)^2}-1)}}$ $\hat{\gamma} = \frac{\hat{\delta}}{\bar{Z}_1}$

4.4.1 Simple kurtosis

The first moment from the covariance expression utilised the unit variance of $\mathbf{X}$, i.e., $\mathcal{E}\{x_k^2\} = 1, \ \forall k = 1, \ldots, d$. The simplest second estimate derived from a standard normal distribution, is that of kurtosis, i.e., $\mathcal{E}\{x_k^4\} = 3, \ \forall k = 1, \ldots, d$. Therefore, taking each dimension of the whitened sample data separately , gives

$$\mathcal{E}\{[\mathbf{\Gamma}^{-\frac{1}{2}}(\mathbf{Y} - \boldsymbol{\mu})]_k^4\} = \mathcal{E}\{Z^2\} \ \mathcal{E}\{x_k^4\} = \mathcal{E}\{Z^2\} \ 3, \tag{4.42}$$

which, when averaged over all dimensions and samples, leads to a second moment estimate of

$$\bar{Z}_2 = \mathcal{E}\{Z^2\} = \frac{1}{3dN} \sum_{k=1}^{d} \sum_{i=1}^{N} [\mathbf{\Gamma}^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu})]_k^4 \ . \tag{4.43}$$

This method unfortunately showed a systematic bias and quite a large variance, perhaps being an average of each data dimension taken individually rather than a truly multivariate expression.

4.4.2 Mardia's multivariate kurtosis

A widely accepted multivariance kurtosis measure is that of Mardia's kurtosis [17]

$$\beta_{2,d} = \mathcal{E}\{[(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{\Gamma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})]^2\} \ . \tag{4.44}$$

The special case of a standard normal variable $\mathcal{N}(\mathbf{0},\mathcal{I})$ gives

$$\beta_{2,d} = d(d+2) \ , \tag{4.45}$$

which correctly gives the well known result of 3 for the 1 dimensional case. Clearly,

$$\begin{aligned}\mathcal{E}\{[(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{\Gamma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})]^2\} &= \mathcal{E}\{[(\sqrt{Z}\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{X})^T\mathbf{\Gamma}^{-1}(\sqrt{Z}\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{X})]^2\} \\ &= \mathcal{E}\{[Z\mathbf{X}^T\mathbf{X}]^2\} \\ &= \mathcal{E}\{Z^2\}\mathcal{E}\{[\mathbf{X}^T\mathbf{X}]^2\} \\ &= \mathcal{E}\{Z^2\}d(d+2) \ ,\end{aligned} \tag{4.46}$$

which can be used to estimate $\mathcal{E}\{Z^2\}$.

However, having already measured the sample covariance to find $\mathbf{\Gamma}$, it would be numerically advantageous to determine the kurtosis of a standard normal with covariance as the combination of $\mathbf{\Gamma}$ terms directly, rather than have to perform the matrix inversion and square root. Some manipulation revealed that

$$\beta_{2,d}\,\mathcal{N}(\mathbf{0},\mathbf{\Gamma}) = 3\sum_{k=1}^{d}(\mathbf{\Gamma}_{kk})^2 + 2\sum_{k=1}^{d}\sum_{j=k+1}^{d}(\mathbf{\Gamma}_{kk}\mathbf{\Gamma}_{jj}) + 4\sum_{k=1}^{d}\sum_{j=k+1}^{d}(\mathbf{\Gamma}_{kj})^2 \ , \tag{4.47}$$

which looks complicated but is fast and trivial to compute with a few programming loops. Defining $\text{kurt}_{\mathbf{\Gamma}} = \beta_{2,d}\,\mathcal{N}(\mathbf{0},\mathbf{\Gamma})$ as the sum of terms in (4.47), leads to an estimate of a second moment of Z as,

$$\bar{Z}_2 = \mathcal{E}\{Z^2\} = \frac{\mathcal{E}\{[(\mathbf{Y}-\boldsymbol{\mu})^T(\mathbf{Y}-\boldsymbol{\mu})]^2\}}{\text{kurt}_{\mathbf{\Gamma}}} = \frac{\sum_{i=1}^{N}\left[\sum_{k=1}^{d}(y_{i,k}-\mu_k)^2\right]^2}{N\ \text{kurt}_{\mathbf{\Gamma}}} \tag{4.48}$$

4.5 Discussion and accuracy tests

For the iterative methods, fine tuning of the starting conditions and convergence tests found a stable system that most often converged in less than 20 iterations. However in some situations, particularly for Gaussian-like data, it would not converge even after 250 iterations, thus taking an extremely long time to evaluate. At this point it would have taken months of processing to evaluate an entire SAR image for the four different models!

Attempts to understand and simplify the procedure led first to the alternative $\mathbf{\Gamma}$ iterative estimation that avoided doing the $\mathbf{\Gamma}$ matrix decomposition and linear regressions, and obtained exactly the same estimate directly from a weighted covariance style expression. Although simpler to understand and expressing the mean and covariance estimates in the same weighted average form, it involved only slightly fewer calculations and so did not improve performance significantly.

Profiling investigations indicated that the majority of the processing time was involved with calculating the Bessel functions. The next modification arose by trying to avoid some of the Bessel function calculations, and is derived by estimating $z_i|\mathbf{y}_i$ by averaging the scale term z_i over the dimensions of $\mathbf{y}_i$. However, for low dimensional data sets, the individual estimates are not accurate enough to be used as the weights and can only be used when averaged. Therefore this method is only useful for $\bar{\eta}$ and not ξ_i, with an overal saving of about 40% of the calculation time. This still meant weeks of processing time.

To be a viable analysis method, the processing time must be reduced to the hour range and so the non-iterative moment methods were investigated next. The covariance gives one moment and a kurtosis measure gives another. The first kurtosis formula essentially looked at each dimensions independently and was not particularly accurate. Further investigations found a better multivariate kurtosis expression, Mardia's multivariate kurtosis, which combines all the cross-terms of the $\mathbf{\Gamma}$ matrix and produces a better estimate. The moment method is a vast improvement in processing time and can process a 1000×1000 image area in about 5 minutes per model, which is a perfectly competitive time frame.

The accuracy of each measure was tested using simulated MK data by comparing the measured and simulated $(\alpha + 1)$ values (Figure 4.2). The iterative method (in red) has the most precision, but has noted bias to under-estimate for large α and fails to measure α down to -1, clipping off at zero. The simple kurtosis (blue) has a positive bias throughout and quite a large variance, whereas the improved Mardia's kurtosis method (green) roughly halves the bias and variance. It is important to realise that the significance of the error in α is not linearly spread and even though the variance looks very large for large α, it is still sufficient to capture the shape. For example, at the high end a measure of 20 ± 6 is still basically Gaussian, and at the lower end the error improves proportionally such that 2 ± 0.2 is probably distinct enough for classification purposes. It is probable that the bias in the iterative method for large α arises from the convergence test stopping before the parameters have tuned asymptotically to the Gaussian-like shape.

Figure 4.3 shows how the accuracy varies with the sample size, also for the three methods as above. Clearly at the value of $\alpha = 1$ the iterative method is most accurate and precise, and again the Mardia's kurtosis method has half the variance and bias of the simple kurtosis measure. There is an obvious trend of larger variance for smaller sample size, as would be expected, but the bias also

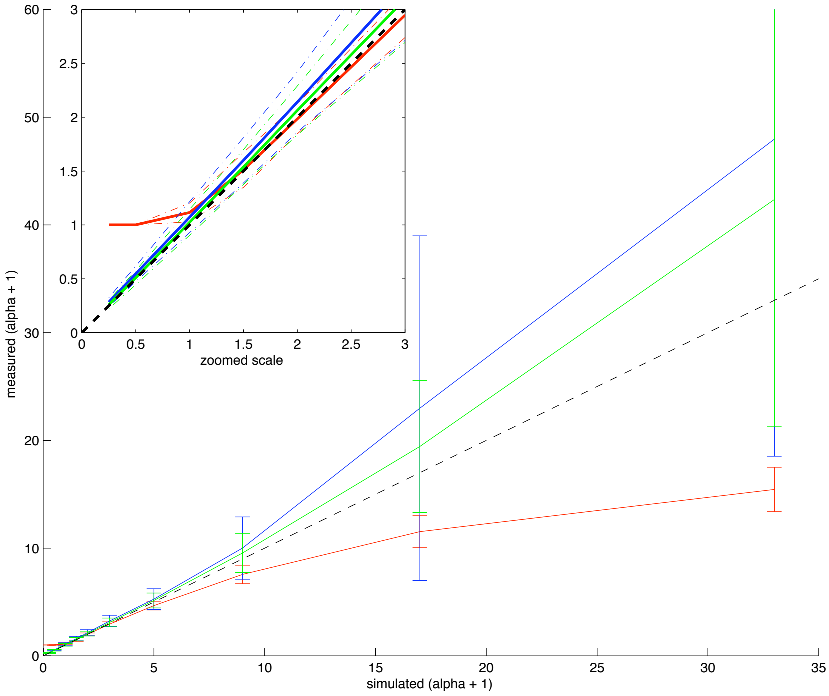


Figure 4.2: Three MK α accuracy measures, $N = 1000$. Dashed black line is ideal result. Iterative method in red, simple kurtosis in blue and Mardia's kurtosis in green. The inset shows the near zero region at a larger scale.

increases for smaller sample size. In all cases the bias for smaller sample sizes estimates towards the Gaussian, and is really just an indication that the higher order statistics need a larger number of samples to be distinguished. It is accepted that methods of moments only work well for large sample sizes, greater than 1000 [19], but Figure 4.3 shows that the variance does not increase dramatically until the sample size is reduced to below 300. This suggests that using the multivariate scale mixture of Gaussians model class reduces the required number of samples because each dimension effectively gives extra statistical information to the global parameters. The figure indicates that perhaps a minimum sample size of 169 may still be useful for the purposes of classification.

4.6 Applying Constraints

The parametric fitting routines used so far, found the maximum likelihood parameters given only the data points, and every parameter in the models was free to be optimised. It is normal for a random sample to vary slightly from the ideal dis-

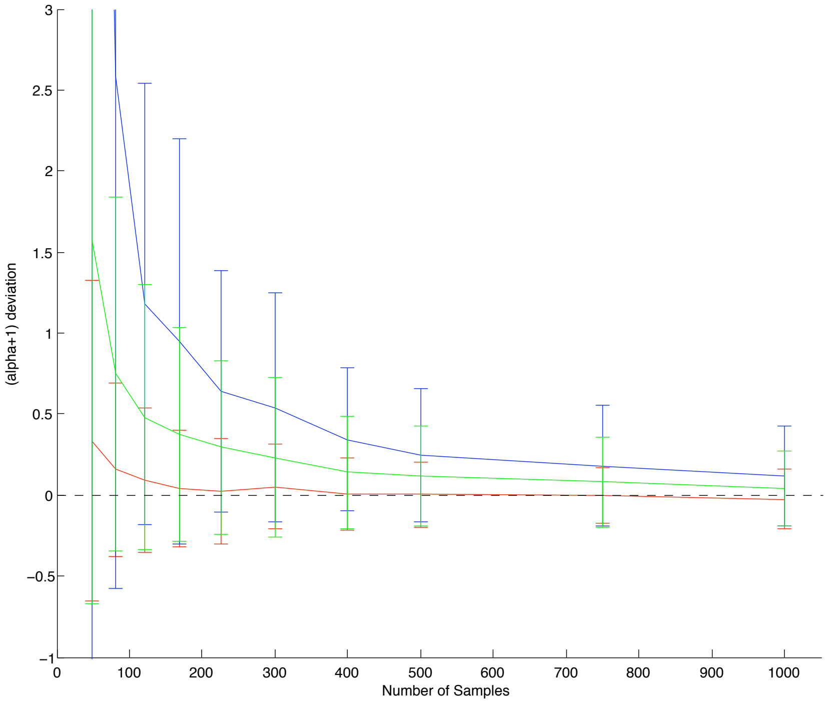


Figure 4.3: Three MK α accuracy measures versus sample size. Iterative method in red, simple kurtosis in blue and Mardia's kurtosis in green. $\alpha = 1, \lambda = 200$.

tribution, purely due to it being a random sample. For example, it is quite likely that a distribution that has exactly zero mean would in fact usually end up with a random sample mean that is nonzero. The closeness to zero being affected by the variance in the sample and the sample size, but rarely would all the random points average to exactly zero. Hence, *if there is some valid reason* that the model mean should be zero, constraining the mean to be exactly zero in the fitting routines would be expected to obtain a better model representation. That is, assuming the small variation from zero is just the random sampling and not part of the model. The mean is just one example of possible constraints, others include the covariance structure and the allowable range of certain parameter values.

It is important to note that there must be some underlying reason that the model should have an additional constraint applied, because otherwise applying the constraint would be incorrectly adding bias to the estimation. Only two additional constraints will be looked at in this study.

4.6.1 Zero Mean constraint

Scattering theory (Section 2.4) implies that the radar measurement of a distributed target is the sum of many individual scattered electromagnetic waves whose resulting signal has a random phase, somewhat like a random walk system [20]. Remembering that the data structure has been created by separating the real and imaginary components of the returned wave into their own dimensions, then the wave data is actually represented in a pair-wise manner. The pair-wise power of the signal is related to the scattering target, however the resulting power will be spread randomly between the real and imaginary parts of the signal. This implies that the (pair-wise) measured amplitude and phase should be independent and randomly distributed around a mean of zero.

The physical reasoning above means that a better estimate of our model parameters should be obtained by setting the mean vector to be exactly zero. This constraint is easily implemented by setting the mean $\boldsymbol{\mu}$ to zero throughout the fitting routines. The other parameters are then optimised based upon the mean being zero, for instance, the covariance matrix would be determined using $\mathcal{E}\{YY^T\}$ and so forth.

The SAR data sets tested did indeed have means very close to zero ($\ll 1\%$ of standard deviation) and this constraint seems to be quite valid. The effect of applying this constraint is likely to be more significant for smaller sample sizes.

4.6.2 Covariance Structure constraints

The covariance structure matrix, $\boldsymbol{\Gamma}$, also has constraints dictated by scattering theory. Being a covariance matrix, it must obviously be symmetric and in fact positive semi-definite. The pair-wise complex number representation with random phase leads to the requirement that each pair by pair minor matrix should have equal primary diagonal elements and zero off diagonal elements.

Some scattering assumptions, in particular certain assumed symmetries [21, 4], imply that the covariance of HH, HV and VV terms should only have off diagonal coupling for (HH, VV) pairs and (HV, VH) pairs, but not for HH and VV with either of VH or HV. Combining this with the complex separation into two real values, leads to a constrained covariance structure of the form

$$\begin{pmatrix} d_1 & 0 & 0 & 0 & 0 & 0 & c_1 & 0 \\ 0 & d_1 & 0 & 0 & 0 & 0 & 0 & c_1 \\ 0 & 0 & d_2 & 0 & c_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_2 & 0 & c_2 & 0 & 0 \\ 0 & 0 & c_2 & 0 & d_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_2 & 0 & d_2 & 0 & 0 \\ c_1 & 0 & 0 & 0 & 0 & 0 & d_3 & 0 \\ 0 & c_1 & 0 & 0 & 0 & 0 & 0 & d_3 \end{pmatrix} . \tag{4.49}$$

The covariance constraints are easily implemented in the routines, by taking the average of the identical terms, for each of d_1, d_2, d_3, c_1 and c_2, from the estimate for $\hat{\boldsymbol{\Gamma}}$, and simply discarding the zero terms. In fact a great deal of processing time can be saved by not calculating the zero terms in the first place. Although the zero-constrained terms are simply being ignored, they do influence any further estimates based upon the $\boldsymbol{\Gamma}$ matrix. For example, in the iterative method, the weights $\frac{1}{z_i}$ are estimated using $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, and the constraint will therefore be taken into account in the feedback process of convergence. In the moment method there is no feedback, however the estimates for $\mathcal{E}\{Z\}$ and $\mathcal{E}\{Z^2\}$ are evaluated using terms of $\boldsymbol{\Gamma}$.

Simulations indicate that the effect of applying the zero and equal pair constraints on the $\boldsymbol{\Gamma}$ estimate is quite striking. Figure 4.4 clearly shows two main points: the effect of the zero-mean constraint (green series) is only small, with slight improvement (compared with the unconstrained estimate in blue) for small sample sizes as predicted; and the effect of the $\boldsymbol{\Gamma}$ constraints (including zero-mean constraint) (in red) is significant throughout and particularly for small sample sizes. One implication is that, if valid, applying this constraint extends the accuracy of the estimation to much smaller sample sizes, removing much of the bias discussed in the previous section (4.5).

For the real test data studied, many of the expected zero terms in the covariance structure were found to be close to zero, although not all could be considered statistically zero (see Table 4.3). That is, terms of order 10^{-2} are likely to be just statistical sample variation, but terms of order 10^{-1} are perhaps too large to be discounted as sampling error. It appears therefore, that the assumption of pair-wise uncorrelated real and imaginary parts is likely to be correct. However the co-polar/cross-polar couplings are not expected from the simple scattering assumptions, so either the assumptions are too simplistic for real data or there is some systematic measurement error leading to this small cross-talk. Whatever the reason, the unexpected nonzero terms will not be investigated further. Although the zero-term constraints will not be applied to the routines to estimate $\mathbf{\Gamma}$, only the four main terms of $\mathbf{\Gamma}$, d_1, d_2, d_3 and c_1, will be taken, averaged over equal pairs, and used for initial classification purposes.

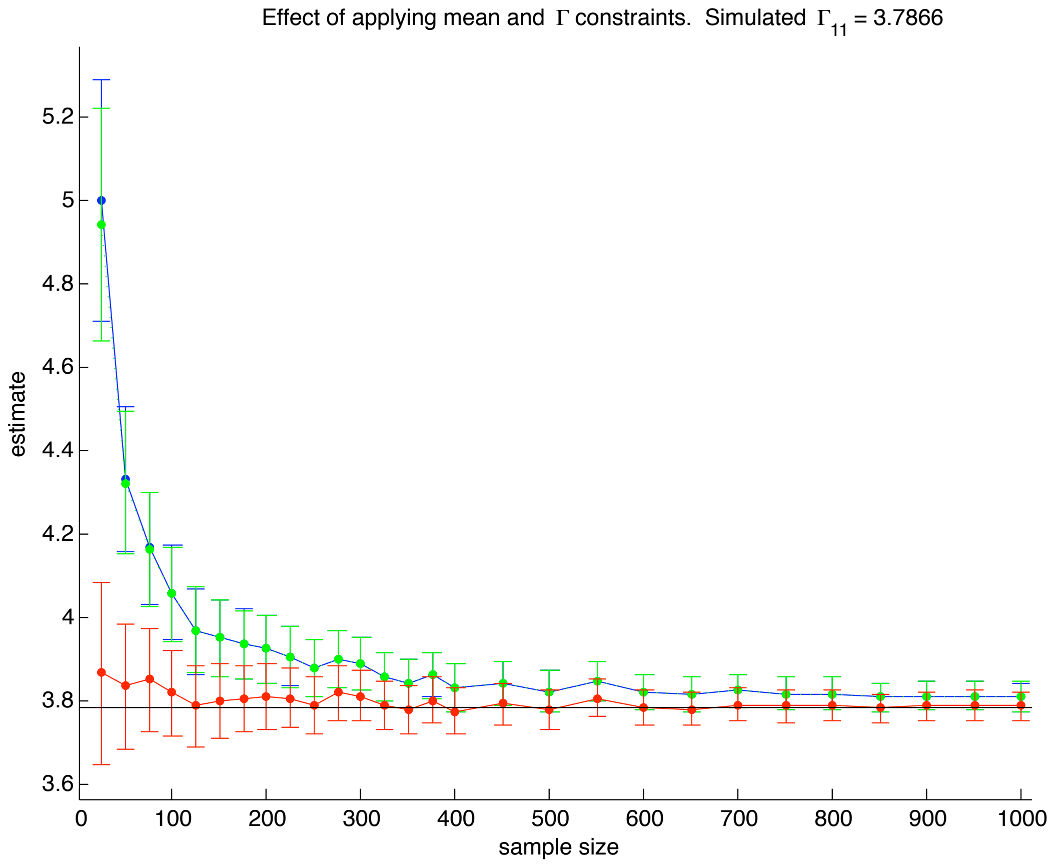


Figure 4.4: Effect of constraints on $\mathbf{\Gamma}_{11}$ estimation from simulated data. Unconstrained in blue, zero-mean constraint in green, and both mean and covariance constraints in red. Note that the green completely overprints the blue line above size 100.

Table 4.3: Covariance of entire data sets, for both C- and L-band data.

```
 C-band Covariance Matrix (normalised)
 4.0389   -0.0089    0.1151   -0.0284    0.0672    0.0210    2.2973   -0.0772
-0.0089    4.0385    0.0188    0.0983   -0.0287    0.0487    0.0788    2.2888
 0.1151    0.0188    0.8667   -0.0029    0.7791    0.0128    0.0520    0.0614
-0.0284    0.0983   -0.0029    0.8653   -0.0180    0.7766   -0.0676    0.0399
 0.0672   -0.0287    0.7791   -0.0180    0.8053   -0.0021    0.1051   -0.0026
 0.0210    0.0487    0.0128    0.7766   -0.0021    0.8024    0.0007    0.0924
 2.2973    0.0788    0.0520   -0.0676    0.1051    0.0007    4.1348    0.0103
-0.0772    2.2888    0.0614    0.0399   -0.0026    0.0924    0.0103    4.1426

 L-band Covariance Matrix (normalised)
 5.1651   -0.0011   -0.1426    0.1814   -0.1377    0.1803    2.1953    0.4538
-0.0011    5.1300   -0.1650   -0.1455   -0.1639   -0.1410   -0.4756    2.1458
-0.1426   -0.1650    1.0616    0.0027    1.0633    0.0160   -0.1203   -0.0524
 0.1814   -0.1455    0.0027    1.0617   -0.0104    1.0639    0.0708   -0.1204
-0.1377   -0.1639    1.0633   -0.0104    1.1065    0.0029   -0.1186   -0.0454
 0.1803   -0.1410    0.0160    1.0639    0.0029    1.1075    0.0630   -0.1183
 2.1953   -0.4756   -0.1203    0.0708   -0.1186    0.0630    5.4698   -0.0305
 0.4538    2.1458   -0.0524   -0.1204   -0.0454   -0.1183   -0.0305    5.4184
```

Chapter 5

Mixtures of several models

There are many reasons why the returned radar signal is not expected to comprise a single target scatterer type. For instance, the size of the radar resolution cell is usually large enough to cover the boundaries between, or enclose, several different types of scatterers. Likewise, the statistical sampling neighbourhood is even larger and may also cover boundaries and enclose more than one target type. A specific target type itself can often be considered to be a mixture of several pure scatterers. For example, a forest region will have characteristic target backscatter that is comprised of some part from leafy canopy scattering, some from vertically aligned trunk-ground scattering and some from rough ground scattering, and the returned signal will be some mixture of all three. Although scattering theory indicates that a large area pure homogenous scatterer should return a Gaussian distributed signal, a small area target or a mixture of several targets may certainly be non-Gaussian [22].

One method to address the idea of a mixture is to decompose the radar measurement into a set of basis components and determine the proportion of each present. There are several different methods to achieve this (a review is given in [21]) and each have different physical interpretations. With regards to the modelling presented here, two distinct approaches could be taken, either pre or post separation. Firstly, the data set could be viewed as a mixture of discrete populations, and separated in some maximum likelihood fashion, probably through iterations, into a sum of several individual pdfs. Either a simple mixture of Gaussians [16], or a more complex mixture of non-Gaussian pdfs. Secondly, the decomposition could be applied after the modelling to the estimated parameters, effectively on the sample averaged parameters. For instance the estimated covariance structure matrix, $\hat{\mathbf{\Gamma}}$, could be separated into basis components and proportion factors. Obviously, the decomposition theorem's assumptions and those of the scale mixture of Gaussians models must not contradict each other.

Although fascinating and potentially very powerful (like *a priori* constraints), these decompositions will not be investigated in this study, and instead the simple case of a scale mixture of Gaussians will be investigated purely for its own merit. Even so, a few of the key concepts of mixture models and their relation to the scale mixture of Gaussians class are discussed below.

5.1 Non-Gaussian modelling as a mixture of Gaussians

The common description of a mixture of Gaussians (called discrete mixture of Gaussians hereafter to avoid confusion with scale mixtures of Gaussians) means that the resulting distribution is the sum of several independent multivariate Gaussians. Each component has a mixture proportion factor plus separate mean and covariance parameters, such as

$$\text{MoG:}\quad f_Y(y) = \sum_{i=1}^{M} \pi_i f_{X_i}(x), \quad \text{where } f_{X_i}(x) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Gamma}_i) \text{ and } \sum_{i=1}^{M} \pi_i = 1. \tag{5.1}$$

Many non-Gaussian distributions can be approximated by discrete mixtures of several Gaussians, which may allow for easier numerical solutions since the Gaussian solutions are well known. As an example, it is quite easy to obtain a very non-Gaussian curve that is a mixture of a very broad and a very narrow Gaussian density with the same means, that combined look more Laplacian-like in appearance. Figure 5.1 is such a plot with a combined Parzen estimate in blue, the Gaussian parametric fit in black and the Laplacian parametric fit in red. The log-likelihood scores indicate that the Laplacian is the better fit (see Section 6.4).

Another example (Figure 5.2) is the two hump case (different means) where it is visually obvious that neither single model is appropriate to describe the data. In this example, it is obvious that the MK or MNIG distribution would also not fit any better (all being symmetric models single peake), and either a mixture of several models or a more complex single model would be required.

Clearly, a mixture of three Gaussians of different widths, but the same means, could make a smoother approximation to a Laplacian curve, and by using more and more components the approxmation could be improved still further. In fact, the scale mixture of Gaussians is exactly the limit of such a series of constrained discrete mixtures into a continuous mixture. That is, an infinite discrete mixture of Gaussians, with the added constraint that each has the same mean and covariance structure, and some particular distribution of widths. Therefore any of the non-Gaussian models could be approximated by a discrete mixture of Gaussians to any order desired, the more components the smoother the representation.

Two immediate questions arise. How many components are needed to adequately model a non-Gaussian model as a discrete mixture of Gaussians? And,

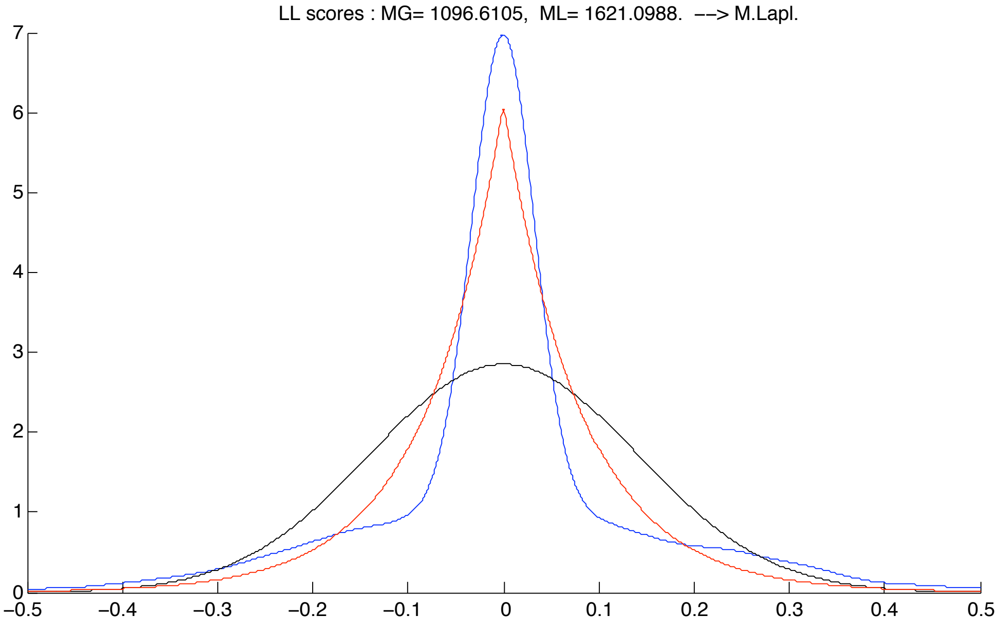


Figure 5.1: Co-incident mixture of two Gaussian densities (parzen estimate in blue), fitted as Gaussian (black) and Laplacian (red).

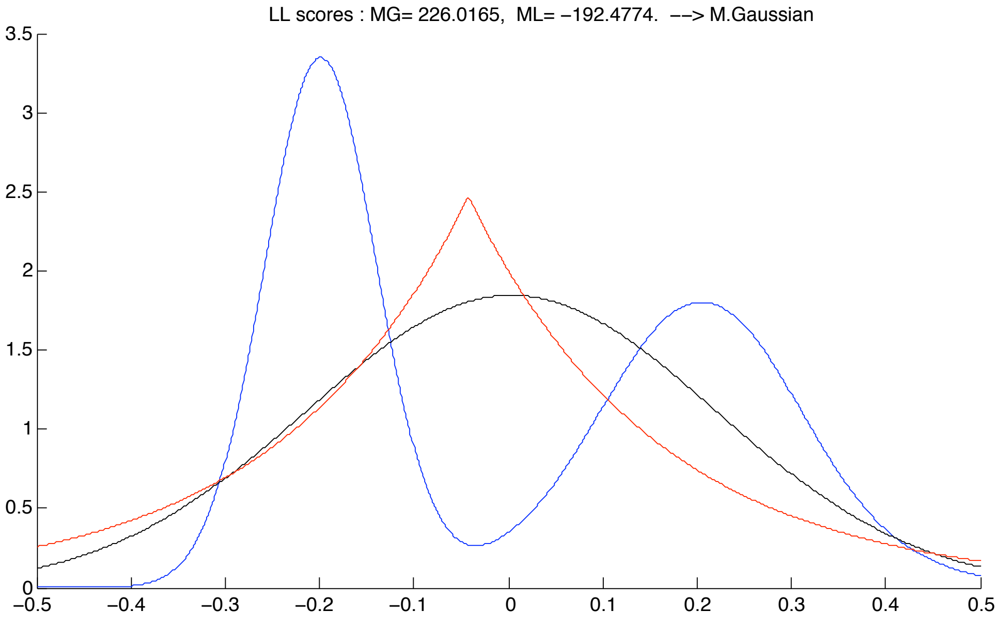


Figure 5.2: Separated mixture of two Gaussian densities, clearly neither individual model fits.

the converse, is the non-Gaussian looking data actually just a mixture of Gaussians, and if so, how many? Unfortunately there are no universal answers to these questions.

5.2 Number of parameters

The data could be modelled as a mix of, for example, five discrete Gaussians, and would then result in a set of three parameters for each component. That is, a proportion factor, a mean vector and a covariance matrix for each component, totalling 15 parameters. The scale mixture of Gaussians method, for instance the MK model, would result in just 4 parameters to describe the same data. Effectively, the comparative scales of each discrete Gaussian fit is compactly combined into the shape parameter of the model. Therefore, since no fundamental number of components is known in advance, it is simply more compact to use the scale mixture models rather than some chosen number of discrete mixtures to model the data. Subsequent classification is also somewhat simplified with one shape parameter, rather than several multidimensional terms.

5.3 Decomposition theorems

The main advantage of separating the data into mixture components would be to allow for decomposition theorems that require different means or covariance structures. For example, the forest interpretation as canopy, trunk and ground components is mostly distinguished by differing covariance structure [4]. Time has not permitted this decomposition to be investigated in this study. It could be implemented using a pre-modelling method that modifies the iterative discrete mixture of Gaussians algorithm to find three components with constraints on their covariance structure matrices. Or as a post-modelling method by separating the estimated covariance structure matrix into the three components. It seems likely that the proportion factors may be directly useful for classification.

5.4 Mixture testing

A practical observation from Figures 5.1 and 5.2 is the change of scale of the log-likelihood scores between the two examples. The data size is the same, yet the first example's scores are significantly larger than the second example's. The log-likelihood score in fact indicates not only which is best, but also whether it is a relatively good fit at all. Thus it is feasible to apply a mixture applicability test to determine if a single model is unlikely to describe the data, and then re-process as a mixture situation. See Section 6.4.3 for details about applying this measure.

The analysis of the SAR data sets indicated a clear Laplacian like distribution near all water-land boundaries. This is interpreted as the case in Figure 5.1 because the water has a very narrow Gaussian like distribution (low intensity) and the land in general has a very broad distribution (high intensity). This highly non-Gaussian mixture region becomes different classes with the simple classifying algorithms used since the shape feature changes so dramatically near the boundary.

It may be more appropriate to attempt to separate these area mixtures, perhaps choosing the component with the largest proportion factor as the representative for that image location. Choosing only the largest representative would then remove much of the blurring and re-classing around high contrast boundaries. The aforementioned mixture applicability test could be used to decide whether to process as a mixture. Very low shape parameter values may also be sufficient to use as a mixture test indicator. This of course, assumes that the data is basically Gaussian other than through area mixtures. Applying it to an homogenous area that happens to be very Laplacian like in distribution, conceivably obtained by flooded bush land for example, would in effect throw away all but the largest component, thus losing potential information. This approach would effectively find only the majority Gaussian scatterer for each neighbourhood, yet may actually be quite reasonable in terms of classification ability. Time has not permitted implementing and testing this technique in this study.

Chapter 6

Goodness-of-fit Testing

Best fitting the parametric models will always result in a solution for each, however it seems obvious that some will fit the data set better than others. Goodness-of-fit testing is a computational method of determining which model best describes the data set, based on some measure of similarity between the model and the data. The model that measures up best with the data set will be chosen to represent the data set statistics. Additionally, the differences between the measured values can give information about how similar the different models are.

Several different approaches are investigated in this study, from simple set distance measures to integrated squared error measures, to the log-likelihood function. The final choice considers the distinguishing ability, estimation accuracy and computation speed to find the most practical solution for testing goodness-of-fit for SAR data modelling.

6.1 Basic set similarity measures

A simplistic approach to testing the fit is to generate a set of points from the different models and compare them to the sample data set using some basic set measures (a general description is found in [23]). Since there is always some randomness in generating a set of data points, it is advisable to average the results from several generated sets. The basic set measures considered were: total summed Euclidean distance (between every point in the sample set and every point in the test set); total summed squared Euclidean distance; and summed maximum Euclidean distance.

The method was only tested between Gaussian and Laplacian models and although it achieved up to approx. 70% accuracy, it required a great deal of computation time and averaging over many trials. Such set measures really only test for compactness of the set, perhaps directly related to variance, which was sufficient to distinguish between the broad bell of the Gaussian and the tight point of the

Laplacian. It is doubtful that they could distinguish between more similar models.

This pure set measure method will not be considered further, because of its poor accuracy and time consuming nature.

6.2 Review of standard methods

The following is summarised from readings from many sources including general statistics course books (for example [24]), internet web searches (for example the Wikipedia, at www.wikipedia.org), and the documentation from the Statistical Toolkit (an open source package found at www.ge.infn.it/statisticaltoolkit).

The general technique used in the statistics literature is that of *hypothesis testing*. The usual two hypotheses are: the H_0 or null hypothesis that the two data sets are from the same population; and the H_1 or alternative hypothesis that they are from different populations. Probabilities are calculated and compared to standard tables and particular confidence levels to decide if H_0 is likely, or whether it should be discarded in place of H_1. Alternatively, the probability measures used can be compared to simply find the most likely alternative out of the choices measured.

Each of these tests compares a measure of the dissimilarity between the sample cumulative distribution function (cdf) F and the test cdf G. That is, when F and G are the same ($G \equiv F$) then the test measure (M) is minimum, usually 0, and when they are orthogonal to each other with no overlap ($G \perp F$) then M is maximum, often normalised to 1.

In the case of testing two (or more) models, e.g. G_1 to G_n, against an unknown sample distribution, F, then the goodness-of-fit test is a relative test where the individual measure with the lowest dissimilarity score is chosen as the best fit, that is

$$\text{choose } G_i \text{ such that } \quad M(G_i, F) < M(G_j, F) \quad \forall\, j \neq i \in 1, \ldots, n. \tag{6.1}$$

Distinguishing power is a relative measure of the usefulness of such tests, and is inversely proportional to the probability of false positive results, type 1 errors. A simple example is that a low power test might measure just one aspect of the functions, whereas a high power test might be based on an integration, and thus have distinguishing contributions from all values.

6.2.1 χ^2 test

The χ^2 test looks at only one aspect of the distribution, e.g. mean or variance, and therefore has low distinguishing power. It is a discrete test that requires counting data into bins, such that there are more than 5 counts per bin. It compares the total squared deviation from the expected count value divided by the expected value. The measure can be written as

$$M_{\chi^2} = \sum_{j=1}^{n} \frac{(y_j - E_j)^2}{E_j}, \tag{6.2}$$

where, y_j is a discrete measured aspect of the data, e.g. $y_j = \text{mean}(x_i \in \text{bin}_j)$ for n bins in the range of x_i, and E_j is the expected value for that aspect in the same bin.

6.2.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test measures the supremum (maximum) deviation of the cdfs,

$$M_{KS} = \sup\{|F - G|\} \tag{6.3}$$

It can be used on unbinned data and is considered medium power since it only tests the maximum point.

6.2.3 Cramer-von Mises test

The Cramer-von Mises test,

$$M_{CvM} = \int_{-\infty}^{\infty} (F - G)^2 f(x) dx, \tag{6.4}$$

is suitable for symmetric or right-skewed distributions, does not require binning data, is centre emphasising, because of the probability density function weight $f(x)$, and high power because it integrates over the whole range.

6.2.4 Anderson-Darling test

The Anderson-Darling test,

$$M_{AD} = \int_{-\infty}^{\infty} (F - G)^2 \frac{1}{F(1 - F)} f(x) dx, \tag{6.5}$$

also uses unbinned data, is tail emphasising, because of the division by $F(1 - F)$, and high power, due to the integration. This was considered the best test in the literature.

6.2.5 Comments

The main tests are based upon the cumulative distribution functions (cdfs), rather than probability density functions (pdfs), and so are not directly applicable to this study. Although not specifically mentioned in the literature, a purely graphical interpretation of these cdf tests implies that with minor modifications they may be equally applicable to the probability density functions.

Most of these tests involve an integral over the distributions and this would have to be estimated from the discrete sample of the data, possibly introducing estimation error or bias. The more powerful tests involve integrating the squared deviation with various weighting schemes to emphasise different regions.

In theory, *any* measure that gives a relative ranking can be used to distinguish the most likely distribution. However, different measures will have different ranges and possibly nonlinear sensitivity in their ranges due to different emphasis. Therefore, the ability to precisely estimate the scores may be just as important as the actual measure used.

In the following section, another measure is analysed that can be shown to be similar to the general integral form discussed above, but applies directly to densities and also has relations to entropy. Testing covers many of the problems that would come up in any of the above techniques, and so can be considered a proxy for any of the squared deviation integrals above.

6.3 Analysis of Cauchy-Schwarz measure

The Cauchy-Schwarz (CS) measure is a relatively new measure [25] and has not been used in this way before, and was tested primarily to see if it was useful. Investigations revealed the similarity to the Andersson-Darling and Cramer-von Mises tests and it is expected to have similar distinguishing power because of the integrated error basis.

The measure is based upon the Cauchy-Schwarz inequality

$$\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geq (\mathbf{x}^T\mathbf{y})^2, \tag{6.6}$$

with the vector inner product replaced with the density function inner product

$$< p, q >= \int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}. \tag{6.7}$$

The inequality can then be stated as

$$\int f^2(\mathbf{x})d\mathbf{x} \int g^2(\mathbf{x})d\mathbf{x} \geq \left(\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x} \right)^2, \tag{6.8}$$

and hence

$$\sqrt{\int f^2(\mathbf{x})d\mathbf{x} \int g^2(\mathbf{x})d\mathbf{x}} \geq \int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}, \tag{6.9}$$

and finally a Cauchy-Schwarz distance defined as

$$CS = \frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\sqrt{\int f^2(\mathbf{x})d\mathbf{x} \int g^2(\mathbf{x})d\mathbf{x}}} \leq 1. \tag{6.10}$$

In this case note that $f \equiv g \Rightarrow CS = 1$ and $f \perp g \Rightarrow CS = 0$ and it is therefore a similarity measure, bounded by [0, 1].

Taking the negative logarithm produces the dissimilarity measure

$$M_{CS} = -\log\left(\frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\sqrt{\int f^2(\mathbf{x})d\mathbf{x} \int g^2(\mathbf{x})d\mathbf{x}}}\right), \tag{6.11}$$

where $f \equiv g \Rightarrow M_{CS} = 0$ and $f \perp g \Rightarrow M_{CS} = \infty$.

Of course the logarithm of products, powers and divisions can be separated into

$$\begin{aligned} M_{CS} = & -\log\left(\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}\right) \\ & -\frac{1}{2}\left(-\log\left(\int f^2(\mathbf{x})d\mathbf{x}\right)\right) \\ & -\frac{1}{2}\left(-\log\left(\int g^2(\mathbf{x})d\mathbf{x}\right)\right), \end{aligned} \tag{6.12}$$

where the first term is known as the cross-entropy, and the other two terms are in fact the Renyi quadratic entropy [26] of f and g respectively.

6.3.1 An integrated squared deviation comparison

Consider the general form of an integrated squared deviation (ISD) measure, similar to the Cramer-von Mises measure, with even (flat) emphasis (i.e., without any $f(\mathbf{x})$ weighting).

$$\begin{aligned} ISD &= \int (f(\mathbf{x}) - g(\mathbf{x}))^2 d\mathbf{x} \\ &= \int (f^2(\mathbf{x}) + g^2(\mathbf{x}) - 2f(\mathbf{x})g(\mathbf{x}))d\mathbf{x} \\ &= \int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x} - 2\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{6.13}$$

The ISD must always be nonnegative because it is derived from the integral of a squared term. In fact $f \equiv g \Rightarrow ISD = 0$ and $f \perp g \Rightarrow ISD = \int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x}$.
So normalising the ISD by dividing by $\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x}$, bounds it in $[0, 1]$. Thus

$$\begin{aligned} ISD' &= \frac{\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x} - 2\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x}} \\ &= 1 - \frac{2\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})dx} \\ &= 1 - \frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\frac{1}{2}(\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x})} \qquad \geq 0. \end{aligned} \tag{6.14}$$

Hence, as a similarity measure

$$0 \leq \frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\frac{1}{2}(\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x})} \qquad \leq 1. \tag{6.15}$$

Taking the negative logarithm produces the dissimilarity measure

$$M_{ISD} = -\log\left(\frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\frac{1}{2}(\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x})}\right). \tag{6.16}$$

Note that all of the comparative information is contained in the upper cross integral and that the lower part is a normalisation factor that is in fact the *arithmetic mean* of the individual integrated squared densities.

Observe also the similarity to the Cauchy-Schwarz measure (6.11), which uses the same cross integral to determine the mutual relation between the two functions, but uses the *geometric mean* as the normalisation factor. Either factor is equally valid to normalise the measure to 1 when $f \equiv g$, because the cross integral becomes $\int f^2$ and the ISD normalisation becomes $\frac{1}{2}(\int f^2 + \int f^2) = \int f^2$, and the CS normalisation also becomes $\sqrt{\int f^2 \int f^2} = \int f^2$. There is actually a well known inequality between the geometric mean and the arithmetic mean

$$\textit{geometric mean} \quad \sqrt{XY} \quad \leq \quad \frac{1}{2}(X+Y) \quad \textit{arithmetic mean}, \tag{6.17}$$

with equality when $X = Y$. This relation leads to the bounding relation

$$\begin{aligned} M_{CS} &= -\log\left(\frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\sqrt{\int f^2(\mathbf{x})d\mathbf{x} \int g^2(\mathbf{x})d\mathbf{x}}}\right) \\ &\leq -\log\left(\frac{\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}}{\frac{1}{2}(\int f^2(\mathbf{x})d\mathbf{x} + \int g^2(\mathbf{x})d\mathbf{x})}\right) = M_{ISD}. \end{aligned} \tag{6.18}$$

One difference between the two measures will be the nonlinear scaling of the measured values, where one is geometrically scaled and the other arithmetically. The difference between the two being greater when the two functions have widely differing Renyi quadratic entropies, meaning that CS is less distinguishing in that case, but roughly equivalent when the entropies are similar.

However as ranking measures, the two are equally valid, and therefore the Cauchy-Schwarz measure is effectively based upon the integrated squared deviation, since both use the same cross integral for the comparative information.

6.3.2 Using the Cauchy-Schwarz measure

The CS measure requires calculation of both the cross integral and squared integrals of the two density distributions, however, in this study the sample distribution function is exactly the unknown that is yet to be determined. The second function is however known from the parametric model in question.

For goodness-of-fit ranking, between f and two or more models $g_1, \dots, g_n$, one half of the term under the square root, the $\int f^2$ part, cancels on both sides of the inequality and does not need to be calculated. The Cauchy-Schwarz goodness-of-fit test then becomes

$$\text{choose } g_i \text{ s.t. } -\log\left(\frac{\int f(\mathbf{x})g_i(\mathbf{x})d\mathbf{x}}{\sqrt{\int g_i^2(\mathbf{x})d\mathbf{x}}}\right) < -\log\left(\frac{\int f(\mathbf{x})g_j(\mathbf{x})d\mathbf{x}}{\sqrt{\int g_j^2(\mathbf{x})d\mathbf{x}}}\right) \forall j \neq i \in 1, \dots, n, \tag{6.19}$$

where $f(\mathbf{x})$ is the unknown sample density function.

Some simplifications and methods to determine the integrals are studied below. The dual Gaussian Parzen method, the large N approximation and numerical integrations.

Dual Gaussian Parzen

The unknown sample distribution can be estimated by a multivariate Gaussian Parzen windowing function, $w_h(\mathbf{x})$, with width parameter h, thus

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} w_h(\mathbf{x} - \mathbf{x}_i). \tag{6.20}$$

If the model function is used to generate a test data set which is also estimated with a Gaussian Parzen window, then the integration is reduced to a discrete double sum.

$$\begin{aligned}\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x} &= \int \left(\frac{1}{N_i} \sum_{i=1}^{N_i} w_{h_f}(\mathbf{x}-\mathbf{x}_i) \frac{1}{N_j} \sum_{j=1}^{N_j} w_{h_g}(\mathbf{x}-\mathbf{x}_j) \right) d\mathbf{x} \\ &= \int \left(\frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} w_{h_f}(\mathbf{x}-\mathbf{x}_i) w_{h_g}(\mathbf{x}-\mathbf{x}_j) \right) d\mathbf{x} \\ &= \frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left(\int w_{h_f}(\mathbf{x}-\mathbf{x}_i) w_{h_g}(\mathbf{x}-\mathbf{x}_j)\ d\mathbf{x} \right),\end{aligned}$$

and the integral can now be seen as a convolution of two Gaussians which results in another Gaussian whose width parameter is the sum of the individual width factors, $h_f + h_g$, and whose operand is the difference of the individual operands, $(\mathbf{x} - \mathbf{x}_j) - (\mathbf{x} - \mathbf{x}_i)$. Thus in the integral, the value x disappears and

$$\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x} = \frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} w_{(h_f+h_g)}(\mathbf{x}_i - \mathbf{x}_j).$$

Similarly, the lower integral becomes

$$\int g^2(\mathbf{x})d\mathbf{x} = \frac{1}{N_j^2} \sum_{j=1}^{N_j} \sum_{j'=1}^{N_j} w_{(2h_g)}(\mathbf{x}_j - \mathbf{x}_{j'}).$$

So the whole CS test between models g_1 and g_2 becomes finding which model g has the lowest measure

$$M_{CS} = -\log \left(\frac{\sum_{i=1}^{N_i} \sum_{j=1}^{N_j} w_{(h_f+h_g)}(\mathbf{x}_i - \mathbf{x}_j)}{\sqrt{\sum_{j=1}^{N_j} \sum_{j'=1}^{N_j} w_{(2h_g)}(\mathbf{x}_j - \mathbf{x}_{j'})}} \right), \tag{6.21}$$

where $w(\mathbf{x})$ is a multivariate Gaussian function with zero mean and scale factor h_f and h_g derived, for example using mean integrated squared error (MISE) [27] methods, from the sample data and test data respectively. The f^2 integral and factors N_i cancel out in the comparison.

The dual Gaussian Parzen technique involves no integration and the double sums are easy to compute with nested programming loops, however the requirement to simulate the model function introduces random variations in the test data set. The variations may be reduced by generating very large data sets, or averaging over several generated sets for each model. However, generating a test data sets from an explicit model function seems to be a step backwards when the parametric formula is known exactly.

Large N approximation

Consider the case when N is large, so that the set of samples $\mathbf{x}_i$ is smoothly representative of the distribution $f(\mathbf{x})$. In this case there will be approximately $Nf(\mathbf{x})d\mathbf{x}$ points $\mathbf{x}_i$ within the range $\mathbf{x}$ to $\mathbf{x}+d\mathbf{x}$ (for multivariate functions this is a volume element). The integration over the whole domain of $\mathbf{x}$ can then be approximated by grouping into bins of size $d\mathbf{x}$ with a flat value $f(\mathbf{x})$ in that bin. The cross-entropy integral then becomes a simple sum over the distribution points $\mathbf{x}_i$.

$$\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x} = \frac{1}{N}\int g(\mathbf{x})(Nf(\mathbf{x})d\mathbf{x}) \Longrightarrow \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i). \tag{6.22}$$

This can also be viewed as the expectation of $g(\mathbf{x})$ with respect to the distribution $f(\mathbf{x})$ and approximated by an average value.

The lower Renyi quadratic entropy integral must still be determined by either analytical integration (if possible) or through some numerical integration method.

Numerical integrations

The cross and Renyi quadratic entropy integrals must be determined in order to calculate the Cauchy-Schwarz measure, where the sample data is only known by a discrete sample set $\mathbf{X}$, and the test models are known through explicit parametric representations.

The dual Gaussian Parzen method simplified to a double sum, at the expense of simulating the test model data, and it was shown that for large N, the cross-entropy integral can be approximated by a simple sum over the sample set $\mathbf{X}$ (6.22).

The cross-entropy integral could also be estimated by a Parzen window function which could be inserted into the integration as a sum of Gaussians. This would be a smooth estimation of $f(\mathbf{x})$ even when N is not large, however it would clearly involve even more computation time.

Except for very simple cases, such as Uniform or Gaussian distributions, the Renyi entropy integral cannot be simplified analytically and Monte Carlo integration techniques must be utilised [28]. Either by generating a data set, $\mathbf{X}$, using the model distribution $g(\mathbf{x})$ and using the formula (6.22) above, or through the usual Monte Carlo integration method in which case

$$\int g^2(\mathbf{x})d\mathbf{x} \Longrightarrow \frac{V}{N_{MC}}\sum_{j=1}^{N_{MC}} g^2(\mathbf{x}_j), \tag{6.23}$$

where V is the size of the multidimensional volume over which the N_{MC} points are simulated. For best results the volume V, where the points are (uniformly) randomly generated, should be matched to the function in question. Ideally, chosen so as not to generate too many zero values, and yet still capture most of the curves details. The method becomes more accurate the larger N_{MC} becomes.

In this case, where $\int g^2(\mathbf{x})d\mathbf{x}$ is desired, the constraint that $\int g(\mathbf{x}) = 1$ can be used to advantage. The Monte Carlo technique can be looped in, for example, 20000 points at a time, noting both the sum of $g(\mathbf{x})$ and $g^2(\mathbf{x})$, and take the value of $g^2(\mathbf{x})$ as soon as the integral of $g(\mathbf{x})$ evaluates to within some set tolerance of 1. It was found, however, that the number of loops required varied enormously, purely dependent on the random scatter of generated points.

6.4 Analysis of Log-likelihood measure

The likelihood function [23] evaluates the probability of obtaining the particular sample data under some assumed distribution function. It therefore gives a probability in the range of zero to one, with zero being impossible and one being a certainty. The log-likelihood simply takes the logarithm thus re-scaling it from minus infinity to zero without changing any relative ordering since the logarithm is a monotonic function. Therefore, choosing the model that achieves the greatest log-likelihood score, maximum likelihood, is essentially choosing the most likely model that generated the data set. As shown below, the log-likelihood score is related to entropy too, and the concept of maximum entropy is a whole field of research in itself, proclaiming that the maximum entropy solution is the solution that applies the least number of additional constraints upon the data.

The likelihood function for a particular data set $\mathbf{X} = \{\mathbf{x}_i\}$ and a distribution $f(\mathbf{x})$ is given by the probability of that combination of $\mathbf{x}_i$ given $f(\mathbf{x})$, which is

$$L(\mathbf{X}, f) = \prod_{i=1}^{N} f(\mathbf{x}_i). \tag{6.24}$$

It is often simpler to take the log-likelihood function as a similarity measure, because the product becomes a sum

$$LL(\mathbf{X}, f) = \log\left(\prod_{i=1}^{N} f(\mathbf{x}_i)\right) = \sum_{i=1}^{N} \log\big(f(\mathbf{x}_i)\big) = M_{LL}. \tag{6.25}$$

A goodness-of-fit test is then the Log-Likelihood Ratio (LLR) test

$$\begin{aligned} LLR(g_1, g_2) = & \log\left(\frac{L[\mathbf{X}, g_1]}{L[\mathbf{X}, g_2]}\right) \\ = & \sum_{i=1}^{N} \log\big(g_1(\mathbf{x}_i)\big) - \sum_{i=1}^{N} \log\big(g_2(\mathbf{x}_i)\big), \end{aligned} \tag{6.26}$$

where a positive result favours g_1 and a negative result favours g_2

The form of this test makes it applicable to testing several functions at a time, namely g_1 to g_n, by choosing the function, g_j, with the largest individual log-likelihood measure (6.25). That is

$$\text{choose } g_j \text{ s.t.} \quad \sum_{i=1}^{N} \log\big(g_j(\mathbf{x}_i)\big) > \sum_{i=1}^{N} \log\big(g_k(\mathbf{x}_i)\big) \ \forall\ k \neq j \in 1, \ldots, n. \tag{6.27}$$

It is immediately obvious that this method requires no data simulations nor integrations. Simply evaluate the model function at each of the sample data points, then log and sum the results. Very fast and efficient.

So the question then is whether this is sufficient? How does it compare to the CS measure or other ISD measures?

6.4.1 LLR as integrated difference

Observe that the LLR (6.26) can be regrouped as

$$LLR(g_1, g_2) = \sum_{i=1}^{N} \Big(\log\big(g_1(\mathbf{x}_i)\big) - \log\big(g_2(\mathbf{x}_i)\big) \Big), \tag{6.28}$$

and for large N, with $\mathbf{x}_i$ representing the distribution $f(\mathbf{x})$, there will be approximately $f(\mathbf{x})N d\mathbf{x}$ points $\mathbf{x}_i$ in the volume between $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$. Therefore the sum over points $\mathbf{x}_i$ tends to a weighted integral

$$LLR(g_1, g_2) \Longrightarrow N \int \Big(\log\big(g_1(\mathbf{x})\big) - \log\big(g_2(\mathbf{x})\big) \Big) f(\mathbf{x}) d\mathbf{x}. \tag{6.29}$$

The log-likelihood measure then represents the difference in the logarithms of the functions, weighted by the sample density, and is clearly not the same form as the squared deviation integrals.

A simple interpretation is that everywhere that $g_1(\mathbf{x})$ is greater than $g_2(\mathbf{x})$ will add to the integral and where $g_2(\mathbf{x}) > g_1(\mathbf{x})$ will subtract. So the overall sign of the integral will show which one won the tug-of-war. Of course both must integrate to 1, being pdfs, and so this test really shows which curve is higher in more places (this already hints towards the concept of entropy). The log scale is likely to really enhance the differences in the tails, where log will get very negative. Also note that the weighting term, $f(\mathbf{x})$, will emphasise the middle of the actual sample distribution, so this measure has a strange mix of centre and tail weights and is likely to have good sensitivity to differences throughout the curves.

The initial factor N is obvious from the original form of log-likelihood as the combined probability of obtaining exactly those values $\mathbf{x}_i$, and will obviously vary with N. Perhaps a more useful measure would be a standardised log-likelihood (SLL) function defined as

$$SLL(\mathbf{X}, g) = \frac{1}{N}\sum_{i=1}^{N} \log\big(g(\mathbf{x}_i)\big) \Longrightarrow \int \log\big(g(\mathbf{x})\big) f(\mathbf{x}) d\mathbf{x}, \tag{6.30}$$

which could compare values independent of the sample size N. However, this is not important in this study, since we only have the single sample set of size N for each test model (See also Section 6.4.3).

This integral (6.30) is a form of Shannon entropy [29], or actually a cross-entropy. Choosing the largest log-likelihood measure, is therefore choosing the distribution with the maximum Shannon entropy consistent with the data set. This also means the distribution with the largest remaining uncertainty, and therefore the least number of additional assumptions, other than the data points, have been introduced.

The integrated difference in log densities weighted by the sample density (6.29), can also be interpreted as the expectation of $\log\big(\frac{g_1(\mathbf{x})}{g_2(\mathbf{x})}\big)$ over the sample distribution $f(\mathbf{x})$, and a positive result implies g_1 has a greater expectation then g_2, and negative the converse.

6.4.2 Relation to Kullback-Liebler divergence

The log-likelihood measure is in fact half of the well known Kullback-Leibler (KL) divergence, whose continuous form is defined as

$$M_{KL}(f, g) = \int_{-\infty}^{\infty} f(\mathbf{x}) \ \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) d\mathbf{x}, \tag{6.31}$$

with properties: $M_{KL}(f, g) \geq 0$, and $M_{KL}(f, g) \equiv 0 \Leftrightarrow f \equiv g$. Note, however, that $M_{KL}(f, g) \neq M_{KL}(g, f)$.

The Kullback-Leibler divergence can be expanded thus

$$\begin{aligned} M_{KL}(f, g) &= \Big(- \int f(\mathbf{x}) \ \log\big(g(\mathbf{x})\big) d\mathbf{x}\Big) - \Big(- \int f(\mathbf{x}) \ \log\big(f(\mathbf{x})\big) d\mathbf{x}\Big) \\ &= H(f, g) - H(f), \end{aligned} \tag{6.32}$$

where $H(f) = -\int f(\mathbf{x}) \ \log\big(f(\mathbf{x})\big) d\mathbf{x}$ is Shannon's entropy, and $H(f, g)$ is a Shannon-like cross-entropy, that is exactly the (negative of the) cross-entropy of the standardised log-likelihood measure (6.30).

Note that, in the case of comparing $M_{KL}(f, g_1)$ with $M_{KL}(f, g_2)$, the common terms $H(f)$ and scale N will cancel out. Hence the log-likelihood ratio test is precisely equivalent to using the (negative) Kullback-Leibler divergence as a similarity measure.

6.4.3 Good-fit threshold test

The standardised log-likelihood measure (6.30) has a useful application. Since it is standardised to the sample size, N, the absolute value of the score can be used as an absolute test of good or bad fitting. All close fits will have a relatively high score, and really bad fits a very low or negative score. A simple threshold may be a sufficiently useful test, for instance to distinguish between a single model or a mixture of several models. The example in Section 5.1 with the two component mixture is a clear case where a threshold could be uesful. However, it is not enough to simply use the standardised log-likelihood score, as that too varies with the shape of the actual data, in fact its Shannon entropy. To standardise this as well, requires actually computing the full Kullback-Leibler divergence.

To compute the KL divergence requires the Shannon entropy of the sample distribution, $-\int f(\mathbf{x})\ \log\ f(\mathbf{x})\ d\mathbf{x}$, which can be computed numerically with Monte Carlo integration methods. In this study the sample distribution $f(\mathbf{x})$ was estimated with a Gaussian Parzen window function to obtain a smooth function for Monte Carlo integration. Again, the intermediate computation of $\int f(\mathbf{x})\ d\mathbf{x}$ can be tested against an ideal value of 1, with some pre-set accuracy, to determine convergence of the entropy integration. These integration methods are rather computational and therefore very slow, but are at least a simple procedural method for multi-dimensional integrals.

An empirical threshold was determined (for 1-D data and 1000 samples) at a level around 0.02 that distinguished even a 5% mixture. This level clearly distinguishes when the sample contains two or more humps with different means, but is less successful when the means are the same. This is mainly due to the random sample variation obscuring the subtle slope change of the mixture, and is improved by taking a much larger sample set to get a smoother sample distribution. The threshold value is different for higher dimensions and varies greatly with sample size. For 1-D data of only 169 samples, the level required was more like 0.1, and a value around 0.4 was needed to for 169 by 8-D data. The level primarily must be larger then the variance expected from repeated measurements of random samples from the same distribution. For instance a level at 2 or 3 standard deviations above the log-likelihood score baseline for that sample size worked well. Figure 6.1 shows the variance levels versus sample size for both 1-D and 8-D data simulations.

Another interesting application of the threshold test is to help determine how many mixtures in discrete mixture of Gaussian modelling are required to fit the mixed data. The approach taken was to successively fit more and more Gaussian mixtures until the KL divergence measure was below the threshold value and could be considered a good fit. The resulting algorithm was quite successful when the sample size was large enough to give a smooth representation of the components, for example, > 200 for each component. It was, however, extremely time

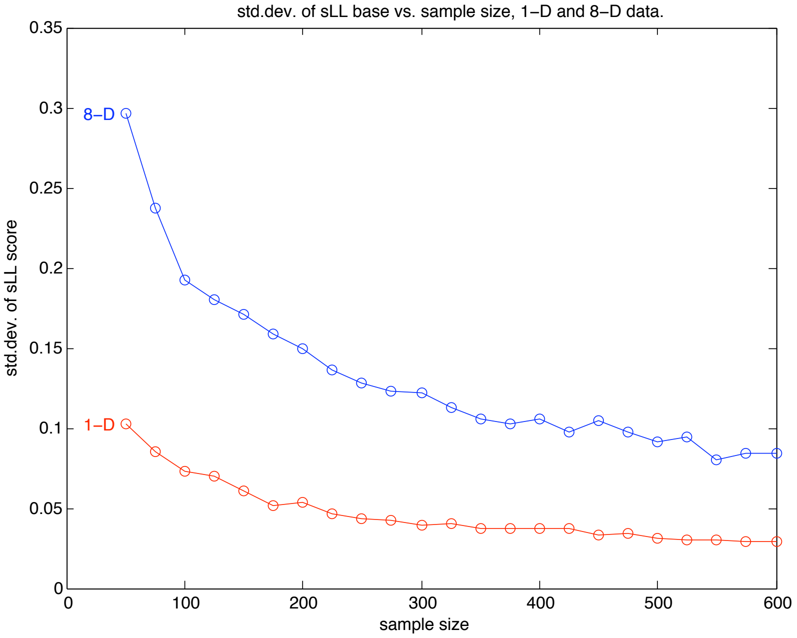


Figure 6.1: Standard deviation of standardised log-likelihood scores for repeated random samples versus sample size.

consuming and not practical for image analysis use.

An faster alternative method is to use the simple log-likelihood score and compare its value when one extra component is added. This means that the normalising part is not required because the test compares function measures over the same sample data. The normalising component is just a fixed offset dependent on the actual sample curve but not the test function. The procedure would first model as one Gaussian, and then successively add an extra component and compare the scores. If the scores are significantly different, again using the threshold value, then keep going with another component. Otherwise revert to the previous number of mixtures, since the additional component was not a significant improvement. This method was fast and worked quite well, usually correctly finding same mean mixtures above 10%. Figure 6.2 shows a progressive mixture of 2 Gaussians (in 5% steps) being tested as a single, double or triple mixture. The standardised log-likelihood scores are relative to the maximum and the threshold line is shown. The threshold clearly separated even a 5% mixture, for 8-D by 625 samples data. It is clearly seen that the third component did not make any improvement, and the mixture of Gaussian algorithm essentially splits one Gaussian into two parts to make the third component, but the test score remains the same (within variance).

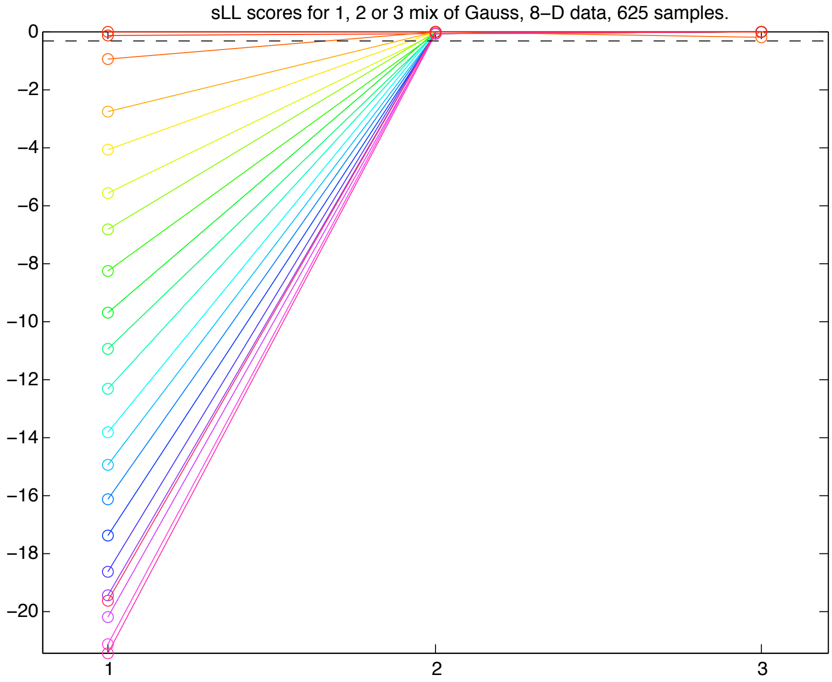


Figure 6.2: Standardised Log-likelihood scores for 1, 2 or 3 mixtures fitted to a progressive mixture of two Gaussians with the same means. Threshold level shown as dashed black line.

6.5 Comparisons

6.5.1 Validation tests

Validation tests can be carried out by first generating data from a known test distribution, best fitting each of the test models, and applying the goodness-of-fit test to them. The validity test is considered correct if it chooses the same original model that was generated. This can be repeated for all models, and with different parameters to build up some statistics about how accuratly the goodness-of-fit test works.

The log-likelihood and integrated CS measures were tested in this way. The integrated CS evaluation was undertaken using the Monte Carlo technique for calculating the Renyi quadratic integral and the large N approximation used for the cross-entropy integral. The parameters were chosen randomly within ranges that gave similar width distributions for each model and also represented a large range of shapes. The width range was chosen to closely match that found in the test SAR data. Both small and large sample sizes, 169 and 10000, were used for these tests, as it is expected that the accuracy varies with sample size N.

Confusion matrices depict columns for each of the generating data sets (Gaussian, Laplacian, K-dist. and NIG in that order), and rows for the chosen test model (in the same order). Thus an ideal test would produce N down all the diagonals and zeros elsewhere. In reality (Table 6.1), we see the number of correctly measured tests (out of 10 for each model) down the diagonal and where they were mistaken for other models on the off diagonals. A crude measure of accuracy is the sum of the diagonal elements divided by the total number of tests, 40 in this case. This score is noted as “diag” under the matrices, along with the computation time.

These matrices show several main points. Firstly, for the 1 dimensional simulation the two methods are roughly equivalent in accuracy, yet in the 8 dimensional case, the Cauchy-Schwarz measure suffers badly, presumably from poor integration estimates. Secondly, the large sample case does seem to concentrate the scores with a few 9/10’s and more zeros, as compared to the small sample case where random sampling may significantly mis-shape the particular distribution instance. Thirdly, the Cauchy-Schwarz implementation takes a lot more time to compute than the log-likelihood measure. A factor of 100 times longer in 1-D, and up to 1000 times longer in 8-D. Lastly, some of the model’s shape relations are visible in the matrices. That is, the Gaussian and Laplacian are never confused, yet the MK and MNIG data are confused with all models, and the MK and MNIG models may score highly on any data set.

The last point is better demonstrated by looking at the graphs of the scores (Figures 6.3 to 6.6). These graphs depict the relative test scores for each of the

Table 6.1: Confusion matrices for log-likelihood measure and integrated Cauchy-Schwarz measure, both 1-D and 8-D by both 169 and 10000 samples.

```
1-D, N = 169
Log-Likelihood measure          Cauchy-Schwarz measure (0.5% integral)
    G     L     K     N                   G     L     K     N
G   6     0     1     2               G   7     0     2     1
L   0     6     2     2               L   0     7     3     2
K   2     3     4     4               K   1     3     2     3
N   2     1     3     2               N   2     0     3     4

diag 18/40, time 0.80187 s.     diag 20/40, time 123.4751 s.

1-D, N = 10000;
Log-Likelihood measure          Cauchy-Schwarz measure (0.5% integral)
    G     L     K     N                   G     L     K     N
G   3     0     0     1               G   3     0     0     2
L   0     9     0     0               L   0     5     0     0
K   7     1     7     6               K   3     5     9     4
N   0     0     3     3               N   4     0     1     4

diag 22/40, time 9.3862 s.      diag 21/40, time 150.2577 s.

8-D, N = 169
Log-Likelihood measure          Cauchy-Schwarz measure (0.5% integral)
    G     L     K     N                   G     L     K     N
G   7     0     0     1               G   2     0     1     2
L   0     5     1     4               L   5     7     5     6
K   1     5     7     4               K   2     3     2     1
N   2     0     2     1               N   1     0     2     1

diag 20/40, time 0.69424 s.     diag 12/40, time 15270.7484 s.

8-D, N = 10000;
Log-Likelihood measure          Cauchy-Schwarz measure (0.5% integral)
    G     L     K     N                   G     L     K     N
G   4     0     0     1               G   0     0     0     0
L   0     8     0     7               L   3     3     8     9
K   0     2     9     0               K   2     7     1     1
N   6     0     1     2               N   5     0     1     0

diag 23/40, time 18.7012 s.     diag 4/40, time 15411.1531 s.
```

tests used to compile the confusion matrices. They plot the difference of each model's score from the top score for each test, divided by the top score. The colouring in each plot is the same with the MG score in black, ML in green, MK in red and MNIG in light blue. There are four model score lines for each of the four generated data sets, although colours plotted last may overprint others. The ideal situation would be a straight line at zero for the data model's colour and good separation from the others.

The figures show that often several scores are very similar, particularly the red and blue of the MK and MNIG. This is not a problem with the separability of the measure itself, but an indication that several models have very similar density functions, and have adjusted their parameters to closely fit the data set. The black and green of the MG and ML are usually well separated and change order for MK and MNIG data depending on the particular value of the shape parameter that was chosen randomly. The MK and MNIG lines are nearly always close to the maximum value in all data sets, indicating that both models are flexible enough to closely fit all shapes. Note that the few out of place points in the 1-D small sample case may be due to random sample variation, for instance the Laplacian data point number 7 seems to be more Gaussian than Laplacian. The few odd results in the 1-D Cauchy-Schwarz implementation are most likely due to poor integration estimates. Observe that both the Log-likelihood and Cauchy-Schwarz measures give virtually the same rankings throughout. The 8-D data tests show markedly more variation, but overall the same picture. Laplacian data clearly measures best with the ML and MK models, and is clearly separated from MG which is often off the bottom of the Laplacian plot. The MK model is often quite a good fit for any of the model's data sets.

Note that the CS integration tolerance of 0.5% was not always achieved within the pre-set maximum number of loops, resulting in great variance and occasional strange scores. This was particularly a problem for the multidimensional data, and then especially for the highly pointed data, Laplacian like, because of the huge numerical range and occasional overflow values.

Note also, that the data generation for MNIG also uses a Monte Carlo method to estimate the inverse Gaussian scale parameter and may suffer from numerical inaccuracies relating to finite bounds. It is conceivable that some of the confusion between the 8-D MNIG data and the ML model may arise if the generating bounds were too small, generating more peak values and less extended tails than true MNIG data.

Another numerical note is that the MK fitting routine was set to limit α to 100 (to restrict the range for later segmentation purposes), which would mean it may not achieve as good a score as expected for the large sample Gaussian like data sets. The MNIG was not limited and thus a good deal of MG data sets scored best with MNIG.

6.5.2 Choice of method

The clear winner is the log-likelihood measure, being the most accurate at higher dimensions, as well as the faster to compute. The slow speed and inherent variation with Monte Carlo integration simply make that method, and others based on integrations, impractical.

Of considerable value, is the fact that several models have a large range of shapes in their parameter space. In particular with Laplacian generated data, the K-distribution and the Laplacian models are virtually the same. In fact the K-distribution can become the Laplacian with the parameter $\alpha = 0$, and so it is not unexpected that they measure up so closely. At the other extreme, with Gaussian data, the K-distribution and the NIG distribution can both have parameters such that they tend towards the Gaussian distribution, and so would also give almost equal measures.

This last fact can be used to advantage by doing away with the need to do several tests, and simply using the K-distribution or NIG distribution for everything. The evidence is that they can clearly replace the Laplacian and the Gaussian with suitable ranges in their parameters, and the two also closely matched in most tests. Some literature suggests that there is some physical basis for SAR backscattering to be K-distributed [2]. The NIG model has a potential advantage that it is never infinite at the peak centre as would be the case with any real data sample, however numerical problems in generating MNIG data mean it cannot be tested as well as the MK model.

Using only one model would certainly simplify the latter stages of trying to classify the image based upon the parameters, as having several models would involve needing several classification schemes, or a branched scheme, and the inherent difficulties in matching up classes from each part.

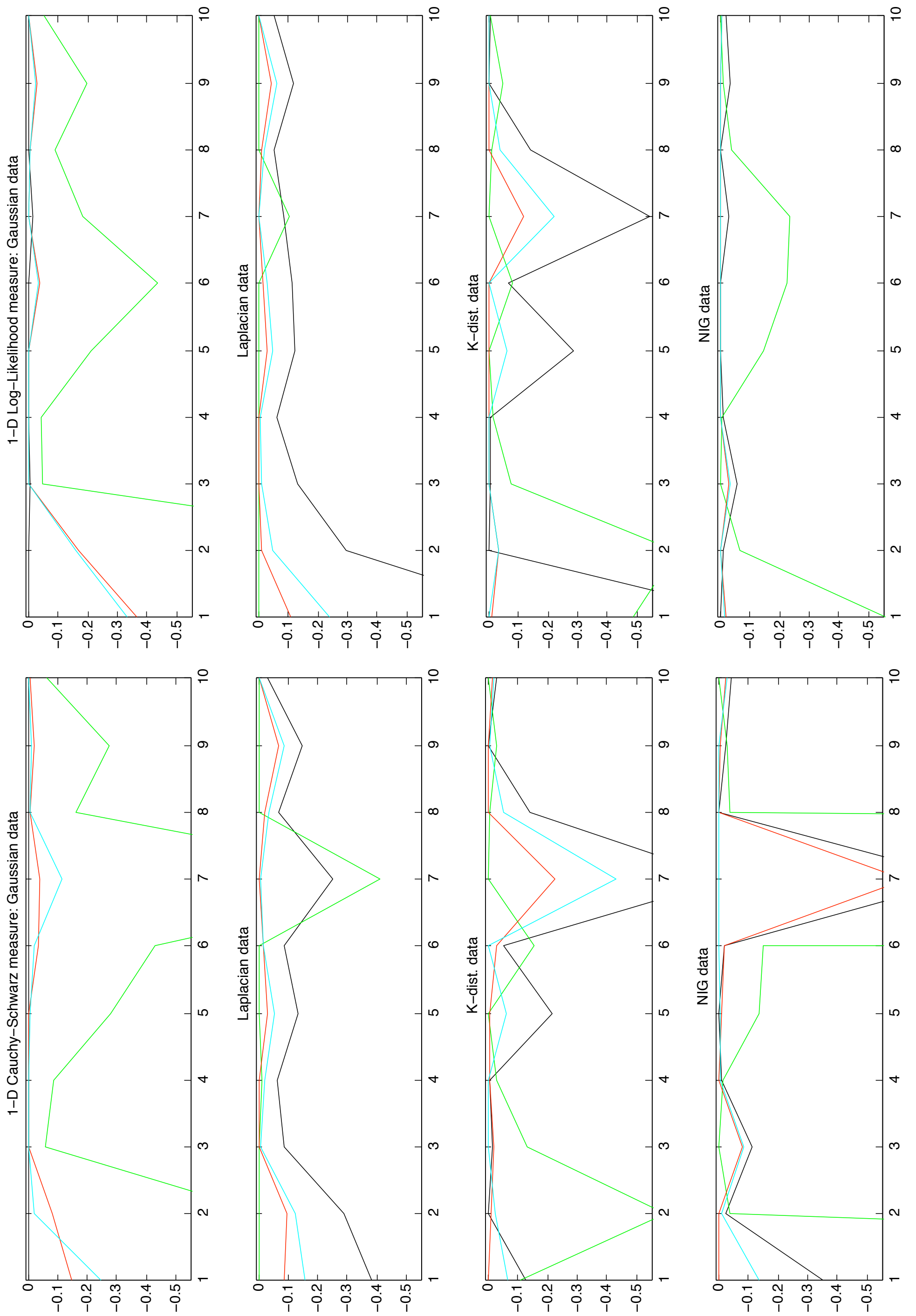


Figure 6.3: Log-likelihood and Cauchy-Schwarz integrated test scores for randomly generated 1-D data with sample size of 169.

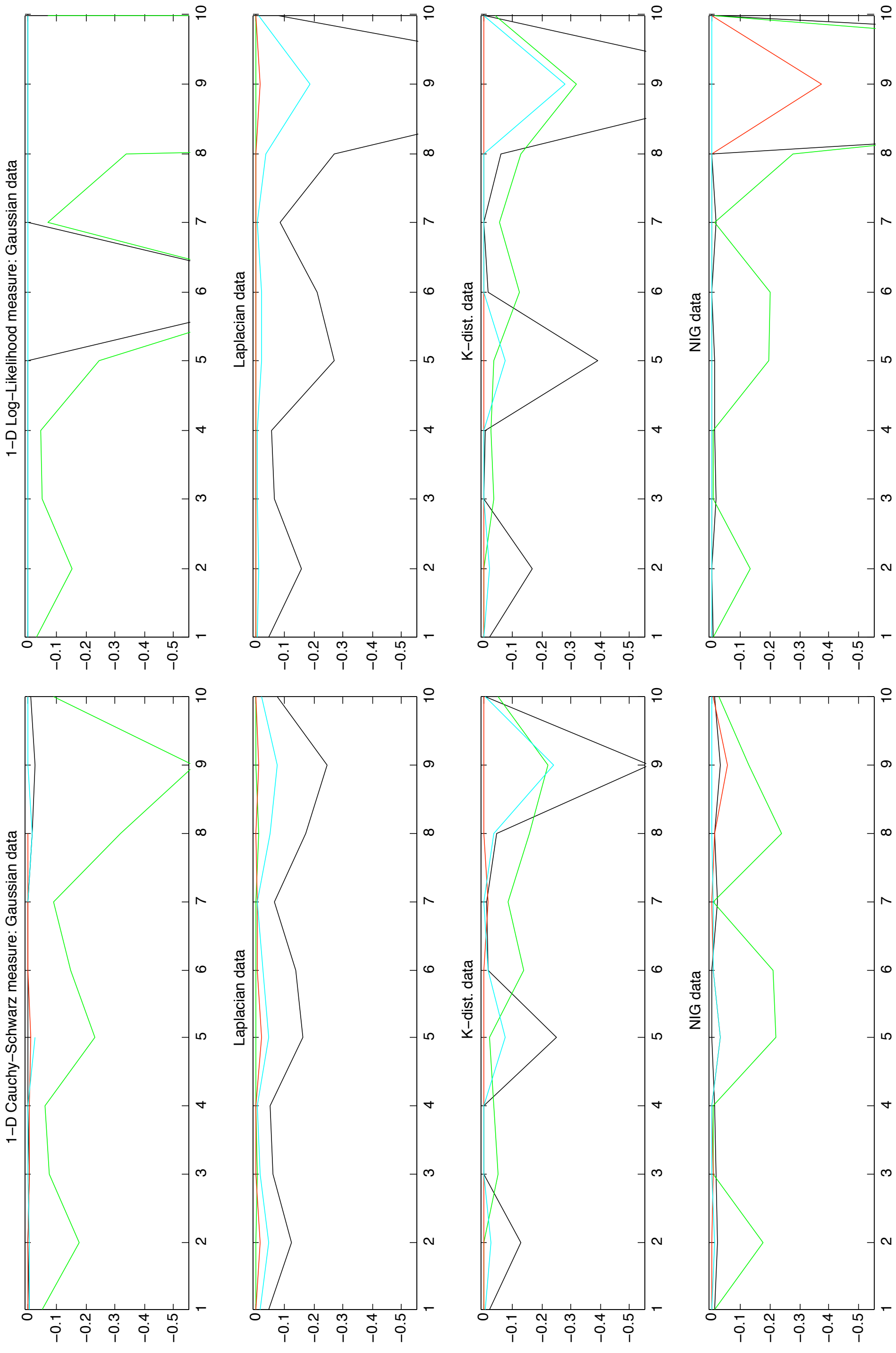


Figure 6.4: Log-likelihood and Cauchy-Schwarz integrated test scores for randomly generated 1-D data with sample size of 10000.

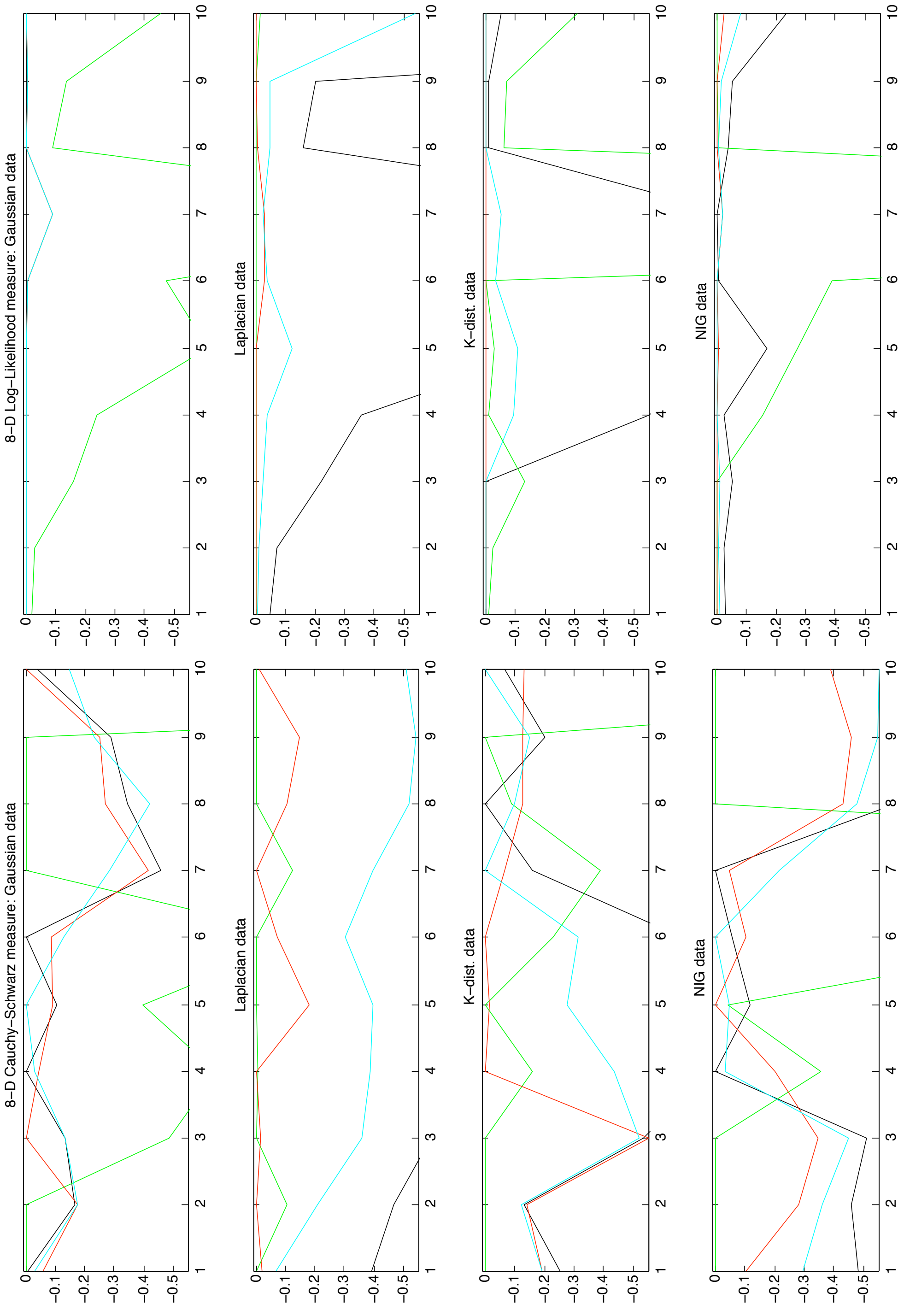


Figure 6.5: Log-likelihood and Cauchy-Schwarz integrated test scores for randomly generated 8-D data with sample size of 169.

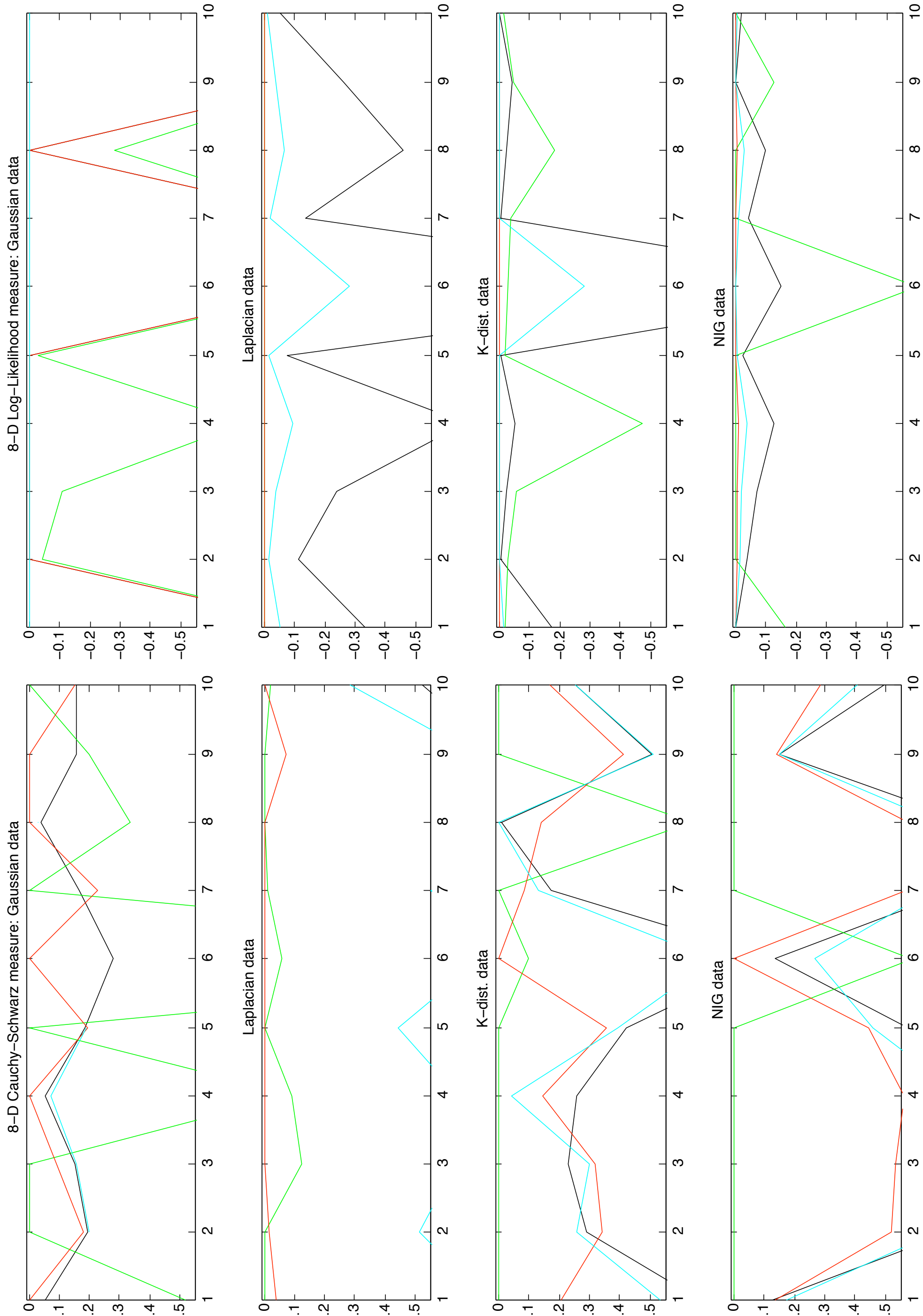


Figure 6.6: Log-likelihood and Cauchy-Schwarz integrated test scores for randomly generated 8-D data with sample size of 10000.

Chapter 7

PolSAR data tests

A test area of 1000 × 1000 pixels was chosen from a fully polarimetric single-look complex scatter matrix image. The region contains both water and land and comes from an EMISAR flight in July 1995 over Bleikvatnet, Norway. The data was transformed into the 8 dimensional structure as described in Section 2.6. An RGB intensity image is shown in Figure 7.1 to help understand the regions. The channels are R $= \sqrt{|S_{hv}|}$, G $= \sqrt{|S_{hh}|}$, B $= \sqrt{|S_{vv}|}$, the square root enhances the brightness. The water is the dark, low intensity, areas and the land is the more structured brighter areas.

The polarimetric SAR data can now be modelled with all four models and log-likelihood goodness tested to see which model is best. A moderate sample size of 441 (21 × 21) was used to avoid the heavy Gaussian bias of small sample sizes. Of particular interest here is whether a single flexible model does fit real SAR data well, so all model scores will be compared to the highest scored model. A good/bad threshold value of 0.5% seemed appropriate as it usually separated models visually different on plots like those in Figures 6.3 to 6.6.

A good way to visualise the best fit results is with the best fit map (Figure 7.2), which is coloured for the model that obtained the highest log-likelihood score for that cell of data points. The colouring scheme is white for MG, green for ML, red for MK and blue for MNIG, and their percentage best score is shown in the title. Notice that the water area is mostly fitted as Gaussian as would be expected for reasonably homogeneous regions (see Section 2.4). The ML model seems to be highly represented along the shoreline, which can be explained by the area mixture of the narrow water distribution and the generally broad land distribution (as mentioned in Section 5.4). On land, the MNIG model is the dominant best fitted model. The overall proportions were 37% MG, 3% ML, 11% MK and 49% MNIG, however these values are probably quite dependent on the actual region and proportion of water versus land. For C-band data, the proportion of MK increased about 8% at the expense of MNIG in the land areas.

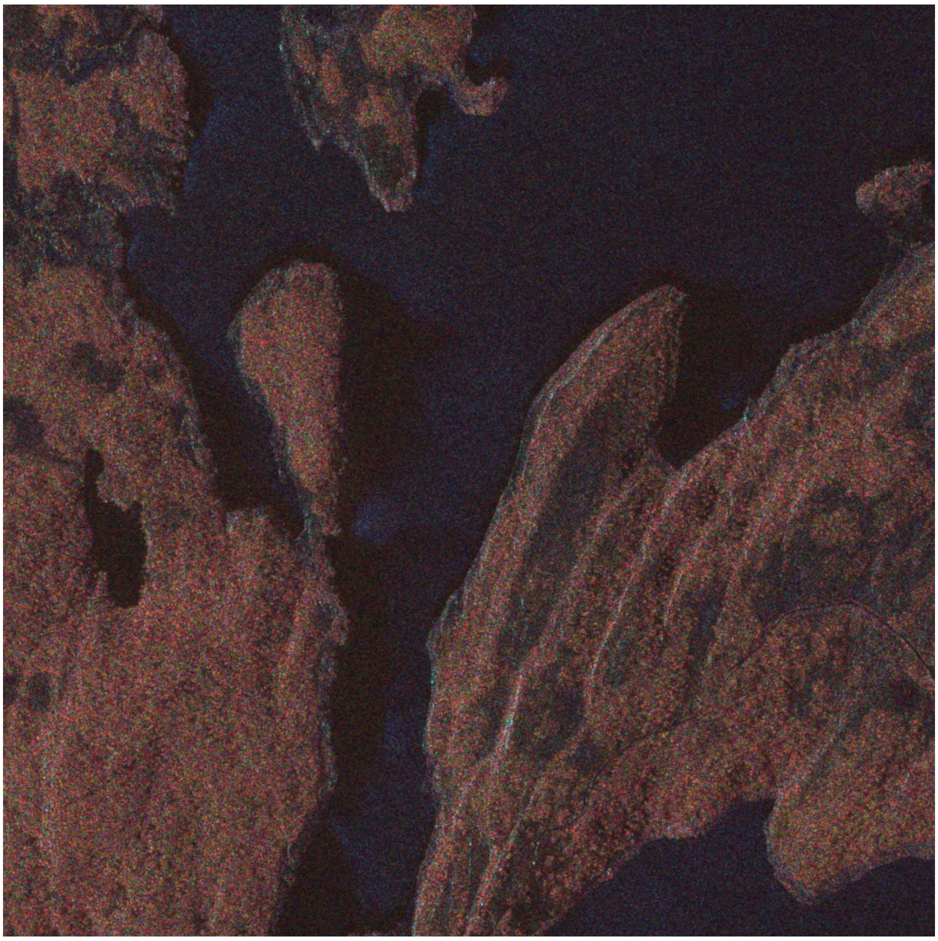


Figure 7.1: L-band RGB intensity image (R=HV, G=HH, B=VV) of test area, consisting of water and land areas.

The close-fit concept is well demonstrated with coverage maps for each model (Figures 7.3 and 7.4), depicting where each model was the best fit in red, within 0.5% log-likelihood score in magenta, and black otherwise where it is considered a poor fit. Given this very simple criterion for good-fit/poor-fit, gives a percent coverage value for each model. That is the MG model has a 74% good coverage, the ML only 8%, the MK model 91% and the MNIG model 94% good fit coverage. The coverage values are roughly the same for both C-band and L-band data. The reverse values, the poor-fit proportions, imply that the MG model *does not* fit real PolSAR data in 26% of this area, much more if just considering land areas, whereas the MK and MNIG models fits poorly in only 9% and 6% of the image, respectively. The main places that these two models fail are the strong mixture boundaries along the shoreline, where the statistical sample does not represent a single target medium. This is a problem that affects all neighbourhood or smoothing style techniques.

Of course, this is a very crude measure but it does show the usefulness of one flexible model. These plots indicate that the most flexible model that can describe nearly all of the real PolSAR data is the MNIG model, with the MK a close second. The MNIG also has the greatest area of best fitted scores.

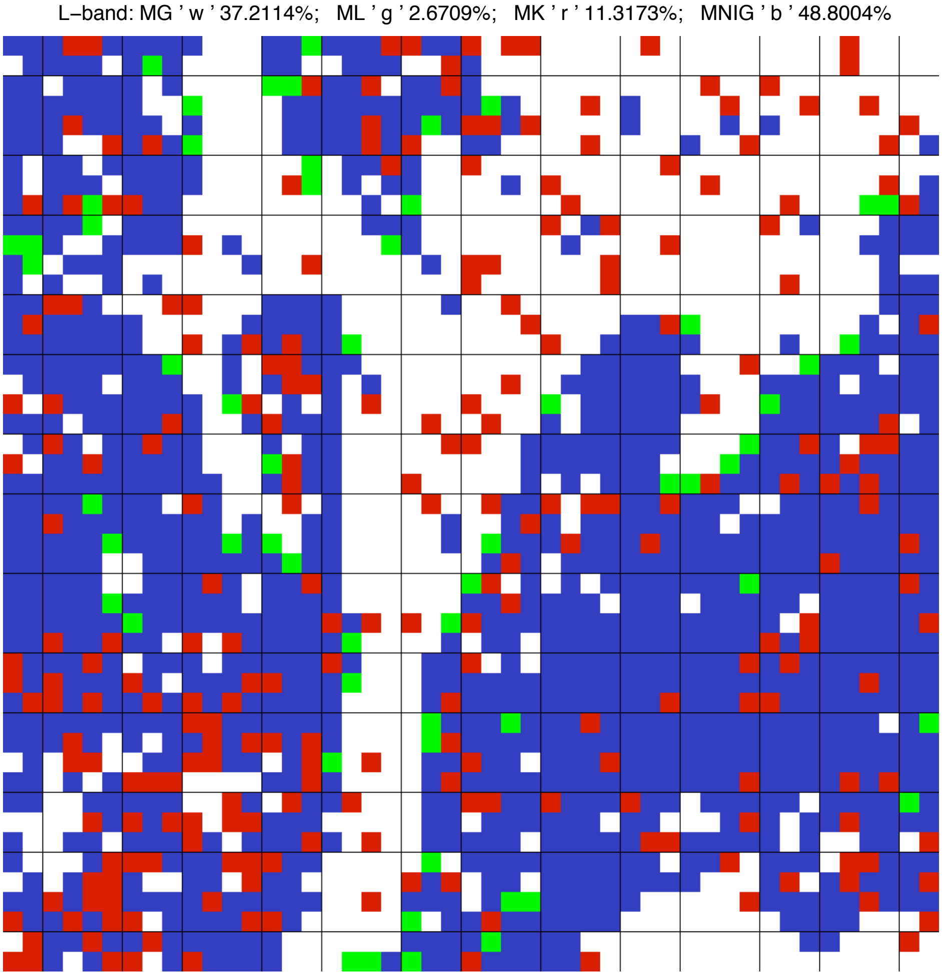
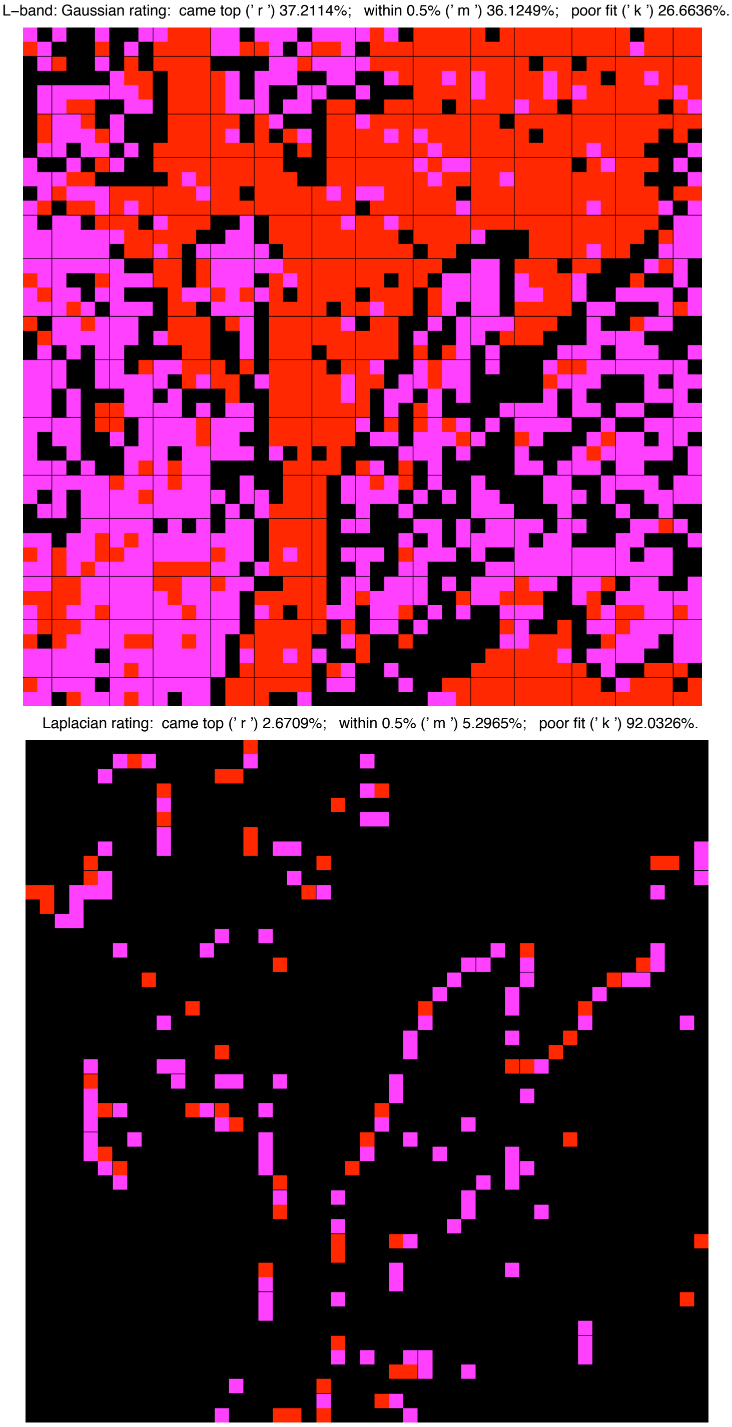


Figure 7.2: Best fit map. L-band, 8-D, 21 × 21 cells, moment method fitting, log-likelihood goodness-of-fit testing. Colour scheme: MG in white, ML in green, MK in red, and MNIG in blue.



L−band: Gaussian rating: came top (’ r ’) 37.2114%; within 0.5% (’ m ’) 36.1249%; poor fit (’ k ’) 26.6636%.

Laplacian rating: came top (’ r ’) 2.6709%; within 0.5% (’ m ’) 5.2965%; poor fit (’ k ’) 92.0326%.

Figure 7.3: Coverage maps for MG (top) and ML models. Best fitted in red, within 0.5% in magenta, and poor fit in black.

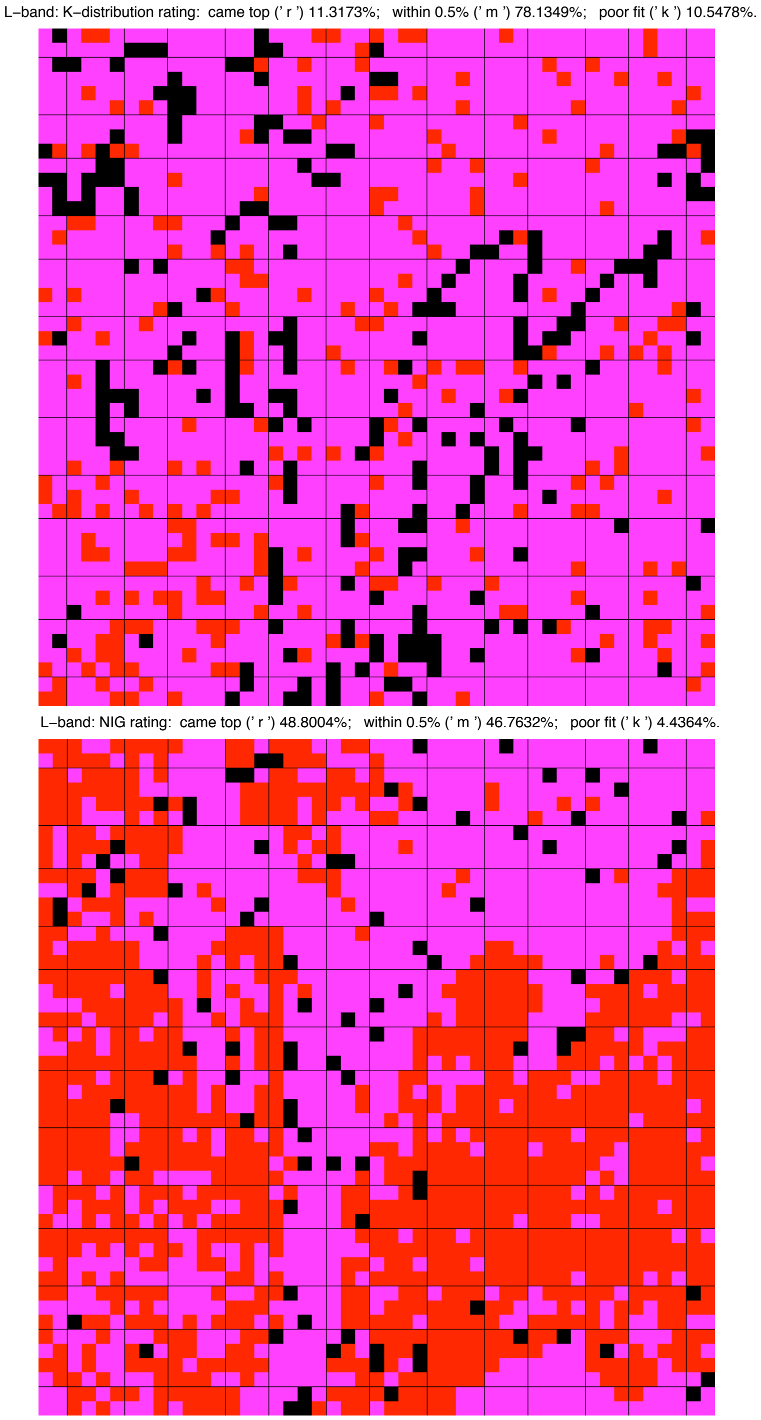


Figure 7.4: Coverage maps for MK (top) and MNIG models. Best fitted in red, within 0.5% in magenta, and poor fit in black.

Chapter 8

Summary I

PartI looked at several quite different tasks and presented certain choices and conclusions that simplify the further goal of segmentation or classification.

Firstly, the four models were introduced with practical parametric estimation methods. Properties of the scale mixture of Gaussians class were discussed with particular emphasis on the global shape and scale parameter interpretation and the polarimetric internal covariance structure matrix. The final choice of method of moments estimation technique proved fast enough for image analysis work and the accuracy was argued to be sufficient to capture the non-Gaussian shapes of the distributions. Scope exists for applying constraints and their benefit was demonstrated with simulations.

Secondly, the concept of Goodness-of-fit testing was investigated with the resulting choice of the log-likelihood measure based upon speed and accuracy for multidimensional sparse distributions.

Lastly, applying the four models and the goodness-of-fit measure to real PolSAR data demonstrated that radar backscatter signals are often not Gaussian in distribution, and furthermore indicated that both the multivariate K and multivariate normal inverse Gaussian distributions are flexible models that are quite able to capture the range of distributions observed.

The flexibility of just one model means that all the time consuming computations of four different models, and goodness-of-fit testing are not actually required. Plus, the resulting single set of parameters make for a far simpler classification scheme that does not require the potentially complicated multiply branched classifier that would otherwise be required.

Although the MNIG model was indicated as the most accurate model in this study, the prevalence of the K distribution in the literature, its close second rating in this study and numerical problems with testing the MNIG model , actually generating test data, all contributed to choosing the MK model for the following segmentation investigations. As will be seen in the Part II, however, this choice becomes irrelevant for this task.

Further work would be required to more accurately characterise the suitability of the MNIG model for PolSAR data, and any potential theoretical implications that these initial results suggest.

Part II

Image Segmentation

Introduction

The investigations of using different non-Gaussian models to represent the PolSAR data set resulted in the conclusion that the fast moment methods for just the one multivariate K distribution model would suffice. Such processing produces an image feature set consisting of two scalar parameters, α and λ, plus a mean vector, $\boldsymbol{\mu}$, and a covariance structure matrix, $\mathbf{\Gamma}$.

This section investigates the feature set produced, interprets some of its values, transforms them into a more linearly spread feature space, and then uses selected features for simple clustering and unsupervised image segmentation. The only validation available for this data set is a comparison to a much coarser land cover map and the natural looking appearance of the classified regions formed.

Time has not permitted further comparisons, however it would be of particular interest to test this feature set for supervised classification (would also require better ground truth data), or to compare classification with alternative feature sets, such as entropy-alpha decomposition [30].

Chapter 9

Ground Truth Data

Comparison shall be made to a vegetation cover map provided by Norut IT of Tromsø, Norway, which was based upon Landsat 5 TM imaging (optical and infrared) from August 1988. The data had been classified into 36 land cover classes, and the resolution was 30 × 30 metre. This image was resampled by Norut IT to 20 × 20 m to match the planned neighbourhood size of the modelling data. However, because the land cover image was geocoded, it did not match the PolSAR data co-ordinates and one of them had to be transformed. The geocoded land cover map was transformed to the PolSAR XY co-ordinate system to avoid any possible sampling effects on the statistical data. This was achieved with a local weighted mean transform, performed at the PolSAR resolution of 1.5×1.5 m, with a nearest neighbour class selection.

The test area's class image (Figure 9.1) contains only 21 classes, with only one class for water. It looks quite coarse when compared to the PolSAR data, even after the smoothing performed by the statistical modelling. However, this is the only ground data available, and will have to suffice. The class definitions list did not become available until a very late stage and so meaningful simplification or regrouping into a texture based, i.e., vegetation height and density, map could not be investigated. The only segmentation verification possible is that of visual judgement when comparing the two class maps.

Keep in mind that the land cover map is based upon imaging from 7 years earlier than the PolSAR data imaging, and is from much later in the season. Also, bear in mind that optical, infrared and radar systems sense quite different surface characteristics. Optical imaging relates to the familiar concept of colour, infrared to blackbody temperature, whereas radar imaging is more related to texture. Therefore, the segmentation of both types of imaging may not be expected to compare directly, although it is anticipated that many classes will be distinguishable on both systems. For instance, grassland and forest are distinguishable by colour in optical imaging, and presumably also by texture in radar imaging.

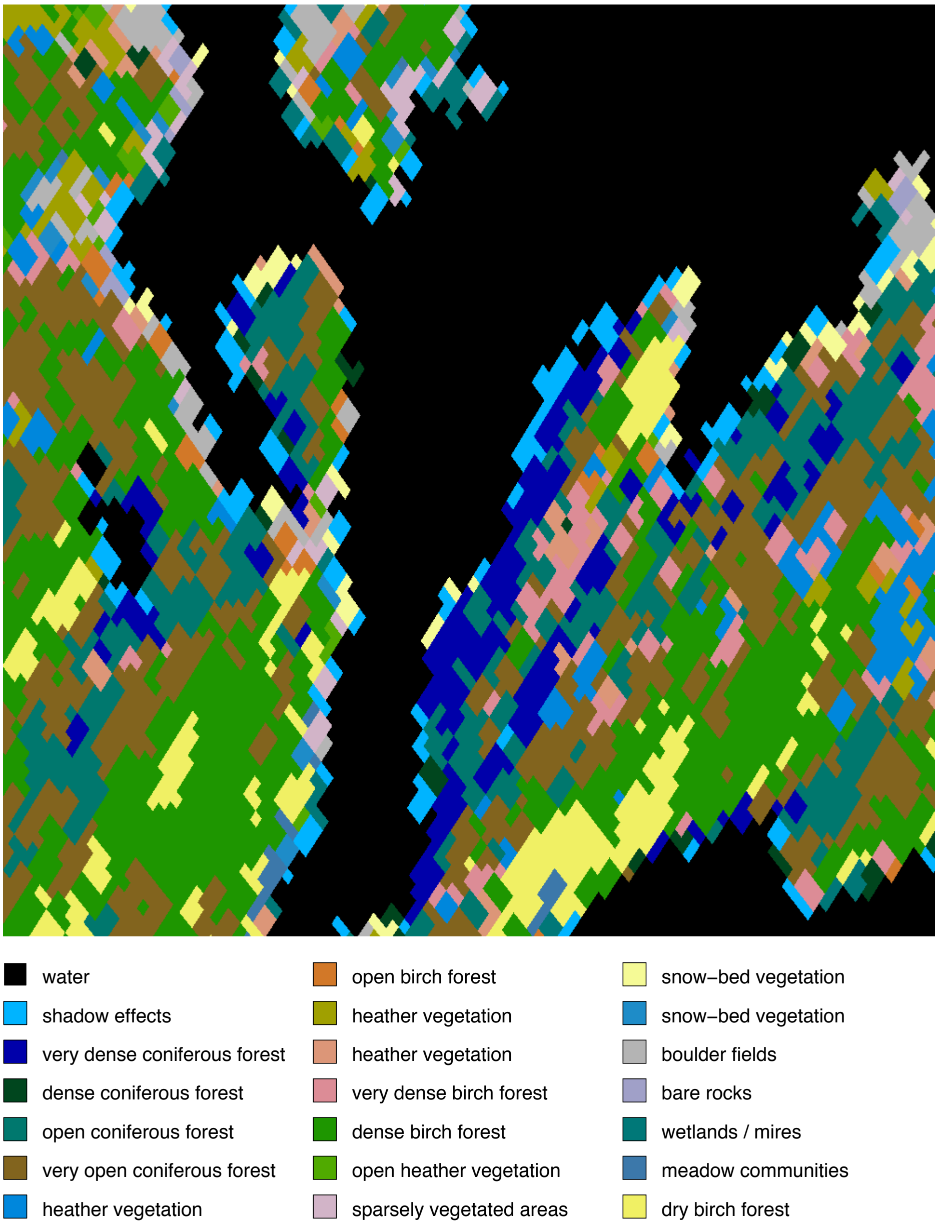


Figure 9.1: Land cover classification image coinciding with test PolSAR data area. Based upon data supplied by Norut IT, Tromsø, Norway.

Chapter 10

Classification Methods

Two unsupervised methods are looked at in this study and shall be referred to as the k-means method and the Bayesian mixture of Gaussians (BMoG) method. Both of these methods are essentially designed to find simple compact clusters. The classification was accomplished in two stages, clustering and classification, because of the large (1 million pixel) image size involved. The clustering stage was performed on a smaller random sample of the full data set, and then the whole image was classified based on the clusters obtained. Five thousand points per cluster were randomly sampled as the training set, representing up to 12 percent of the data, depending on the number of clusters being sought. The image is varied enough that this sampling appears to be representative of all clusters.

The approach taken in this study is to manually transform the feature set into a roughly linearly spaced feature space, followed by a simple clustering method. There exist many more advanced clustering algorithms, many including built-in kernel transformations such as spectral clustering [31], that may produce better results. However, time has not permitted them to be investigated here. The simple methods on pre-transformed data, are at least sufficient to show the general potential of the feature set.

The two methods discussed here can produce quite different class boundaries as is shown in the 2-D example for each. Additionally, both methods are able to produce disjoint classes and cut through clusters, depending on the situation. It was observed that on the normalised data using all 5 dimensions, that both methods produced roughly similar results.

10.1 k-means / iso-data algorithm

A description of the commonly used k-means or iso-data clustering algorithm, as well as the nearest neighbour classification can be found in [23]. The principle is to find the set of M centres that minimises the total squared Euclidean distance measure over all the data points. The classification then labels the data to the index of the nearest centre vector. Different distance measures may be used, but only the squared Euclidean distance is guaranteed to converge to the optimum.

It it recommended that the feature dimensions are normalised first, so that each dimension gets equal weighting, otherwise the segmentation may be preferential in the feature with the largest variance only. The usual normalisation is to centre the data to its mean vector and divide each dimension by the variance of that dimension. Thus each dimension becomes a mean zero and variance 1 data set.

The k-means method has a tendency to produce M roughly equally scaled segments and often appears to simply cut-up the data into blocks, even right through visible globular clusters. This was particularly noticeable before the dimension normalisation was applied, and prompted the search for another clustering method that did not cut-up globular looking clusters. If M is too large, the k-means method seems to simply cut clusters in half to make more clusters. A 2-D example is shown in Figure 10.1 and shows both of the two cases mentioned above. At the top a 3 class clustering is increased to a 4 class clustering with the red cluster seemingly cut down the middle. The lower 8 class clustering indicates the apparent straight-cut segmenting that seemed characteristic of the k-means method.

10.2 Bayesian Mixture of Gaussians clustering

The BMoG method was considered after viewing transformed feature space plots depicting roughly globular clusters, and not wanting to cut them up like the k-means method seemed to do. The method assumes that the clusters posses a Gaussian density profile, but may be (multi-dimensional) elliptical due to the included covariance matrix term. It is perhaps related to using the k-means method with the Mahalonobis distance measure, but with a different iteration measure. The clustering will find M centres with covariance and then the entire image may be classed with a Bayesian maximum likelihood classifier to set each data point to the most likely Gaussian cluster centre (general concepts covered in [23]).

The BMoG method also has a strange side effect, where it may produce disjoint clusters or even annular ring clusters. If a broad density cluster has another high density cluster within it, then points near the high density centre becomes one class and the remaining data around it becomes the other class. Although correctly optimised for discrete mixtures of Gaussians, these disjoint and ring clusters may not be physically meaningful given some interpretations of the feature space parameters. If M is too large, the BMoG method seems to duplicate a cluster into two half density roughly equal centres, with quite a range of resulting class boundary patterns. At least the high density centres will be appropriately grouped and only the smaller number of remaining points may find themselves in these oddly shaped clusters. Figure 10.2 shows a similar 2-D example to that used with the k-means clustering with, arguably, better class boundary judgement. The step from 3 to 4 classes seems to (correctly) have found the side lump on the red cluster, rather than just splitting it down the middle. The disjoint example is visible in the 8 class example (bottom) where the dense red class protrudes through the broader blue and green classes, leaving a small number of blue points above the red group.

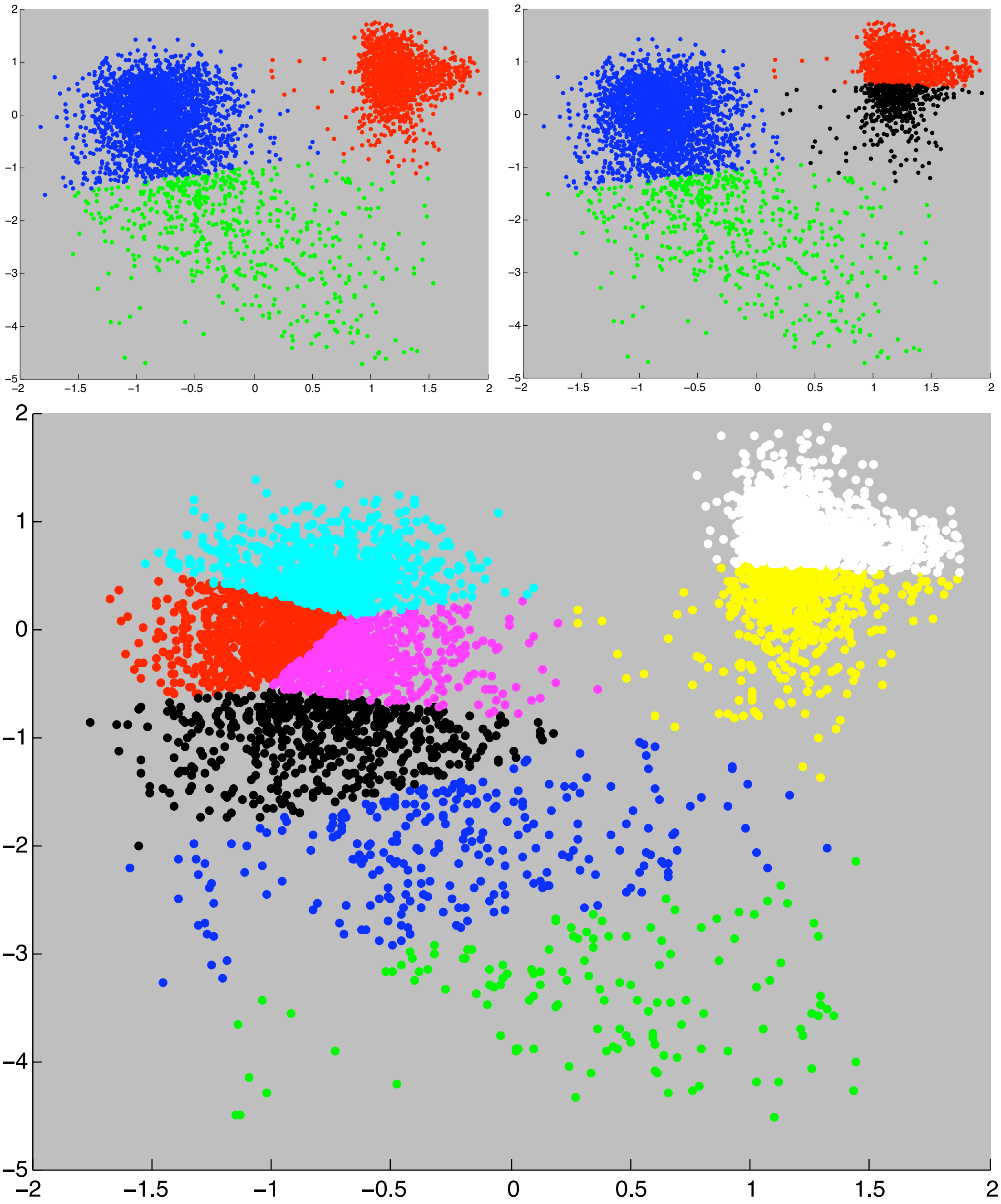


Figure 10.1: A simple 2-D k-means clustering example. (top) An apparent cluster (red) is split in two (red/black) to make another cluster, and (bottom) the cutting-up of the data space.

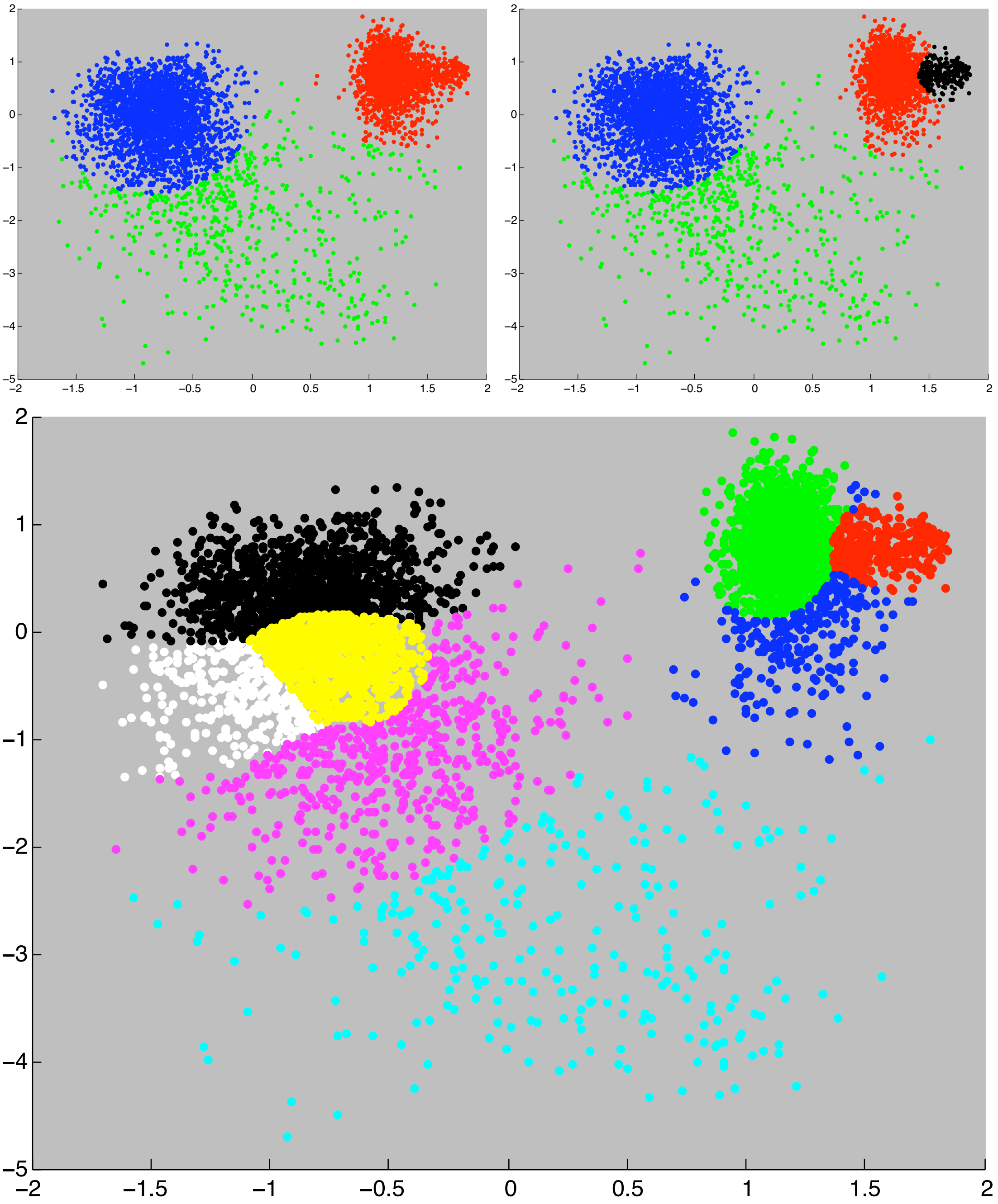


Figure 10.2: 2-D BMoG clustering example. (top) An apparent cluster (red) is split at side bulge to make another cluster (black), and (bottom) the BMoG curved segmenting of the data space, with the disjoint cluster (blue) example.

Chapter 11

MK Parametric Feature Space

The MK parameter set contains two scalars, a vector and a matrix parameter. A single feature vector can be obtained by selecting both scalar parameters plus specific terms from the vector and matrix parameters. Theoretical considerations (Section 2.4) and empirical evidence (Section 4.6) suggest that the mean vector is not particularly important, being mostly very near zero anyway, and the theoretically most relevant terms from the covariance structure matrix are the diagonal terms and the VV by HH covariance term. That is, the four main averaged terms, d_1, d_2, d_3 and c_1, from Section 4.6.2. The four $\mathbf{\Gamma}$ terms may be reduced to three, by dividing by one term, since the covariance structure matrix is already normalised.

This first feature set shall be denoted as the Alpha-Lambda-Gamma (ALG) set, consisting of the MK model's α, λ and 3 $\mathbf{\Gamma}$ matrix terms, that is

$$\text{ALG:} \quad \begin{bmatrix} \alpha \\ \lambda \\ G_1 \\ G_2 \\ G_3 \end{bmatrix} \quad \text{, where} \quad \begin{array}{l} G_0 = (\Gamma_{1,1} + \Gamma_{2,2})/2 \\ G_1 = (\Gamma_{3,3} + \Gamma_{4,4} + \Gamma_{5,5} + \Gamma_{6,6})/(4G_0) \\ G_2 = (\Gamma_{7,7} + \Gamma_{8,8})/(2G_0) \\ G_3 = (\Gamma_{1,7} + \Gamma_{2,8})/(2G_0) \end{array} \tag{11.1}$$

The initial clustering attempts were performed directly on the ALG data as given in (11.1), using the k-means method with 8 classes. The results showed one class for all the land data and 7 different water classes and was not particularly useful. The simple reason is that the range of values of each dimension were so different and the high value lambda terms, order 10^6, were dividing the classes on that scale, with all the land detail being lost within one class near zero. This prompted the need to view and then transform the data in some way prior to clustering.

Figure 11.1 (top) shows the raw α and λ features plotted against each other and clearly demonstrates the large range difference involved, the seemingly angular relation between the two, and the concentration of points around the origin. The first transformation (middle, Figure 11.1) derives from the theoretical moment estimates from the MK model (Section 4.1.1) and is to plot α vs. $\lambda/(\alpha+1)$, yet still shows nonlinear bunching up near zero α. The last transformation (bottom, Figure 11.1) shows several roughly symmetric globular clusters and seems to be a suitable input for the simple clustering algorithms. Note that only a few thousand random points are plotted and thus only the main clusters densities are visible.

The last transformation equations are plotting $\log(\alpha+1)$ vs. $\log(\frac{\lambda}{\alpha+1})$ and are essentially the log of the distribution's shape and width parameters. These are derived from the empirical distribution measures, $\bar{Z}_1$ and $\bar{Z}_2$ from Section 4.4, and have almost completely undone the complex transformations that obtained α and λ from $\bar{Z}_1$ and $\bar{Z}_2$ in the first place. Substituting $\bar{Z}_1$ and $\bar{Z}_2$ into these expressions gives a relation of $\log(1/(\frac{\bar{Z}_2}{\bar{Z}_1^2}-1))$ vs. $\log(\frac{1}{\bar{Z}_1})$.

Starting directly from the empirical measures, $\bar{Z}_1$ and $\bar{Z}_2$, and using an inverse instead of a log transformation for the first ratio term yields a slightly better cluster plot. Figure 11.2 shows the raw $\bar{Z}_1, \bar{Z}_2$ space (top) with the α, λ derived transform (middle) and the alternative $\bar{Z}_1, \bar{Z}_2$ transform (bottom).

The final transform's two terms are $-\log(\bar{Z}_1)$ and $(\bar{Z}_1^2)/\bar{Z}_2$. The first term is basically $-\log(\text{width})$ and the latter essentially $(\text{width}^2)/\text{kurtosis}$ which is a commonly used measure of non-Gaussianity. The expressions are purely in terms of the sample distribution measures $\bar{Z}_1$ and $\bar{Z}_2$ and pay no regard to parametric model parameters. This is an important interpretation and worth re-stating. Transforming the parametric scalar terms to improve cluster distinguishing power, has effectively bypassed any particular parametric model, and reverted to purely empirical distribution measures of *width* and *non-Gaussianity.*

This of course means that no matter which model was chosen, the MK or MNIG or some other underived model, the distinguishing ability is already contained purely in the raw moment measures of the distribution. The other $\mathbf{\Gamma}$ matrix terms are also derived purely from the normalised sample covariance and are therefore independent of any particular *scale mixture* model. This discussion holds for scale mixture of Gaussians class of models, and may not be the case for other classes of parametric models. The main required assumption from this model class is that the dimensions are all equivalently distributed apart from internal covariance structure, leading to the global shape and width terms.

Essentially this non-Gaussian modelling leads to one extra feature parameter than a purely Gaussian approach would produce. That is, the non-Gaussianity measure, in addition to the purely Gaussian estimates of width and covariance.

The $\mathbf{\Gamma}$ terms, G_1 to G_3, are also inter-related, and some simple transformations produce a better spread cluster space. The choice of transforms is perhaps intuitive guesswork and, although not perfect, seem at least suitable as input for the simple clustering algorithms. The choice was simply to log the two diagonal terms, and leave the cross term alone since it contained some negative values. Figure 11.3

shows the three raw $\mathbf{\Gamma}$ term combinations down the left and the three transformed combinations down the right.

The new feature space, which shall be called the Zed-Gamma (ZG) feature set, can be described by

$$\text{ZG:} \quad \begin{bmatrix} -\log(\bar{Z}_1) \\ (\bar{Z}_1^2/\bar{Z}_2) \\ \log(G_1) \\ \log(G_2) \\ G_3 \end{bmatrix}, \tag{11.2}$$

where G1 to G3 are as in (11.1), and the $\bar{Z}_1, \bar{Z}_2$ terms refer to the first and second moment estimates of Z, which are in fact the mean sample variance and Mardia's sample kurtosis (see Section 4.4).

Note that a purely multivariate Gaussian analysis of the same data would have produced the same feature set minus the non-Gaussianity term, lets call it the Width-Gamma (WG) set. This is because of the main assumption that all dimensions are simply scale mixtures of Gaussians. The nonuniform scaling information goes completely into the non-Gaussianity term, the remaining terms are just the Gaussian estimates for width and covariance.

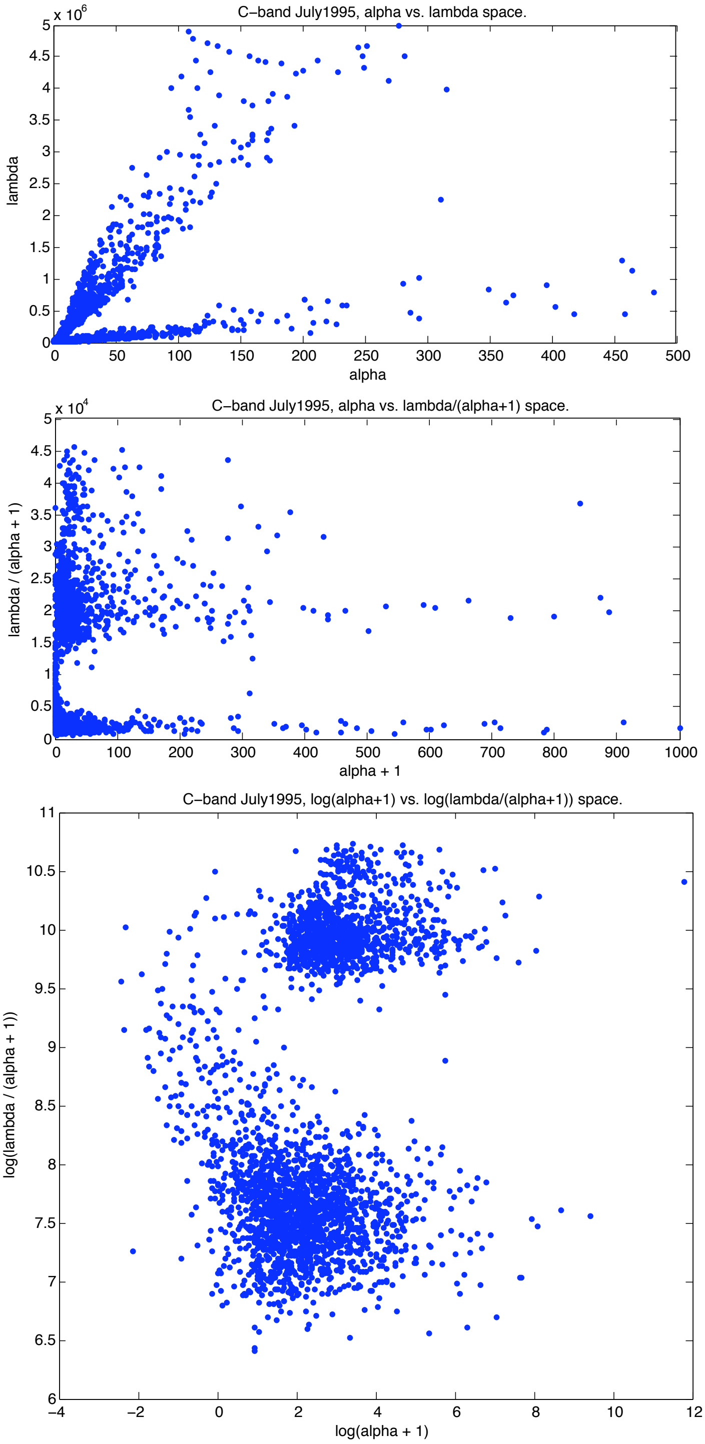


Figure 11.1: MK α and λ feature space transformation plots.

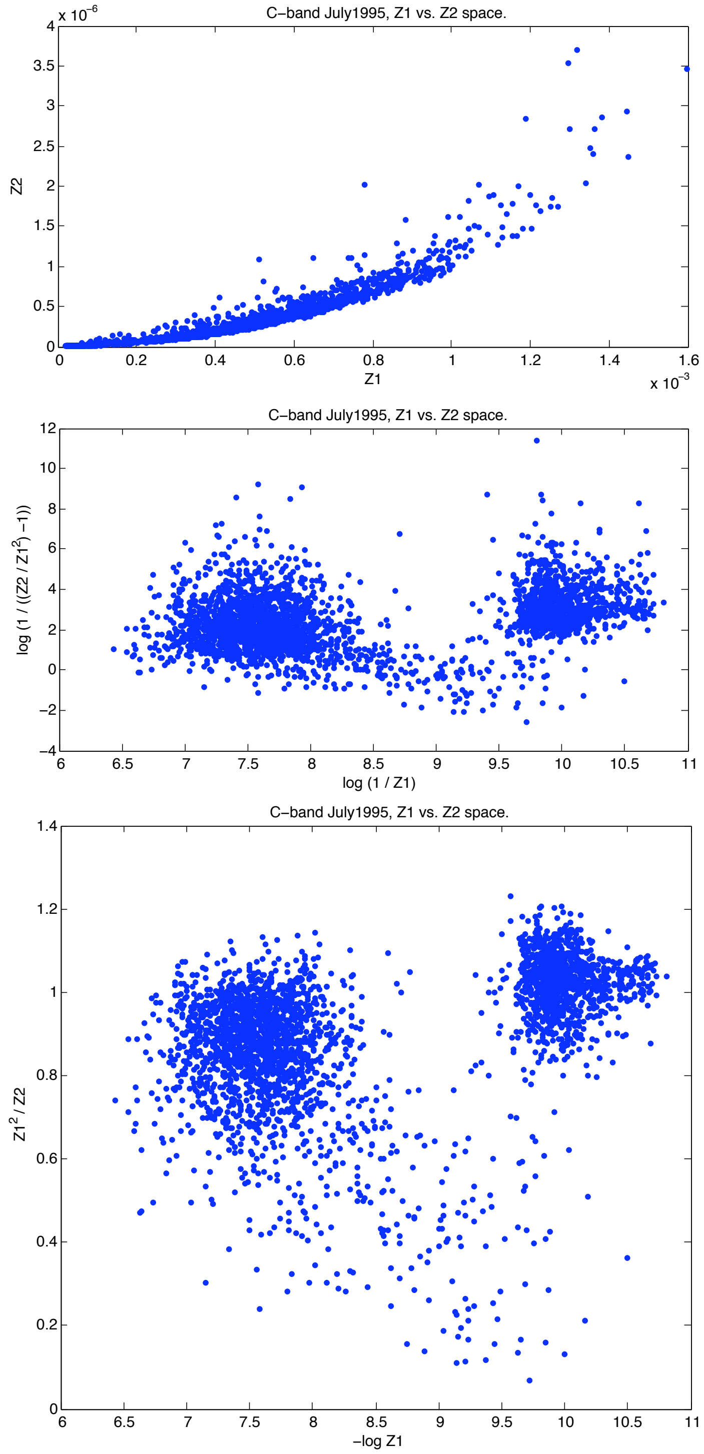
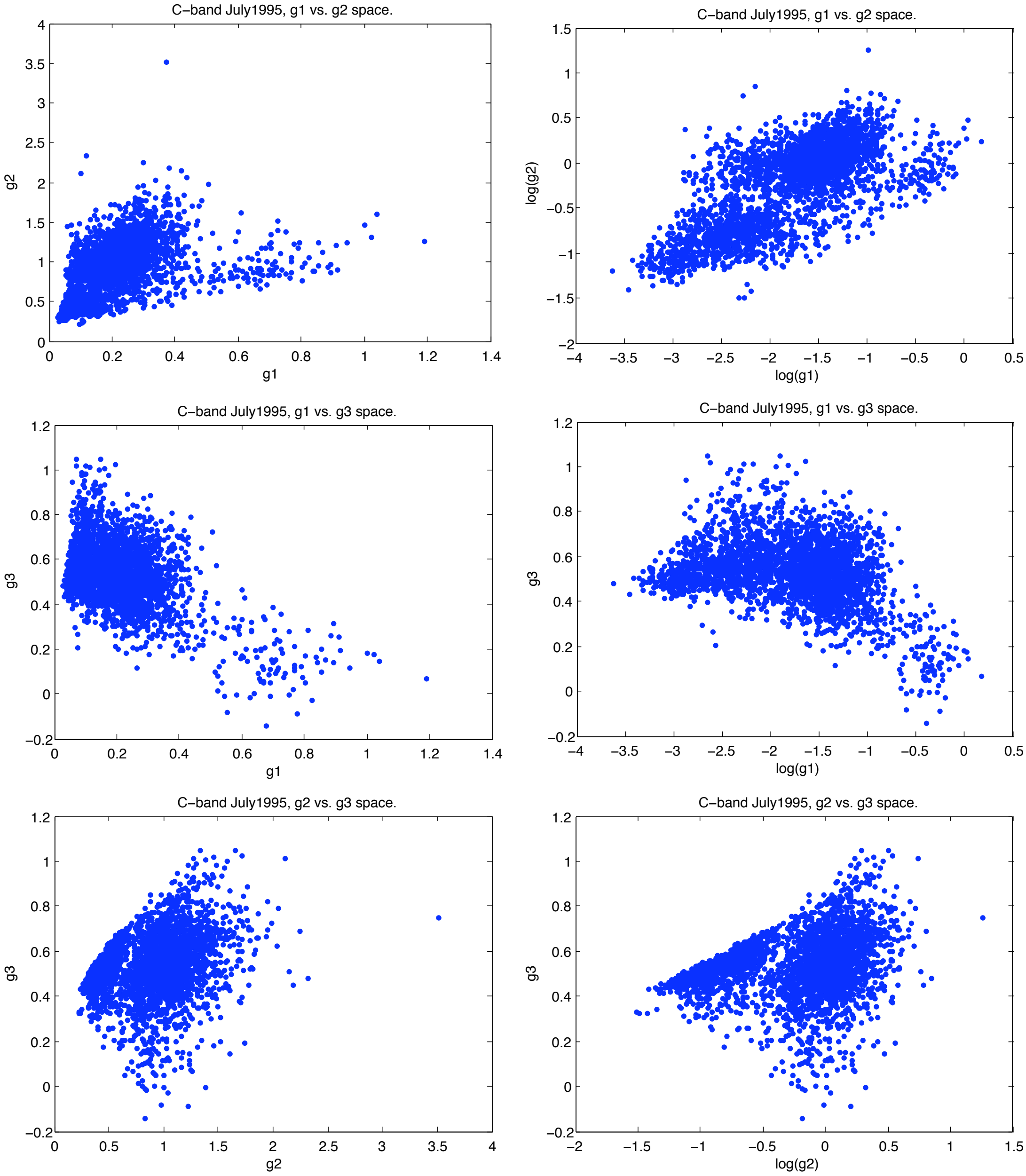


Figure 11.2: Moment estimates, $\bar{Z}_1$ and $\bar{Z}_2$, feature space transformation.



C−band July1995, g1 vs. g2 space.

C−band July1995, g1 vs. g2 space.

C−band July1995, g1 vs. g3 space.

C−band July1995, g1 vs. g3 space.

C−band July1995, g2 vs. g3 space.

C−band July1995, g2 vs. g3 space.

Figure 11.3: **Γ** term feature space transformations.

Chapter 12

Segmentation Results

Many different options have been discussed, in the modelling procedures and the different feature sets, that may be applied to create segmentation maps of PolSAR data. The modelling can be performed at different neighbourhood sizes, the constraints may or may not be applied, and there exists data from both L-band and C-band frequencies. Additionally, it would be worth investigating how the segmentation compares to a more traditional triple intensity polarimetric based segmentation, and also to a single intensity nonpolarimetric (co-polar HH) segmentation. Comparison may also be made to the multivariate Gaussian based analysis, since the WG data set is simply a subset of the ZG data set.

Unfortunately, it is not possible to obtain a quantitative measure of the effectiveness of any option because of inadequate ground truth data. Each option will simply be imaged for visual comparison and informal observations discussed. If not specifically stated the comparisons refer to the July C-band data set. Note that the colour scheme is manually chosen to attempt a similar colour appearance as the land cover map, yet the actual colour chosen does not necessarily represent the defined land cover class.

12.1 Clustering methods

Figure 12.1 depicts images made with both the k-means and BMoG methods for varying numbers of clusters. It is clearly seen that land/water is equally well separated in both methods, but the land based class regions differ. For larger numbers of clusters, the differences seem smaller with the BMoG method possibly creating slightly more solid regions, for example, compare the yellow and pink regions.

It is not clear whether the overall patchy appearance is a real phenomenon, perhaps relating to real vegetation density changes, or an artefact of the analysis. The land cover map has too coarse a resolution to answer this question.

The 25 class maps have a similar number of land classes as the land cover map, however, the 16 class maps have sufficient detail for visual comparison of the main features and, to reduce complexity, further comparisons will be made with only 16 classes.

12.2 Neighbourhood sizes

The size of the neighbourhood averaging, i.e., the statistical sample size, was increased to mimic the coarse resolution of the land cover map, and hopefully improve class separation due to decreased parameter variance. The previous images at a 13×13 (20 m square) neighbourhood are replicated, together with maps at the increased 21×21 (32 m square) neighbourhood in Figure 12.2. The land cover map is included for visual reference. The true fluctuation scale is not known because the radar data if of higher resolution than the land cover map. However, it is easier to compare the smoother 21×21 map to the land cover map and this size will be used for the other comparisons.

12.3 Constraints applied

Figure 12.3 compares the clustering obtained with the zero-mean and $\mathbf{\Gamma}$ matrix constraints applied during analysis, to the previous unconstrained analyses. The colour differences make it difficult to see, but distinct differences exist. The constrained k-means image seems to have more detail as seen by the extra light brown patches on the left hand land area and the pink/grey-blue separation on the right. The constrained BMoG image does not show as much variation although a few areas are more solid. It may be possible that the differences are caused by differences in the randomly sampled training data affecting the clustering, or it may reflect less variation in the $\mathbf{\Gamma}$ matrix estimates leading to tighter clustering. Once again, the land cover map is inadequate, and the regions too complex, to accurately test the usefulness of applying the constraints. It was also shown that the constraints had far more effect for smaller sample sizes, and may simply not show up at the 21×21 smoothing level.

12.4 Intensity based maps

The modelled feature set can be compared to a smoothed intensity based clustered image. That is, by taking the absolute value of the complex scattering coefficients, averaging over the local 21×21 neighbourhood, and combining the cross-polar terms, with $\sqrt{2}$ factor, to produce a triple intensity feature set. The dimensions are then normalised by their individual variance and the k-means and BMoG segmentation methods are applied. Early radar systems only used one co-polar polarisation and this can be simulated for comparison by analysing just the HH intensity term. The triple intensity segmentation maps are shown above the HH-intensity maps in Figure 12.4. The land cover map and the smoothed 3-intensity brightness image are included for reference.

A surprising amount of distinguishing information appears to be present, even in the single intensity image, although there is a tendency to produce smooth joined regions surrounded by boundary regions, particularly for the k-means method. The single HH map also has a marked reduction in the number of visible classes, as some clusters converged to contain only very small numbers of points or became (invisibly) thin shell-like boundaries.

Some aspects of the main features seem to be present in the land cover map, however, the long connected regions and layers like contours may actually derive from the topography of the region. It is well known that backscatter intensity is related to reflection angle and the hillside curvature directly reflecting back to the imaging antenna will appear the brightest. The included brightness image clearly shows the long thin bright areas that become one class and although the topographic map of the area has not been studied to confirm this, it certainly appears to represent natural topography. Of course, vegetation cover is also influenced by the local topography, so it is unknown how much of the effect may be real land cover and how much topography. Overall, the HH-intensity map does not have quite as rich variation as the modelled ZG data images, but the 3-intensity map has a similar level of detail. Unfortunately, it is unknown which is the better clustering.

12.5 Gaussian Analysis

The WG data set is just the ZG set minus the non-Gaussianity term. The WG image is compared to the ZG image in Figure 12.5. There are visible differences (apart from the colours), for instance the shoreline classes that the non-Gaussian term highlights, perhaps more ragged region boundaries in the WG map, and also some more pronounced shape changes in some areas. The pink regions in the ZG images, become far more extensive brown regions in the WG images.

The shoreline enhancement is not actually a desirable effect, because it will incorrectly create different classes at the land/water interface over the thickness of the neighbourhood smoothing. This problem of area mixing exists in many methods, including, no doubt, the WG set, but the non-Gaussian term highlights it here. The usefulness of the non-Gaussianity term in the land areas cannot be measured without far more rigourous testing with better ground truth data.

12.6 Polarimetric information

Since the non-Gaussian term has shoreline problems, and the intensity term has topology problems, it seems worthwhile to compare the purely $\mathbf{\Gamma}$ matrix segmentation. That is, the normalised covariance structure matrix that holds all the polarimetric information of the data. Being normalised to determinant 1, it is expected to be invariant to the mentioned problems. Segmentation maps created purely from the 3 $\mathbf{\Gamma}$ matrix terms are compared to the full ZG set in Figure 12.6. The G only image is clearly far more patchy with far fewer large connected regions. The (ZG to G) difference is far more striking than the (ZG to WG) difference of Figure 12.5. The $\mathbf{\Gamma}$ only segmentation does depict some of the major features of the land cover map, e.g. the green is predominantly on the left and lower right. Clearly the intensity term is crucial in joining otherwise only marginally differing areas, however, it has still not been validated which is the more correct segmentation.

12.7 C-band and L-band maps

The different radar frequencies have characteristic differences too. The longer wavelengths have better penetration power and so thick vegetation may appear quite differently in the C and L bands, as is clearly seen in the green/brown forest region and the pink areas of Figure 12.7. Presumably, the usefulness of the penetrative power depends upon the desired task, and no further comparison can be made without more defined requirements and ground data.

12.8 Combined C and L-band map

Better distinguishing power may be expected by using both C and L bands together, and this can be accomplished by simply putting both data sets together into one 10 dimensional feature set. The CL-ZG set segmentation maps are shown in Figure 12.8. As mentioned already, the ground truth data is inadequate for detailed validation. In this case the BMoG image is very smooth with simple regions, to the extent that it appears to have less detail than either the k-means clustering or each single band BMoG image, however, more rigourous testing is required to evaluate this image.

12.9 March and July maps

The available data included data for three different months in 1995: March, May and July. Initial investigations focused upon the July data because it was assumed to represent the summer land cover, probably without snow cover, whereas the

March and May images probably include a great deal of snow cover and frozen lakes. It is an interesting aside to compare the different seasons, and at least indicates that snow is distinguishable with this technique. Figure 12.9 shows both the C-band and L-band images for the March data and should be compared to equivalent July images of Figure 12.7. There appears to be quite a change in the water and shoreline regions, where (presumably) there are heavy deposits of wind-blown snow, and the narrow between the two main land areas is almost closed in. The (assumed) forested areas are still clearly distinguishable, and the heather area (pink/ligh-blue in the middle right of image) now appears similar to the assumed snow areas as might be expected. It is quite likely that snow has a particular polarisation signature that could be used to class the data, however, there has been no time to investigate this in the current work.

An additional view of the seasonal differences is shown in Figure 12.10, which shows the $\bar{Z}_1, \bar{Z}_2$ feature space for both July and March and for C-band and L-band. The main separation is probably that of land and water, yet sub-clusters exist within each. The main cluster on the right contains two reasonably obvious density centres, and it is interesting to note the differences of these from C to L bands and from July to March. These two are closer together in C-band than L-band, with the left-most of L-band almost merging with the other main cluster. The July to March images show that the C-band has greater change than the L-band, with reduced separability and more spreading between the clusters. There is good potential to use such plots to help characterise class signatures, but time did not allow this to be investigated in this study.

12.10 Segmentation figures

All the segmentation figures have been collected together in the following pages.

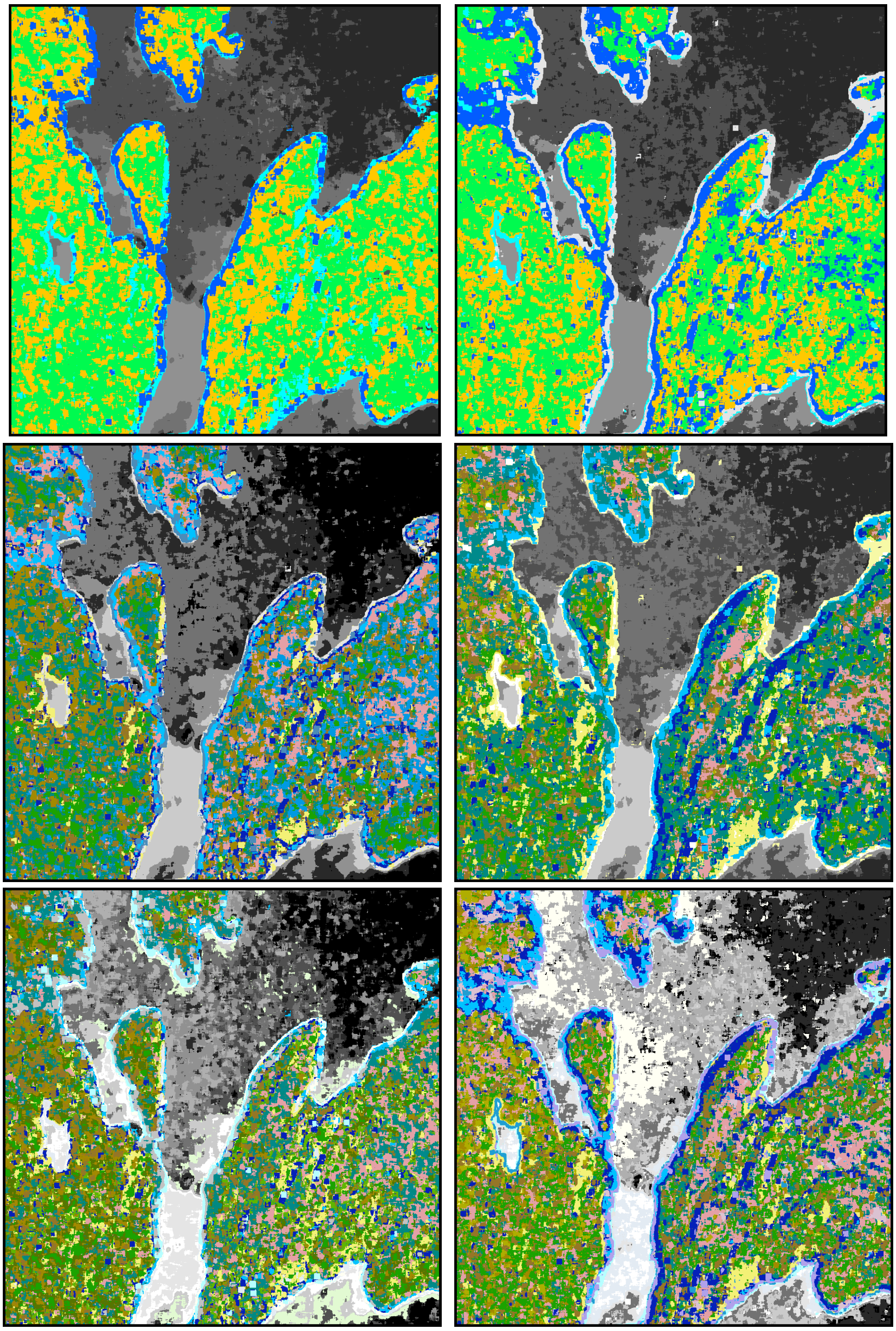


Figure 12.1: k-means (left) and BMoG (right) maps for each of 8, 16 and 25 classes, ZG data set, 13 × 13 neighbourhood.

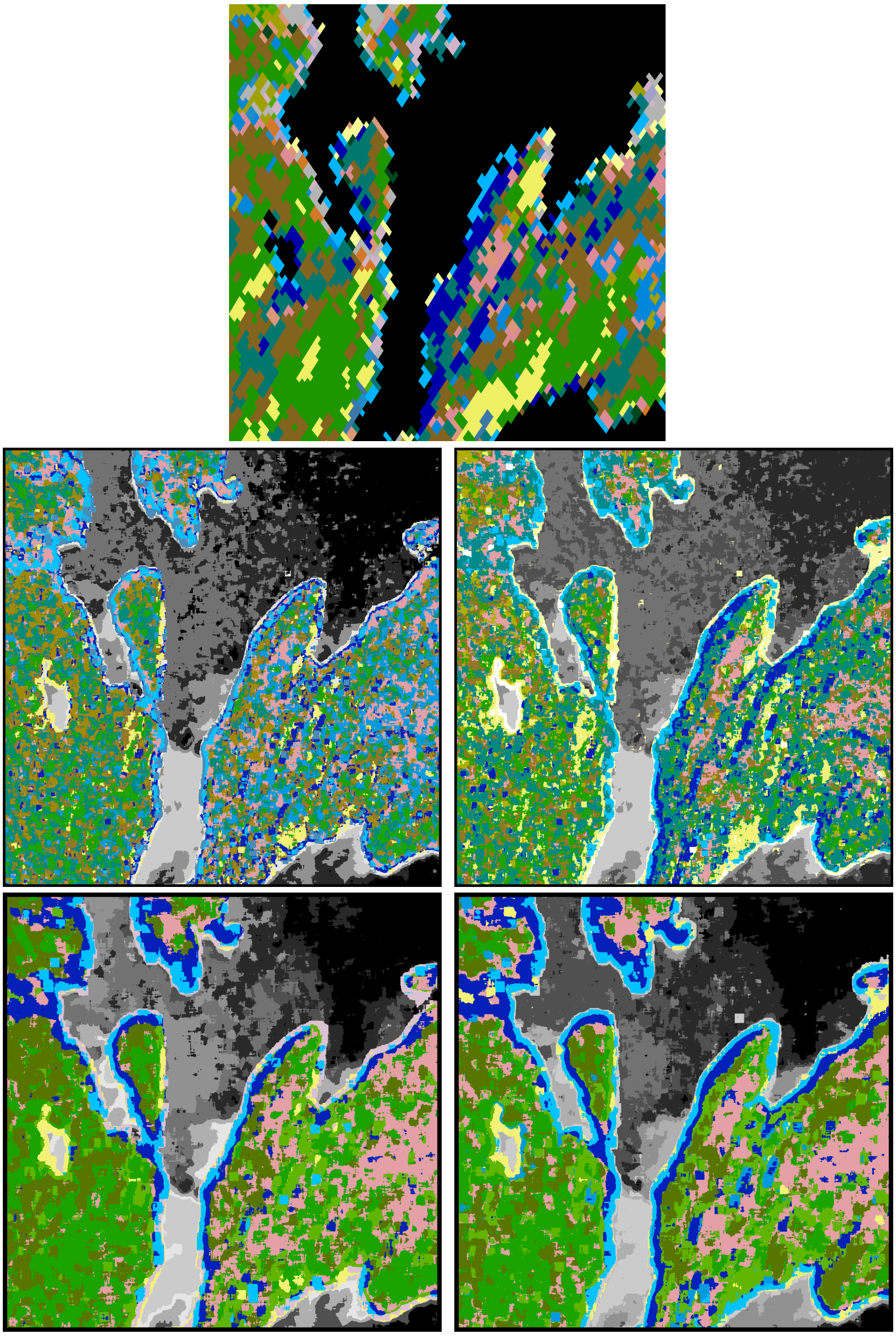


Figure 12.2: Land cover map (top), k-means (left) and BMoG (right) maps at 13 × 13 (middle) and 21 × 21 (bottom) neighbourhoods, ZG set.

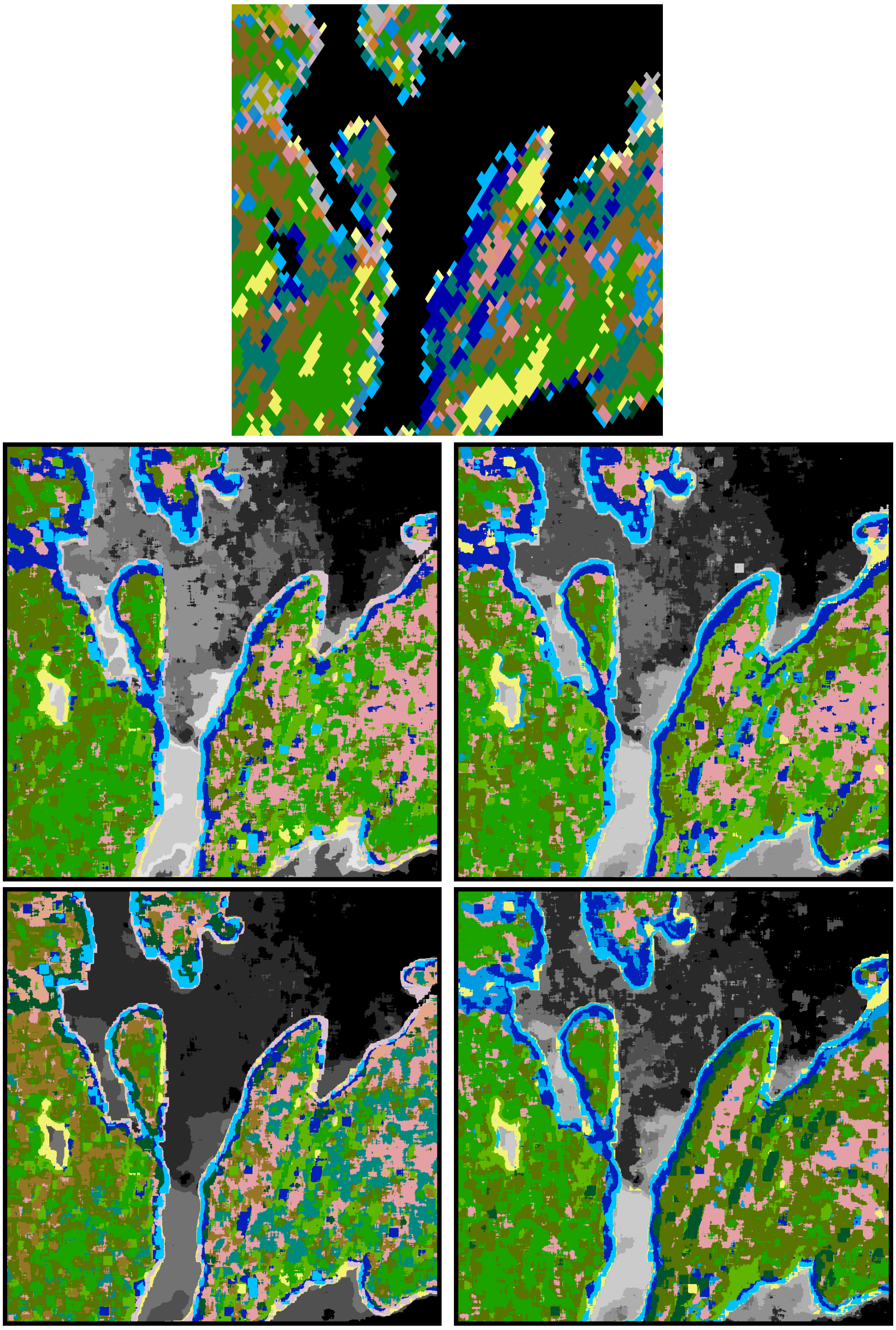


Figure 12.3: Land cover map (top), k-means (left) and BMoG (right) maps without constraints (middle) and with constraints applied (bottom), ZG set, 21 × 21.

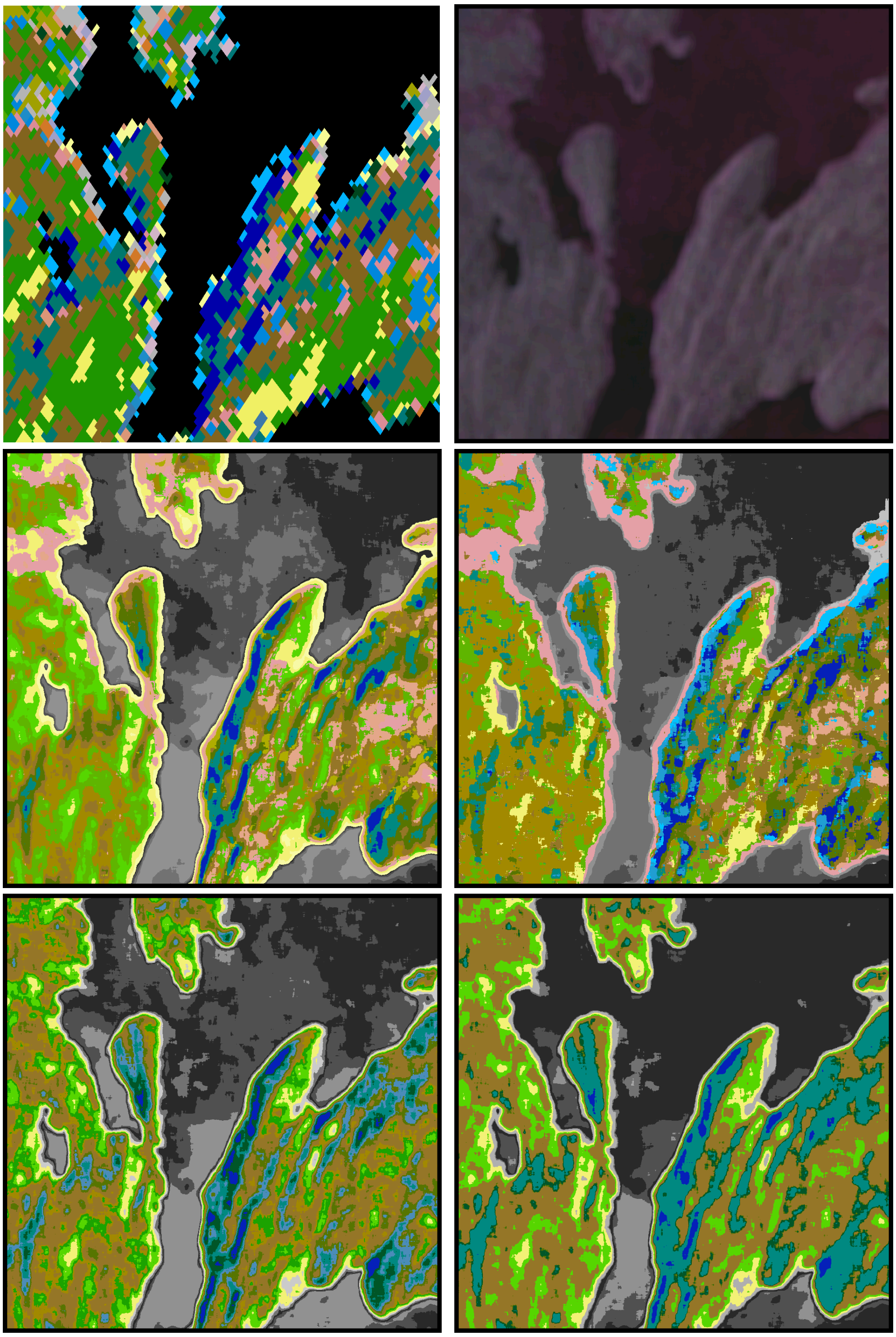


Figure 12.4: Land cover map (top-left) 3-intensity brightness image (top-right), k-means (left) and BMoG (right) 3-intensity maps (middle) and HH-intensity (bottom).

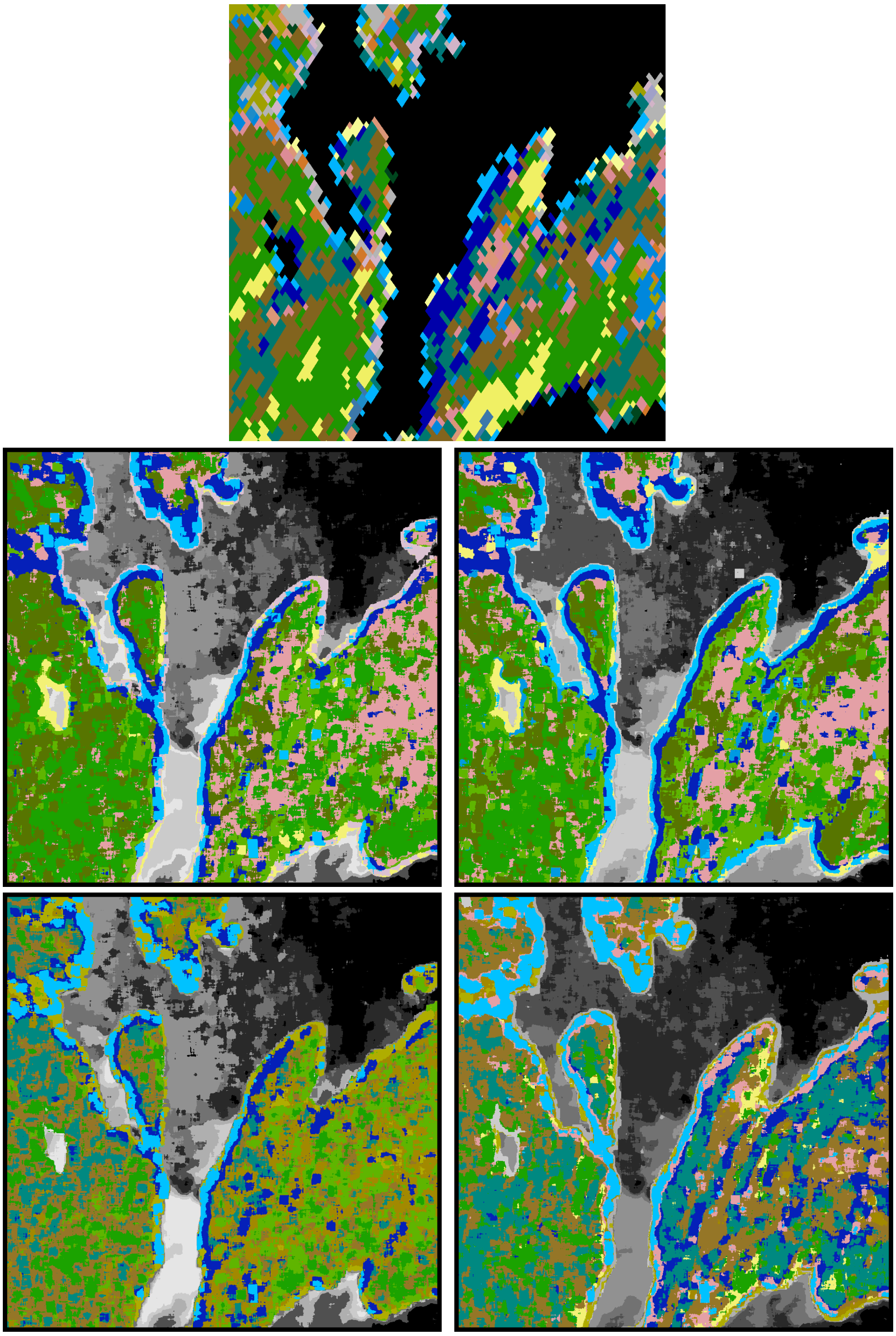


Figure 12.5: Land cover map (top), k-means (left) and BMoG (right) maps for ZG set (middle) and WG set (bottom).

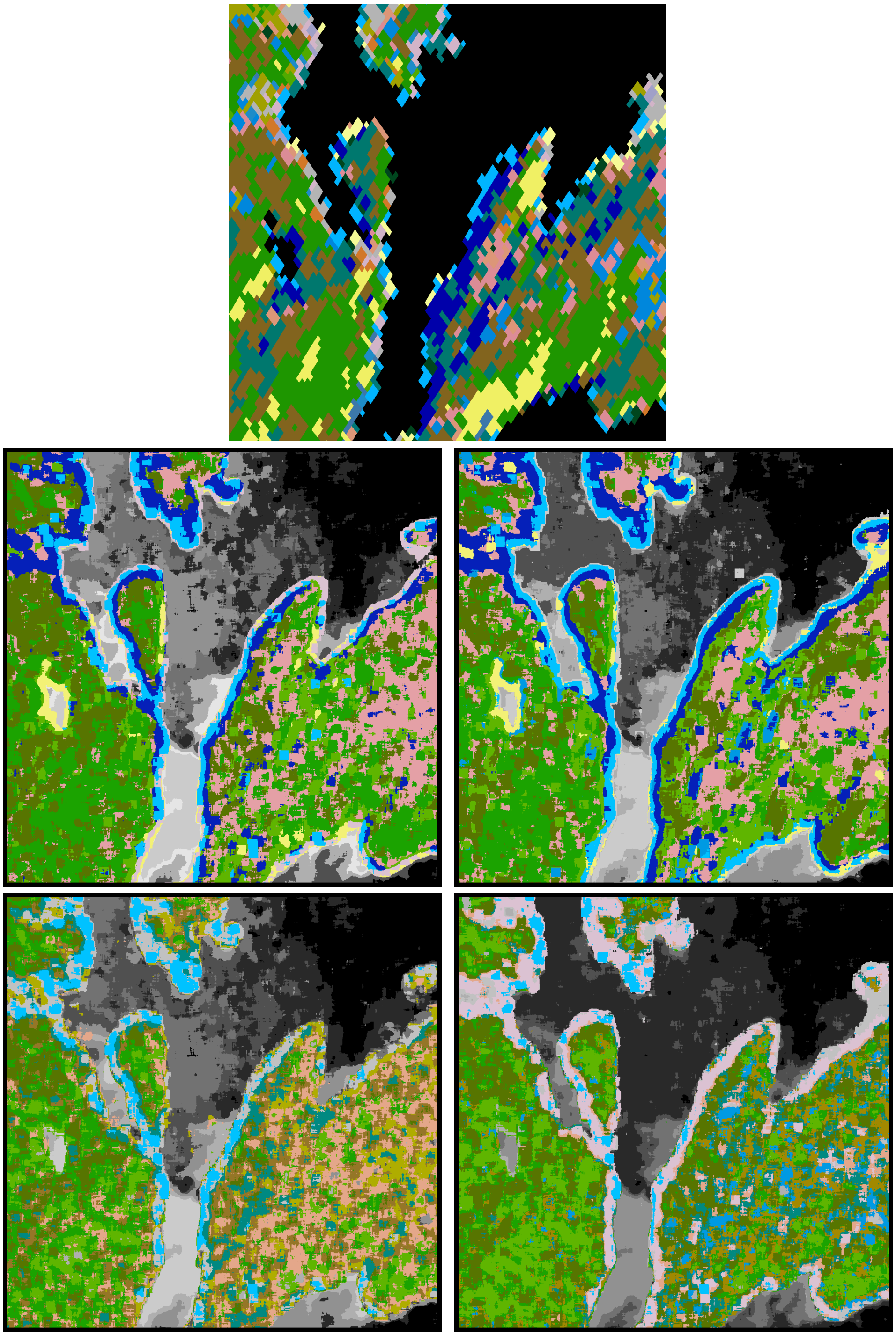


Figure 12.6: Land cover map (top), k-means (left) and BMoG (right) maps for ZG set (middle) and Gamma set (bottom).

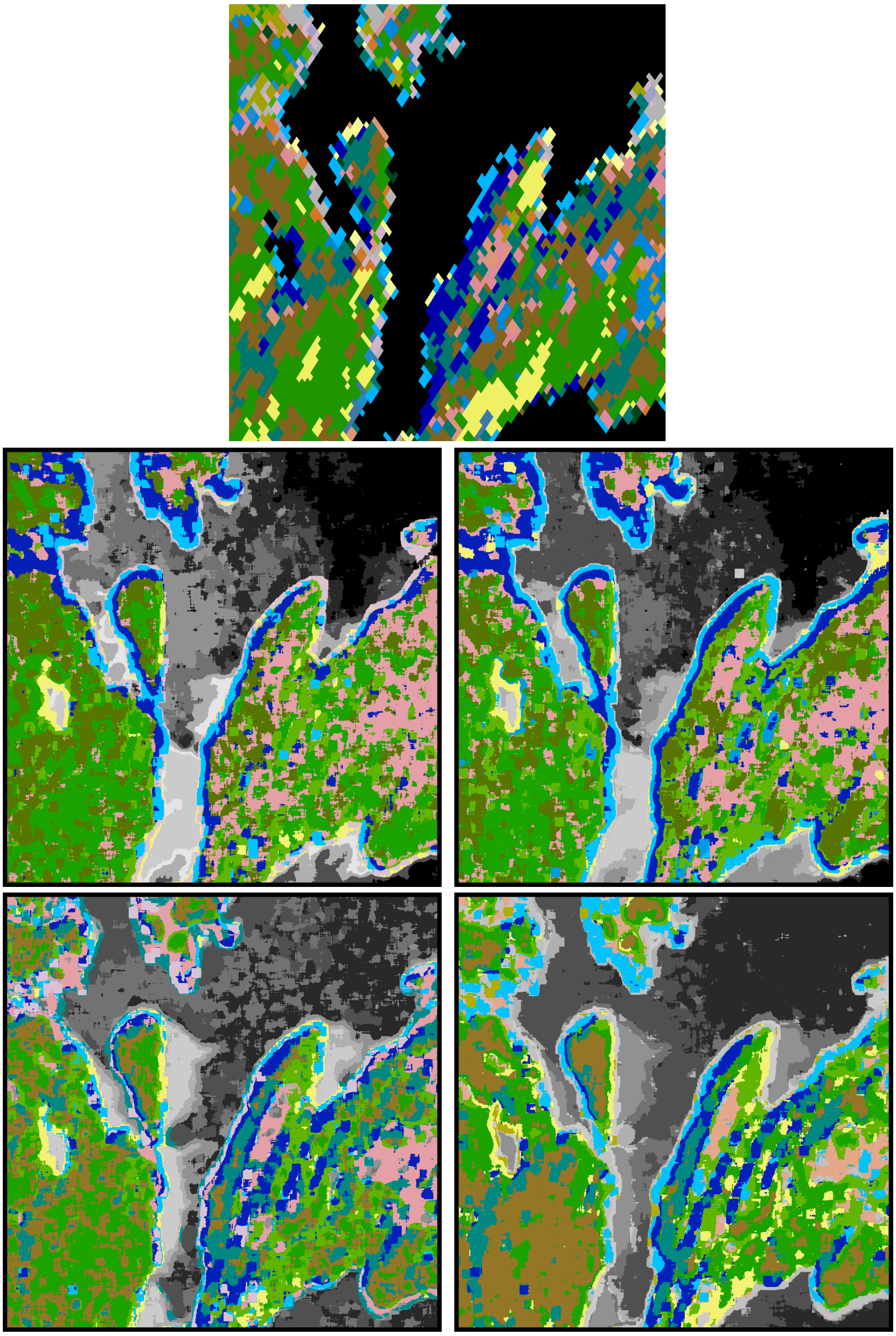


Figure 12.7: Land cover map (top), k-means (left) and BMoG (right) maps for C-band (middle) and L-band (bottom).

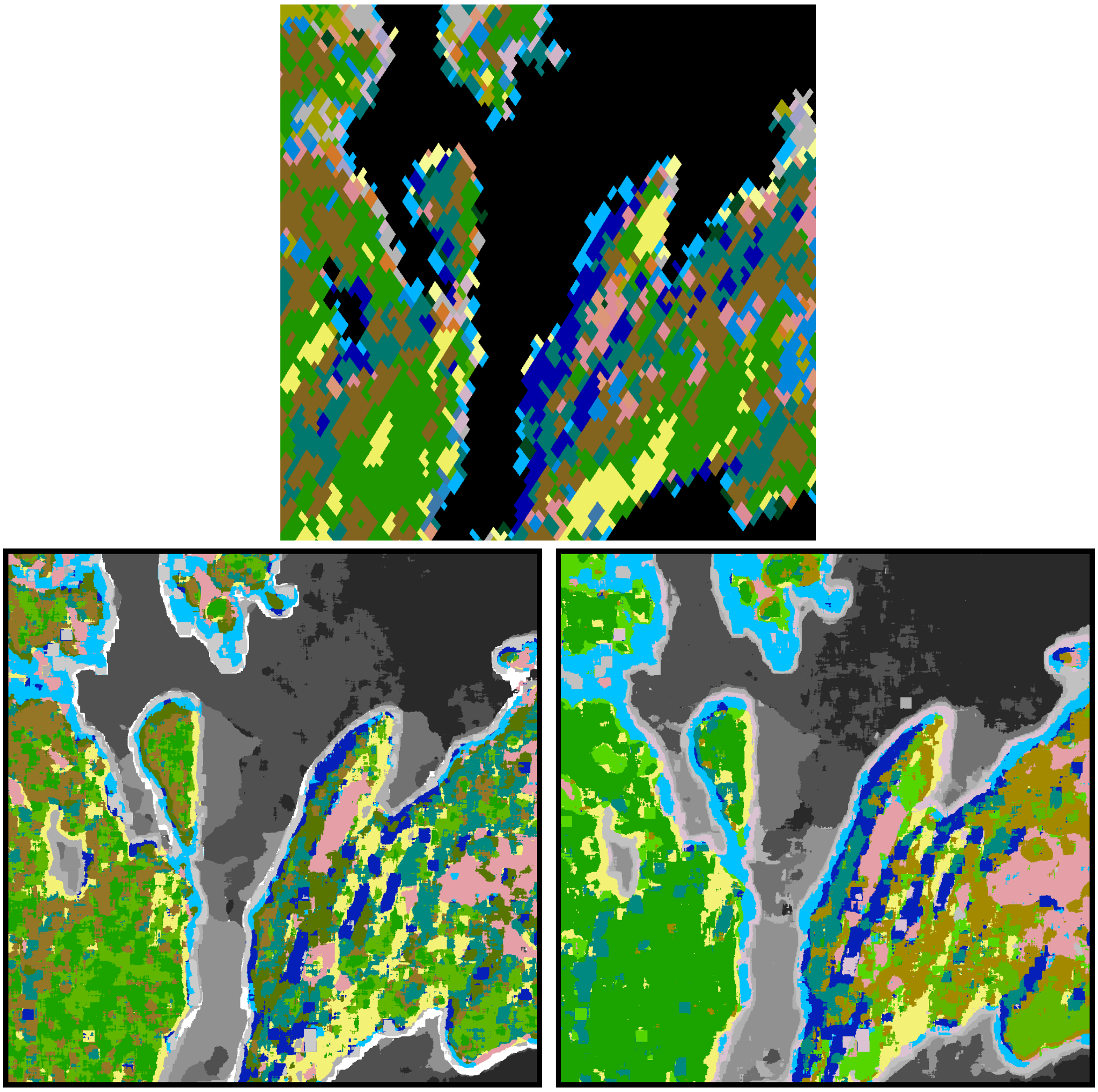


Figure 12.8: Land cover map (top), k-means (left) and BMoG (right) maps for combined C and L band ZG sets.

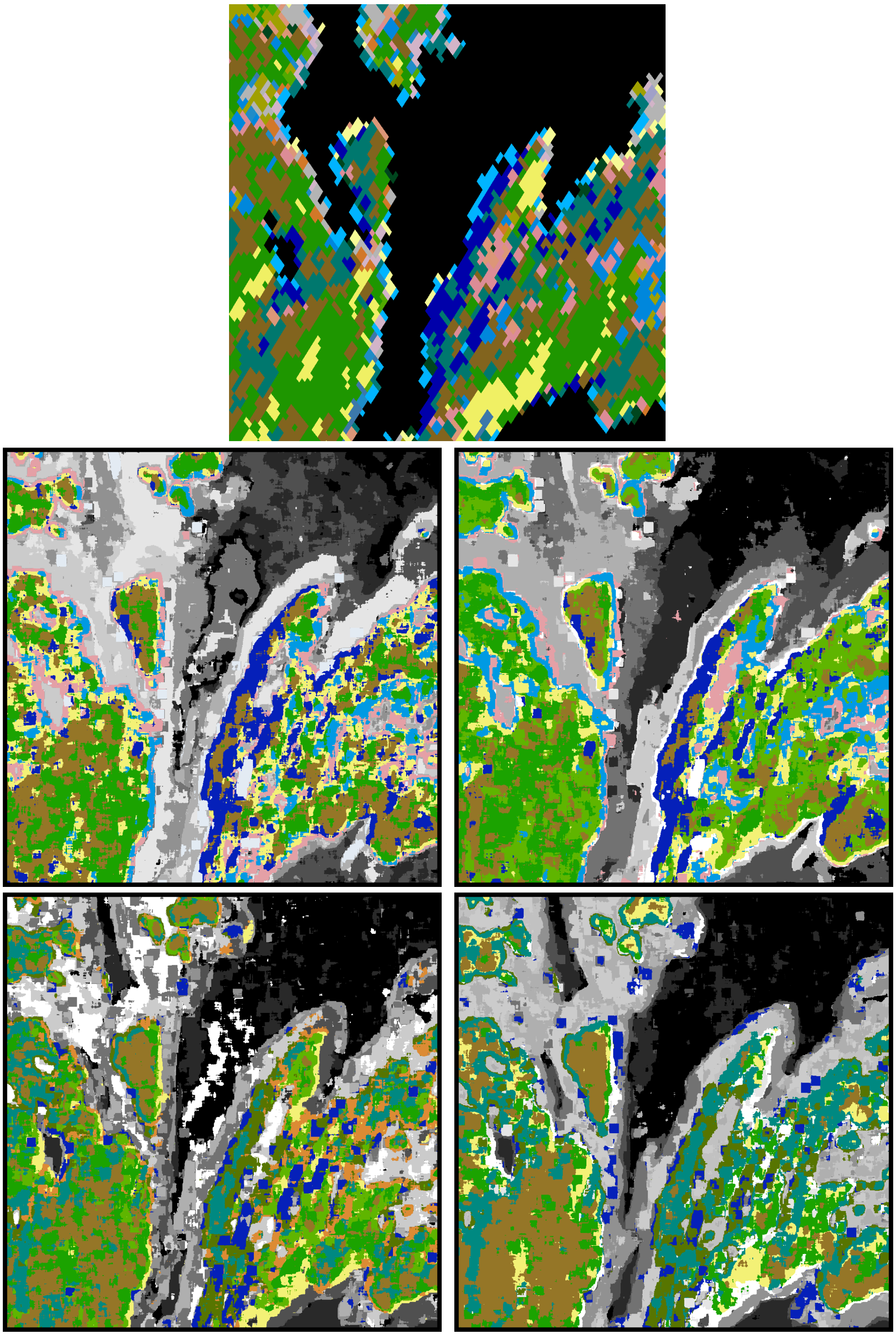


Figure 12.9: Land cover map (top), k-means (left) and BMoG (right) maps for C-band (middle) and L-band (bottom) from March ZG data.

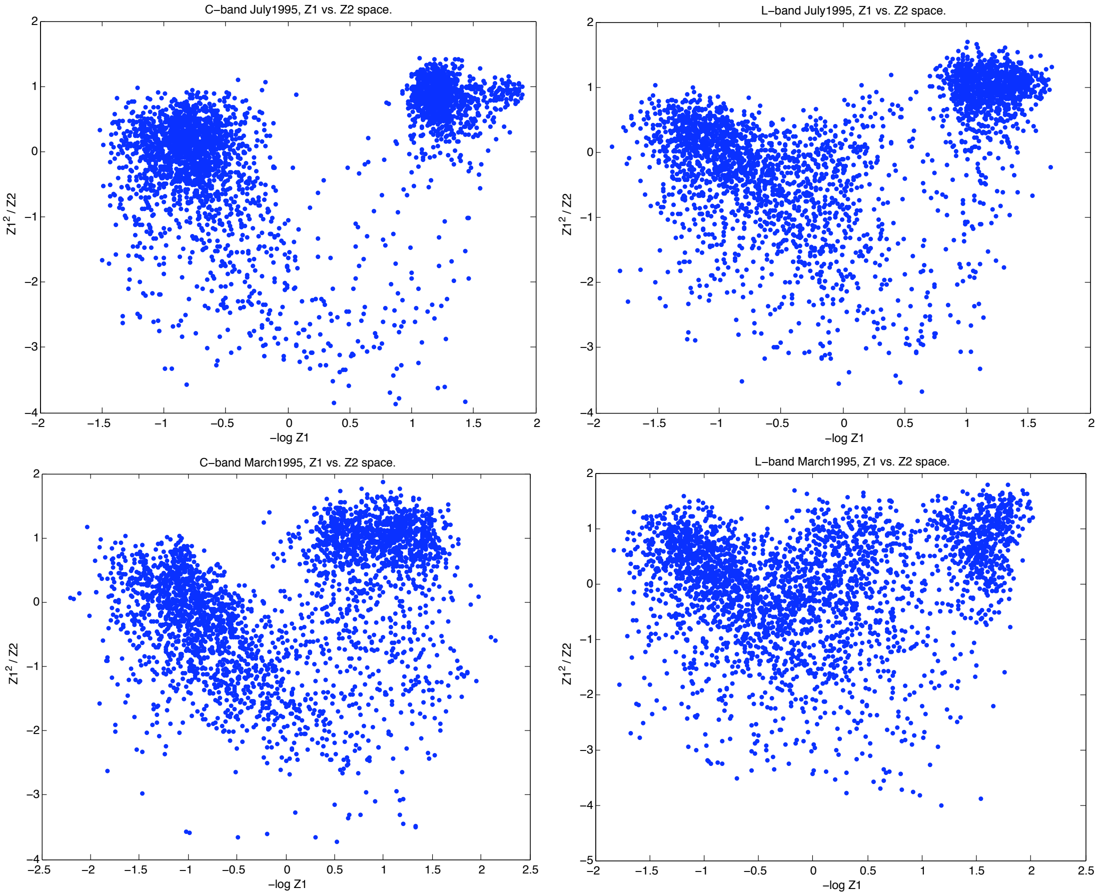


Figure 12.10: July (top) and March (bottom) Z space plots, for both C-band (left) and L-band (right) data sets.

Chapter 13

Summary II

Part II analysed the MK feature set resulting from the modelling and discovered that functional relations exist between the features. Largely visual methods were used to find suitable nonlinear transforms that reduced the dependencies and resulted in visual globular clusters. It is expected that the simple clustering methods would produce better results on such pre-transformed data, and more complex kernel methods may not be required.

The interpretation of the transformed feature space found a remarkable result that the new feature set had actually become detached from any specific scale mixture model. The transforms effectively changed the model's two scalar parameters back into empirical measures of *width* and *non-Gaussianity* of the sample distribution. This suggests that a general non-Gaussian analysis may be performed without requiring any specific model distribution. The main assumption remaining, from the scale mixture of Gaussian class, is that each dimension has an identically shaped distribution and width governed by the internal covariance structure matrix. The new feature set becomes an essentially multivariate Gaussian analysis, for width and $\mathbf{\Gamma}$ matrix elements, with the additional non-Gaussianity term.

When the modelling and clustering were performed on the test PolSAR data, the resulting images were found to be of far higher resolution with far more complexity than the available ground truth class map. This limited the possibility of rigourous testing and the only real comparison possible was by a crude visual judgement. Several comparisons are presented where some of the modelling and clustering options are compared, however, the overall effectiveness cannot be adequately evaluated.

The initial results are certainly realistic and encouraging, and further testing would certainly be worthwhile. To achieve this evaluation, it is suggested to find a test area with fully polarimetric SAR data that has a ground data classification map of at least similar resolution, preferably including high resolution optical data. It would be highly beneficial to include agricultural farming or planted forest areas that would have clearly defined boundaries and uniform vegetation areas. It may then be possible to obtain a quantitative accuracy measure by counting the classification in selected areas, and thus evaluate all the options that were simply viewed in the previous chapter.

Chapter 14

Conclusions

In hindsight this project resulted in two contrasting results, since the first part studied several parametric models in some detail and then the second part concluded that the model's details are not required for image segmentation. Both results are however important in different situations.

The modelling studied the class of multivariate scale mixtures of Gaussians and the interpretation of global shape, scale and internal covariance structure proved helpful. The concept of flexibility is another useful aspect of the models, when the objective is to fit a range of real data distributions, and it was shown that some model's distribution space includes that of other simpler models.

Two parameter estimation methods are compared, an iterative method and a moment based method. Accuracy and speed are compared with simulations and the moment based method was accepted as suitable for image analysis work, since it was very fast. The slower iterative method may prove useful for theoretical evaluations of the modelling, since it was more accurate.

Methods to test a multivariate model's goodness-of-fit were evaluated, with an integrated difference method and a likelihood based method compared in detail. The log-likelihood method was shown to be fast and accurate for high dimensional data. The four models were evaluated with the log-likelihood measure with several interesting results. The PolSAR data set was demonstrated to be highly non-Gaussian, as has been noted by others [1], and both the MK and MNIG models were shown to be quite good fits over the majority of the data, also observed by others [2, 15]. The inter-comparison of both the MK and MNIG however shows that the MNIG is a slightly better representative model for PolSAR data, which may have theoretical implications for scattering theory. Having one flexible model that represents virtually all of the data distributions means that only one model is needed for image processing and a single consistent parameter set results for the entire image.

Feature space investigations revealed nonlinear relations between the model's parameters. Theoretical moment expressions and some experimentation found suitable transforms that produced roughly linear spaced globular clusters that would be suitable for very simple clustering methods. Interpretation of the transforms revealed that the new features were independent of the specific model and were representatives of the sample distribution *width* and *non-Gaussianity.* This implies that the technique is a generalised scale mixture of Gaussians modelling method, whose main assumption is the global shape characteristic over all dimensions. The feature set is then interpreted as the multivariate Gaussian analysis for distribution width and covariance structure, with the additional term of non-Gaussianity.

The modelling, clustering and image segmentation were performed on a test PolSAR data area, and various procedural options are discussed. Unfortunately, the complexity of the natural environment and lack of good resolution ground data, made it impossible to evaluate the technique. Initial results look promising, with smoothed regions visually matching the main features of the available land cover map.

The method requires further rigourous testing and several suggestions are listed below.

Good ground truth data will be the key to evaluating the technique and two suggestions can be made. Firstly, seek test data that includes man-made regions such as agricultural fields or planted forests, to greatly simplify evaluation or to allow for quantitative measurement in specific areas. Secondly, the comparison truth data should be of similar resolution to the test data or there will always remain questions of whether details are real or artificial.

It would be interesting to perform the more accurate iterative analysis once to compare its clustering ability with the moment method's results. It would be anticipated that the improved accuracy may produce "cleaner" clustering results.

The estimation procedures discussed the application of various *a priori* constraints that were shown to be beneficial. The extension of these constraints, possibly combined with EM-style mixture modelling, could be a good way to introduce simple decomposition theorems, and could be worth investigating further.

Parametric characterisation of the results in terms of physical signatures may prove insightful, and possibly show parametric invariance. It may then be possible to perform one supervised clustering scheme to produce a template that may then be applicable to all subsequent analyses, even those from different areas.

Perhaps the goodness-of-fit threshold testing technique for mixture testing could be used to avoid some of the shoreline area mixing problems by determining when an area mix seems to be involved. The majority mixture component could then be analysed to represent that image location.

The best-fit results may be worth more rigourous testing, to confirm the preference for the MNIG model over the MK model, and pursue the theoretical implications. One key observation that may influence the fitting of real finite sample distributions, is the always finite peak of the MNIG, as opposed to the possible infinite peak of the MK model.

Finally, the scale mixture of Gaussians modelling was not restricted to PolSAR data, and it would be interesting to apply it to another field. Whether the same simplification to Gaussian modelling plus non-Gaussianity would be useful elsewhere would be an interesting study, as would application to very high dimensional spectral analyses.

References

[1] E. Jakeman and P. N. Pusey. A model for non-Rayleigh sea echo. *IEEE Trans. Antennas Propagat.*, 24, 6:806–814, November 1976.

[2] S. H. Yueh, J. A. Kong, J. K. Jao, R. T. Shin, and L. M. Novak. K-Distribution and polarimetric terrain radar clutter. *J. Electro. Waves Applic.*, 3:747–768, 1989.

[3] T. Eltoft, T Kim, and T-W. Lee. Multivariate scale mixture of Gaussians models. In *Proceedings of ICA 2006, Charleston, SC, USA*, March 2006.

[4] A. Freeman and S. L. Durden. A Three-Component Scattering model for Polarimetric SAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 36, 3:963–973, May 1998.

[5] O. E. Barndorff-Nielsen. Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scand. J. Statist.*, 24:1–13, 1997.

[6] W. G. Rees. *Physical Principles of Remote Sensing.* Cambridge University Press, Cambridge, UK, second edition, 2003.

[7] F. T. Ulaby and C. Elachi. *Radar Polarimetry for Geoscience Applications.* Artech House, 1990.

[8] C. Olivar and S. Quegan. *Understanding Synthetic Aperture Radar Images.* Artech House, 1988.

[9] D. R. Sheen and L. P. Johnston. Statistical and Spatial Properties of Forset Clutter Measured with Polarimetric Synthetic Aperture Radar (SAR) . *IEEE Trans. Geoscience and Remote Sensing*, 30, 3:578–588, May 1992.

[10] A. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B*, 36:99–102, no. 1 1974.

[11] T. Eltoft, T Kim, and T. Lee. A multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13, 5:300–303, May 2006.

[12] T. Eltoft. The Rician inverse Gaussian distribution: A new model for non-Rayleigh signal amplitude statistics. *IEEE Trans. Image Process.*, 14:1722–1735, November 2005.

[13] D. M. Drumheller and H. Lew. Homodyned-K Fluctuation Model. *IEEE Trans. Aerospace and Electronic Systems*, 38, 2:621–632, April 2002.

[14] E. Jakeman and R. J. A. Tough. Generalized K distribution: a statistical model for weak scattering. *J. Opt. Soc. Am. A*, 4, 9:1764–1772, September 1987.

[15] T. A. Øigård, A. Hanssen, R. E. Hansen, and F. Godtliebsen. EM-estimation and modeling of heavy-tailed processes with the multivariate normal inverse Gaussian distribution. *Signal Processing*, 85, 8:1655–1673, 2005.

[16] A. A. D'Souza. Using EM To Estimate A Probability Density With A Mixture Of Gaussians. *Internet publication*, 2003.

[17] K. V. Mardia. Measure of Multivariate Skewness and Kurtosis with Applications. *Biometrica*, 57, 3:519–530, December 1970.

[18] J. A. Blimes. Factored sparse inverse covariance matrices. In *Proceedings of ICASSP*, June 2000.

[19] I. R. Joughin, D. B. Percival, and D. P. Winebrenner. Maximum Likelihood Estimation of K Distribution Parameters for SAR Data. *IEEE Transactions on Geoscience and Remote Sennsing*, 31, 5:989–999, September 1993.

[20] J. W. Goodman. Some fundamental properties of speckle. *J. Opt. Soc. Am.*, 66, 11:1145–1150, November 1976.

[21] S. R. Cloude and E. Pottier. A Review of Target Decomposition Theorems in Radar Polarimetry. *IEEE Transactions on Geoscience and Remote Sensing*, 34, 2:498–518, March 1996.

[22] E. Jakeman. On the statistics of K-distributed noise. *J. Phys. C*, 13:31–48, 1980.

[23] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, San Diego, 1999.

[24] R. J. Larsen and M. L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, New Jersey, 1986.

[25] J. Principe, D Xu, and J. Fisher. Information Theoretic Learning. *Unsupervised Adaptive Filtering*, Volume 1, Chapter 7, 2000.

[26] A. Renyi. On Measures of Entropy and Information. *Selected Papers of Alfred Renyi, Akademiai Kiando, Budapest*, 2:565–580, 1976.

[27] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[28] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, second edition, 2004.

[29] C.E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol. 27:379–423, 623–656, JUL, OCT 1948.

[30] S. R. Cloude. Target decomposition theorems in radar scattering. *Electronics Letters*, 21, 1:22–24, January 1985.

[31] R. Jenssen, T Eltoft, and J. C. Principe. Information Theoretic Spectral Clustering. *IEEE proceedings on Neural Networks*, 1, July 2004.

[32] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing.* Addison-Wesley, Reading, Massachusetts, first edition, 1992.

[33] Henry Stark and John W. Woods. *Probability and Random Processes with Applications for Signal Processing.* Pearson Prentice Hall, New Jersey, third edition, 2002.