UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

DEPARTMENT OF PHYSICS AND TECHNOLOGY

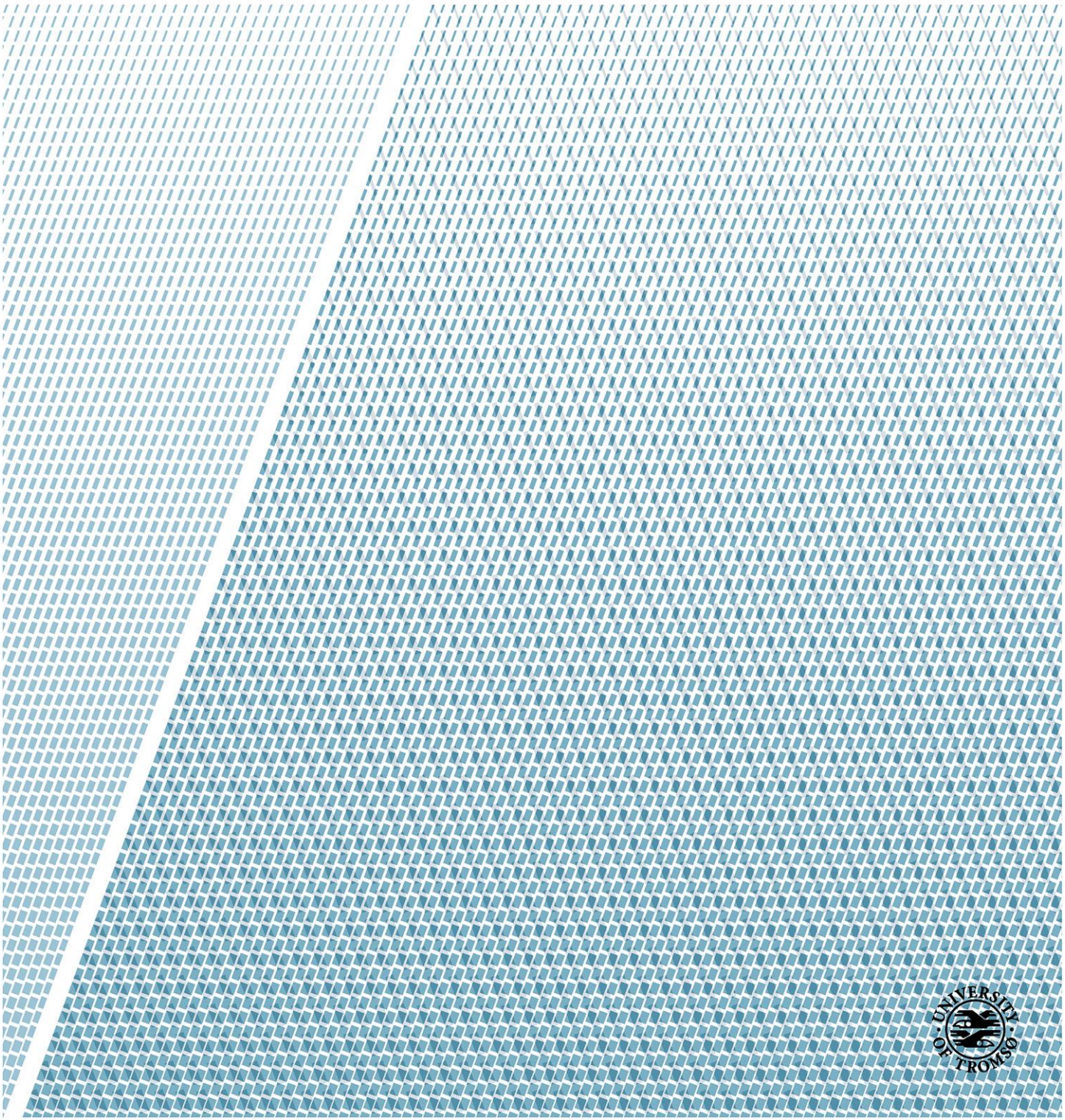
A Classification Strategy for Multi-Sensor Remote Sensing Data

An analysis and implementation of Meta-Gaussian classification schemes

—

Arja Beate Kvamme

FYS-3941 Master's thesis in applied physics and mathematics
August 2017



Abstract

In integrated remote sensing, one of the objectives is to create reliable services by combining information from various data sources. The combination of multiple data sources is often denoted "data fusion", and is a topic that has high interest in remote sensing applications. In this thesis, we devise a classification strategy for multi-sensor remote sensing data, based on the strategy presented in the paper "On the Combination of Multisensor Data Using Meta- Gaussian Distributions" [1]. The classification method uses data fusion through a transformation of variables into a multivariate Meta-Gaussian distribution, and correct assumptions or estimates of the marginal probability density functions is an important key in this transform. We found that using general parametric probability density functions, or kernel estimates were valid in a supervised classification setting, with no need to specify individual marginals based on the true underlying distribution. Further, we found that classification based on the Meta-Gaussian function, using transformed variables, surpassed that of a standard multivariate Gaussian function. Unsupervised classification based on the same strategy was implemented in a generalized mixture decomposition algorithmic scheme framework. Current results are positive, and indicate that this method has potential when it comes to combining multi-sensor remote sensing data.

Acknowledgements

So long, and thanks for all the fish. Also, big thanks to my supervisors Torbjørn Eltoft and Anthony Doulgeris, who gave me just the right amount of space and fear of disappointment. I truly hope I managed to live up to your expectations. Thanks to the coffee machine for keeping me going through long days and short nights, thanks to the computer for not dying on me, Bjørnbert INC for providing server access, Kristoffer for pushing me towards the finish line, the few people in my class for giving me a competitive edge. And all the people I met during my studies, both here at UiT, and at UNIS. And an extra special thank you to the proofreaders Cornelius, Johannes and Magnar, I hope your time was well spent.

Contents

List of Figures

List of Tables

Abbreviations

SAR	Synthetic aperature radar
VNIR	Visible and near infrared
IR	Infrared
NIR	Near infrared
NDSI	Normalised difference snow index
SWIR	Shortwave infrared
SLC	Single look complex
MLC	Multi look complex
PDF	Probability density function
CDF	Cumulative density function
NQT	Normal quantile transform
PolSAR	Polarimetric SAR
GMDAS	Generalized Mixture Decomposition Algorithmic Scheme
PolSAR	Polarimetric synthetic aperture radar
ESA	European Space Agency
GTC	Geometric terrain correction
RTC	Radiometric terrain correction
SNR	Signal to noise ratio
USGS	United States Geological Survey
PCA	Principle component analysis
DTDR	Data transformation and dimensionality reduction

Chapter 1

Introduction

In the world of today, information is key, and there is an abundance of it. Which is why a simple equation for indexing webpages, the PageRank [2] turned out to be the humble start of one of the now largest companies in the world [3], Google. But accessing information is one thing. To be able to take fully advantage of the information that is available, and utilizing it is something entirely different.

One of the problems faced by the remote sensing community today is not the lack of information— nor its accessibility, thanks to the public access policy held by a lot of data providers, such as the European Space Agency (ESA) [4] and the United States Geological Survey (USGS) [5] of which most have an easy way to find specific data over specific regions.

The problem is processing, converting what the satellite measures into something that is more interpretable than plain radiance or backscatter values. For human interpretation, this could be something as easy as creating a RGB image using optical bands, or adjusting the contrast to make objects of interest more visible. Visual interpretation was for a long time the only way to analyze images, and it wasn't untill the first Landsat mission that digital image representation became relatively accessible [6].
For an automated classification procedure however, the visual appearance of the data is not necessarily a factor. This is an advantage, because we, humans, have sensory limitations. We are not able to see beyond the visible spectrum, and we can't simultaneously process millions of values from numerous sources looking for connections. Unsupervised classification, also known as clustering, can be used to find such connections.

At the same time, the data that is input into any classification algorithm

will of course have an impact on the end result. Carefully selected features derived from the original data can help improve classification accuracy, and at the same time give them more validity by connecting them to a physical property. In terms of unsupervised classification, such features can also help us to interpret the classification results.

In this thesis, we aim to test a classification method that combines data from different sensors, through a transformation into Meta-Gaussian variables. Using both raw data and generated features, we will test the method to see if we can improve classification results.

1.1 Earlier Work

The methodology that is used in this thesis largely stems from the paper *"On the Combination of Multisensor Data Using Meta-Gaussian Distributions."* by Storvik et al [1]. Here they propose a method to combines images obtained from different sensors, creating a joint distribution, and at the same time preserving any correlation between the images. This method is further described in section 2.2, "Data fusion using Meta-Gaussian distributions". Similar methods for data fusion using copulas, have been around since 1940 [7], although the term "copula" did not arise until 1949 [8]. Copulas are multivariate probability distributions, whose marginals are all uniform distributions. They are popular in fields such as finance [9] climate modelling [10,11], and is also used in remote sensing [12], where the dependency between different marginal structures is required, and difficult to model using other conventional distributions.

In [1], they only considered the case of supervised classification, and in their testing procedure, the marginal probability distribution functions were considered to be Gamma, K or Gaussian. We hope to expand on this.

1.2 Objective

The aim of this thesis is to build upon the previous work, [1], by Storvik et al, which was also recreated in the pilot project "Classification strategy for multi-sensor data using Meta-Gaussian distribution" [13] and extend the method to include:

- A clustering step to support unsupervised classification.

- Generalization of the marginal probability distribution functions(PDF). In many cases real data may not be well described by parametric models. In these cases, non-parametric, kernel based approximations of the PDF may prove to be better alternatives.

- Extensive testing on a multitude of data, both real and simulated.

Results from single-sensor classification and multi-sensor classification will be compared. The end goal would be to develop a multi-class classification algorithm based on the Meta-Gaussian data fusion method.

1.3 Structure of the thesis

- **Chapter 1: Introduction**
  Was a general introduction into some of the challenges faced by the remote sensing community, generally and in an Arctic perspective. We also presented the objective of this thesis.

- **Chapter 2: Data Fusion**
  We describe the main motivation behind data fusion and how it is used in the Meta-Gaussian. A brief description of some typical probability distributions in remote sensing, as well as some more general, is also included, due to the importance of the marginal probability distribution functions in the Meta-Gaussian.

- **Chapter 3: Remote Sensing Data**
  Reviews the principles of optical and SAR imaging, in terms of acquisition and use.

- **Chapter 4: Pattern Recognition**
  Introduces some general principles of pattern recognition, as well as the specific classification methods used in this thesis.

- **Chapter 5: Data and Features**
  A description of features, and the different simulated and real multi-sensor data sets used.

- **Chapter 6: Implementation and Results**
  Describes how the two different classification schemes were implemented, and presents the different experiments that were conducted, as well as their results.

- **Chapter 7: Conclusion**
  Concludes the thesis. We summarize the findings of this thesis, and present some ideas for future work.

1.4 Remote Sensing of the Arctic

The Arctic is an important part of the world, politically, sociologically, economically and climatologically. It is defined as the part of the Earth that is inside the Arctic circle, which is currently, as of 1 May 2017 at 66.33 degrees latitude, and moving further north at a rate of approximately 15 meters per year. Eight countries have areas inside the Arctic circle, and over the years there have been many disputes over territorial claims. In figure 1.1 a map of the Arctic is shown. According to the report *Snow, Water, Ice, Permafrost in the Arctic* [14], the Arctic has seen a 50% decline in the extent of sea ice , as well as a 75% loss of volume in the last 30 years. According to the predictions made in this report, we could have summers without any Arctic sea ice by 2040.

Remote sensing is an important tool when it comes to observing the cryosphere. The area that lies above the Arctic circle has many attributes such as harsh weather, low temperatures, and a general remoteness, which makes fieldwork and in situ measurements unfavourable. The different interests in the Arctic often require a larger field of view and usually a high temporal resolution. Remote sensing is actively used in applications such as glacier monitoring, ship and iceberg detection, ice maps, measuring the melt period and extent, and total mass loss or gain over an area. Currently, there are many different forms of remote sensing techniques that are used for assessing the state of the Arctic, such as:

- Optical imagery
- Laser altimeters
- EM-birds (Electro-magnetic)
- Radar altimeters
- Gravimetric imaging (GRACE)
- SAR

Each and every technique with its own limitations. Passive optical imagery in the visible domain cannot function in the polar night, due to lack of solar

illumination, and are unable to penetrate cloud cover. EM-birds flown by helicopters are expensive to use, and have limited coverage. Gravimetric imaging currently has a very coarse spatial resolution, and is mainly used for estimating total mass over large regions. Radar and laser altimeters are mainly limited to measuring the distance between an object and the sensor.

1.4.1 Classification of Sea Ice

Mapping of sea ice has both commercial interests, and environmental interests. Sea ice extent, and the amount of multi-year ice are important indicators of climate change, and plays major roles in the modelling of temperature prognosis. Sea ice and snow both have an higher albedo than water, and the decrease of it may speed up global warming. In addition to this, having an accurate and recent ice map is important for ships travelling in the Arctic region, not only for scientific, but also recreational and other commercial uses. The recent decline in sea ice has opened up many new shipping routes, that can decrease the travel times.

Up till now, such ice maps have been created manually, but in recent years there has been a efforts in creating automated procedures for the classification of sea ice, and consequently generating sea ice maps [16–18]. This is believed to enable an increase in both the temporal and spatial resolution of the product when compared to manual maps. The Norwegian Meterological Institute currently have such a automated system operative [16], but they emphasise that this is currently only to be considered as a supplement to the manually created ice maps.

For such a product to be operational, and have any benefit over the manual maps, it requires a high confidence in the automated results, and as of now, most of these attempts are still being compared to the manual maps where those are available.

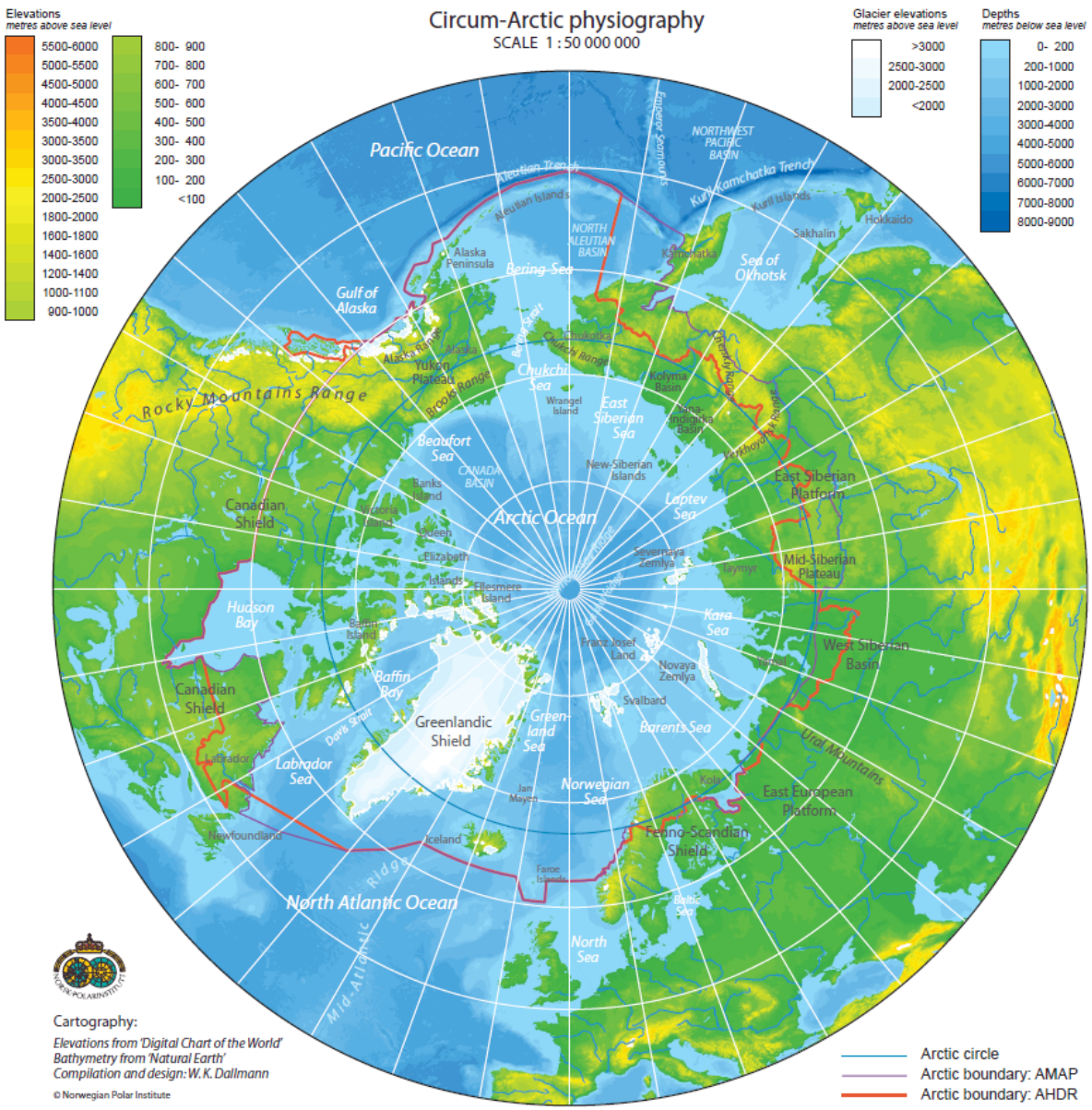


Figure 1.1: Map of the Arctic and surrounding areas. Arctic circle at the time is shown in blue.
Image credit: Cartography: Elevations from "Digital Chart of the World". Bathymetry from "Natural Earth". Compilation and design by Winfried K. Dallmann. Copyright Norwegian Polar Institute [15]

Chapter 2

Data Fusion

2.1 Background on Data Fusion

Data fusion is a concept that has been around for a long time, and it is still a highly relevant subject, in part due to the abundance of data that now is available, and its benefit in terms of utilizing this data.

In the field of remote sensing, this data can come in the form of satellite imagery from a multitude of sensors. Other sources of data are digital elevation models, weather records, graviometric data, and aerial photography to name a few. This data can be represented in many different ways, such as binary, continous, and labeled. These data sources are useful on their own, but when they are combined with each other potentially they can potentially give a better end result, whether it is in terms of classification or determination. Typically data fusion is either a means to achive *a more reliable determination* or, to *generate an interpretation of the scene not obtainable with data from a single sensor* [19]

The main motivation behind data fusion is either that of increasing the reliability of a decision, or through the combination of different sensor data, make decisions that were not possible through single sensor data.

In terms of sensor fusion, which is the focus in this thesis, the process of data fusion is often divided into three different sub-genres, depending on in which part of the processing level the data that is being fused.

1. Pixel level fusion

Images, or image bands are fused on a pixel to pixel basis. This can be done through a variety of methods, the simplest being a straightforward merging of the data. Ie, consider a pixel being represented by two different pixel values, I_1 and I_2. A simple fusion method would be to collect these two into a feature vector, $I = [I_1, I_2]$. This could be built upon to include a weighting factor on the different bands, ie. $I = [\rho_1 I_1, \rho_2 I_2]$ which could account for the reliability of the sensor, or the importance it is given. Other, more sophisticated methods such as Markov Random Fields and Simulated Annealing to derive cost functions for the fusion process, Artificial Neural Networks, and Wavelets, [20] have also been used. Evaluation is then performed on the fused data.

2. Feature level fusion

Features are extracted either independently from each image band, or they may be formed by a combination of several bands. An example of this is through a commonly used method of data transformation, the Karhunen-Loève transform [21]. Perhaps more commonly known as the principle component analysis (PCA) [22, 23]. It transforms multidimensional data into a data set that is linearly uncorrelated. For SAR data, features derived from a combination of polarimetric channels will often be more valuable for classification than using the raw backscatter values. The selected features are then fused, similar to the fusion performed on pixel level. Evaluation is performed on the fused data. [24]

3. Decision level fusion

Pixels are identified/classified separately for each of the images, forming decisions for each pixel, and each image. The decisions of the separate images are then combined to either give a better classification, or more robust decisions [24], such as, if a pixel has a normalized difference vegetation index (NDVI) at, or above 0.6, and cloud mask of 0, classify as vegetation.

Hybrid methods

In multi-sensor remote sensing data, some correlation between channels depicting the same area of interest is natural, and expected, but exploiting this

natural correlation has not been of high focus in the world of data fusion. There are many reasons for this, and the most prominent one is that modelling multivariate distributions (of which a point is assumed to belong to), that has varying marginal distributions is not easily done. The method of data fusion proposed in [1] aims to preserve this correlation and use it in the classification. In doing so, the method cannot be said to strictly belong to the pixel level fusion, nor to the feature level. The data is fused at pixel level, and based on this, multivariate Meta-Gaussian distributions are formed. The classification itself is done in a Bayesian framework, on transformed data, which is essentially feature vectors, created based on the transformation found after the fusion at pixel level. In such, the method considered here can be deemed as both a pixel level fusion, and a decision level fusion method.

2.2 Data Fusion Using Meta-Gaussian Distributions

The Meta-Gaussian distribution in the form that is used in this thesis, was first proposed in [25], as a way of representing a joint distribution for detected[1] radar images, whilst at the same time preserving the correlation between different image bands. This method was also reviewed in [13], but we repeat it here in a modified form for the sake of understanding some of the mathematics and derivations behind the concept. The combination of multisensor data using the Meta-Gaussian distribution is suggested by Storvik et al [1] as a method to improve classification on data that is a combination of images from different sensors. It allows for data from different marginal distributions to be joined in a multivariate distribution, the so called Meta-Gaussian distribution. In order to do this, they suggest these three simple steps.

1. Transform the marginal data to the standard Gaussian distributed variables.

2. Model the dependences between the marginal data through correlations of the transformed data.

3. Derive the distribution of the original data by using the inverse transform of 1) assuming the dependence given in 2).

[1]Detected SAR images are amplitude images created by applying a non-linear transformation to the sum of the squares of the in-phase and quadrature components of the radar signal [26]

The main motivation for this method is that it gives us a simple way of modelling dependencies between the different components through the correlation matrix $\mathbf{C}$ of the distribution function that is found.

First we need to know the probability distribution of the data, and its cumulative distribution function. To transform the marginal data, that we now assume has a marginal probability density function g_j, and a corresponding cumulative distribution function G_j, into a new marginal probability density h_j with corresponding cumulative distribution function H_j, we can use the standard transformation rule for random variables.

$$y_j(H^{-1}(G_j(x_j)) \tag{2.1}$$

As we want to transform the data into a standard Gaussian density ($y \sim N(0,1)$) we simply choose H to be the cumulative distribution of the standard Gaussian distribution function. This gives us:

$$y_j = y_j(x_j;\gamma_j) = \Phi^{-1}(G_j(x_j;\gamma_j)) \tag{2.2}$$

where γ_j is a vector containing the parameters of the distribution g_j. Since the transform function Φ is continuous and non-decreasing, there will be a one-to-one correspondecy between y_j and x_j, which allows us to go back to our original distribution in a later step. In theory, any continuous, injective function that retains a one-to-one relationship could be used. For instance, in [27], the uniform distribution was used as a transformation function in a similar method using multidimensional copulas. The transform between y_j and x_j is given by:

$$x_j = x_j(y_j;\gamma_j) = G_j^{-1}(\Phi(y_j);\gamma_j) \tag{2.3}$$

where Φ is the cumulative distribution function of the standard normal distribution,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \tag{2.4}$$

If we now define

$$\mathbf{y}(\mathbf{x};\gamma) = (y_1(x_1;\gamma_1), .., y_p(x_p;\gamma_p))^T \tag{2.5}$$

where $\gamma = (\gamma_1, \gamma_2, ...\gamma_p)$. Through our transformation, we now have a set of variables, $y_i \sim N(0,1)$, and the dependency between these transformed variables are found through the correlation matrix, $\mathbf{C}$ of the transformed data $\mathbf{y}$. The distribution for $\mathbf{y}$ is then given by:

$$\tilde{f}(\mathbf{y};\mathbf{C}) = \frac{e^{-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}}}{|2\pi\mathbf{C}|^{1/2}} \tag{2.6}$$

which we recognize as the multivariate normal distribution, with zero mean.

So, moving on to what we really want to find, namely the distribution, $f(\mathbf{x};\gamma,\mathbf{C})$ for our "raw" data, $\mathbf{x}$ [28]. This can be found through the distribution of $\mathbf{y}$, and the transformation between $\mathbf{y}$ and $\mathbf{x}$.

$$f(\mathbf{x};\gamma,\mathbf{C}) = \tilde{f}(\mathbf{y};\mathbf{C})|\mathbf{J}| \tag{2.7}$$

We already know the distribution of $\mathbf{y}$, and all we need to find is the determinant of the Jacobian matrix for the transformation between $\mathbf{y}$ and $\mathbf{x}$.

We start by finding the Jacobian matrix, $\mathbf{J}$, of the transformation from $\mathbf{y}$ to $\mathbf{x}$. The transform from $\boldsymbol{y}$ to $\boldsymbol{x}$ is in the $R^p \leftarrow R^2$ space, The Jacobian of the transform is given by:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1(y_1;\gamma_1)}{\partial y_1} & \frac{\partial f_2(y_1;\gamma_1)}{\partial y_2} & \dots & \frac{\partial f_p(y_1;\gamma_1)}{\partial y_p} \\ \frac{\partial f_1(y_2;\gamma_2)}{\partial y_1} & \frac{\partial f_2(y_2;\gamma_2)}{\partial y_2} & \dots & \frac{\partial f_p(y_2;\gamma_2)}{\partial y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(y_p;\gamma_p)}{\partial y_1} & \frac{\partial f_2(y_p;\gamma_p)}{\partial y_2} & \dots & \frac{\partial f_p(y_p;\gamma_p)}{\partial y_p} \end{bmatrix} \tag{2.8}$$

Which equates to zero for any off-diagonal elements, For diagonal matrices, such as this one, the determinant of the Jacobian is then given as

$$|\mathbf{J}| = \prod_{j=1}^{p} \mathbf{J}_{jj}, \tag{2.9}$$

that is, through taking the product of the diagonal elements.

$$|\mathbf{J}| = \prod_{j=1}^{p} \frac{\partial(\Phi^{-1}(G_j(x_j;\gamma_j)))}{\partial y_j} \tag{2.10}$$

We then calculate the partial derivatives of the Jacobian, starting with using the core rule on

$$\frac{\partial(\Phi^{-1}(G_j(x_j;\gamma_j)))}{\partial y_j} = \frac{\partial(\Phi^{-1}(U)}{\partial y_j}\frac{\partial G_j(x_j;\gamma_j)}{\partial y_j} \tag{2.11}$$

where we have used U to the denote the core $G_j(x_j;\gamma_j)$. Recall that for an inverse function, such as $\Phi^{-1}(\cdot)$, we will have

$$\frac{\partial \Phi^{-1}(U)}{\partial U} = \frac{1}{\frac{(\partial \Phi(U)))}{\partial U}(\Phi^{-1}(U))} \tag{2.12}$$

And for CDFs such as G_j and Φ, we have that the derivatives are equal to the PDFs. This gives us for the determinant of the Jacobian

$$|\mathbf{J}| = \prod_{j=1}^{p} \frac{\partial(\Phi^{-1}(U))}{\partial y_j} \frac{\partial G_j(x_j;\gamma_j)}{\partial y_j} = \frac{1}{\phi(\Phi^{-1}(G_j(x_j;\gamma_j))} g_j(x_j;\gamma_j) \tag{2.13}$$

and we recognize $\Phi^{-1}(G_j(x_j;\gamma_j)) = y_j(x_j;\gamma)$, so it all condenses into

$$|\mathbf{J}| = \prod_{j=1}^{p} \frac{g_j(x_j;\gamma_j)}{\phi(y_j(x_j;\gamma_j))} \tag{2.14}$$

We get the distribution for $\mathbf{x}$, given the parameters $\phi = (\phi_1, ..., \phi_p)$ and $\mathbf{C}$

$$f(\mathbf{x};\gamma,\mathbf{C}) = \frac{e^{-\frac{1}{2}\mathbf{y}(\mathbf{x};\gamma)^T\mathbf{C}^{-1}\mathbf{y}(\mathbf{x};\gamma)}}{|2\pi\mathbf{C}|^{1/2}} \prod_{j=1}^{p} \frac{g_j(x_j;\gamma_j)}{\phi(y_j(x_j;\gamma_j))} \tag{2.15}$$

This can then be simplified by using the fact that

$$\prod_{j=1}^{p} \phi(y_j(x_j;\gamma_j)) = \frac{e^{-\frac{1}{2}\mathbf{y}(\mathbf{x};\gamma)^T\mathbf{I}\mathbf{y}(\mathbf{x};\gamma)}}{|2\pi\mathbf{I}|^{1/2}} \tag{2.16}$$

Through

$$\prod_{j=1}^{p} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.17}$$

And we have that $\sigma = 1$ and $\mu = 0$ for all y_j , such that

$$\prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} \tag{2.18}$$

For the denominator we have that

$$\prod_{j=1}^{p} \sqrt{2\pi} = \sqrt{2\pi}\sqrt{2\pi}\sqrt{2\pi}...\sqrt{2\pi} == \sqrt{|2\pi\boldsymbol{I}|} \tag{2.19}$$

Where $\boldsymbol{I}$ is the identity matrix, of dimension $p \times p$ and $|\cdot|$ indicates the determinant. For the nominator, we have

$$\prod_{j=1}^{p} e^{-\frac{1}{2}y_j(x_j;\gamma_j)^T\mathbf{I}y_j(x_j;\gamma_j)} = e^{-\frac{1}{2}(y_1(x_1;\gamma_1))^2}...e^{-\frac{1}{2}(y_{p-1}(x_{p-1};\gamma_{p-1}))^2}e^{-\frac{1}{2}(y_p(x_p;\gamma_p))^2} \tag{2.20}$$

multiplication of an exponential with the same base term is equivalent to adding the exponents.

$$e^{-\frac{1}{2}((y_1(x_1;\gamma_1))^2+y_2(x_2;\gamma_2))^2+...+y_p(x_p;\gamma_p))^2} \tag{2.21}$$

So that our final expression for the distribution of $\mathbf{x}$ is given by

$$f(\mathbf{x};\gamma,\mathbf{C}) = \frac{e^{-\frac{1}{2}\mathbf{y}(\mathbf{x};\gamma)^T(\mathbf{C}^{-1}-\mathbf{I})\mathbf{y}(\mathbf{x};\gamma)}}{|\mathbf{C}|^{1/2}} \prod_{j=1}^{p} g_j(x_j;\gamma_j) \tag{2.22}$$

Classification is then performed using the classical Bayes rule

$$\hat{z}_i = \arg\max_k \pi_k f_k(\mathbf{x}) \tag{2.23}$$

2.2.1 Estimation of Parameters

What is the importance of the marginal model? One of the main obstacles in the method of data fusion using Meta-Gaussian distribution introduced by [1] was that the parametric model for the marginal distribution needed to be specified. This will either require some knowledge about the data beforehand, or simply testing the procedure with different parametric models and then choosing the one with the best goodness of fit.

Alternative ways of choosing these models can be through either:
1. Using a more flexible model that is assumed to fit for all marginals.
2. Using a non-parametric model that will adept to fit each marginal.

The objective of this section is to investigate and determining the effects of these three choices. Ideally, a flexible parametric model, or a non-parametric model, which will be flexible, is what we would like to be possible.

2.2.2 Estimation of marginals

A vital part of this method , as is with all methods that are based on distributions, is the estimation of parameters for the marginals. There are many ways to do this, and in the paper [1] two such methods are discussed, and will also be presented below. When estimating the parameters, we assume that we have a set of training data,

$$T = \mathbf{x}_{k,i}, i = 1, ..., n_k, k = 1, ..., K \tag{2.24}$$

in which $\mathbf{x}_{k,i}$ is the observation vector, where k denotes the class membership, and i denotes the observation, n_k is the number of observations from

each class k. Independence is assumed between classes, whilst dependence is assumed within the observation vectors, and is modelled through the Meta-Gaussian distribution. The last assumption is that none of the class-specific distributions share the same parameters.

With these assumptions in place, the parameters can be estimated separately for each of the classes. To simplify the notation, we consider the case of one class, and can suppress the class index k. We can then write our observation vector, $\mathbf{x}_{k,i}, i = 1, ..., n_k, k = 1, ..., K$ as $\mathbf{x}_i, i = 1, ..., n_k$.
Below are the two methods suggested for the estimation of parameters, the common maximum likelihood methods, and the simpler Estimating Equations (EE) method.

Estimation using the method of Maximum Likelihood

Maximum Likelihood estimation of the unknown parameters γ and $\mathbf{C}$, still assuming independence between observation vectors $\mathbf{x}_i$ conditional on class information. The log-likelihood function for data within a class is then

$$l(\gamma, \mathbf{C}) = \sum_{i=1}^{n} \log(f(\mathbf{x_i}; \gamma, \mathbf{C})) \tag{2.25}$$

Where f is our Meta Gaussian distribution function, given in equation 2.22, and n is the number of samples in the class. Writing this out gives us

$$l(\gamma, \mathbf{C}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{n}{2} tr[(\mathbf{C}^{-1} - \mathbf{I})\mathbf{S}(\gamma)] + \sum_{i=1}^{n} \sum_{j=1}^{p} \log(g_j(x_{i,j}; \gamma_j)) \tag{2.26}$$

Where

$$S(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}(\mathbf{x}_i; \gamma)\mathbf{y}(\mathbf{x}_i; \gamma)^T \tag{2.27}$$

We wish to maximize this log-likelihood function, $l(\gamma, \mathbf{C})$, to obtain our ML estimates for γ and $\mathbf{C}$. Taking into account the constraints that exist for $\mathbf{C}$, which is a corrolation matrix, we have that the diagonal elements must be equal to 1, the off-diagonal elements must be between 1 and -1, inclusive, it must be symmetric, and it must be positive semi-definite. [29] Due to the constraints on our correlation matrix $\mathbf{C}$, and the constraints that may exist for γ depending on the marginals that are used, a direct optimization can be difficult. Storvik et al therefore proposed to rewrite $\mathbf{C}$ in a way that allows for a simplification of the constraints that are posed for $\mathbf{C}$ [1]. Given that $\mathbf{C}$ is a correlation matrix, we can write

$$\mathbf{C} = \mathbf{D}^{-1/2}\mathbf{M}\mathbf{D}^{-1/2} \tag{2.28}$$

where $\mathbf{M}$ is a positive definite symmetric matrix, and $\mathbf{D}$ is the diagonal matrix of $\mathbf{M}$. Using the properties of $\mathbf{M}$, we can express it as $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is a lower triangular matrix. This in turn means that $D_{jj} = M_{jj} = \sum_{r=1}^{j} L_{jr}^2$. And, in turn, for all $k \leq j$ we then have that

$$C_{jk} = \frac{\sum_{r=1}^{\min j,k} L_{jr} L_{kr}}{\sqrt{\sum_{r=1}^{j} L_{jr}^2}\sqrt{\sum_{r=1}^{k} L_{kr}^2}} \tag{2.29}$$

By inserting this into our equation for the log-likelihood function, we get

$$l_\gamma(\gamma, \mathbf{C}) = \frac{n}{2}\log|\mathbf{D}| - \frac{n}{2}tr[|\mathbf{L}\mathbf{L}^T|^{-1}\mathbf{D}^{1/2}\mathbf{S}(\gamma)\mathbf{D}^{1/2}] + \frac{n}{2}tr[\mathbf{S}(\gamma)] + \sum_{i=1}^{n}\sum_{j=1}^{p}\log(g_j(x_{i,j};\gamma_j)) \tag{2.30}$$

Which we can then use to optimize using a general optimizer.

Estimating Equations

The EE method can be said to be a variant of Maximum Likelihood (ML) classification, or, ML can be said to be a variant of the EE method depending on which way you look at it. It is essentially a more generalized method of estimation than the ML method. The estimating equations is a statistical metod that is used to fit model parameters to existing data. This is done by solving a set of simultaneous equations consisting of the sample data and the unknown parameters [29].

In our case, this is done through the following steps: Assuming that each marginal density has unknown parameters γ_j and denoting $\gamma = (\gamma_1, ..., \gamma_p)$, where p is the number of different image layers, or channels that are to be merged. For now we assume that the marginal is a parametric probability density function that needs to be specified beforehand, e.g. the Gamma PDF. The goal is then to estimate γ and the correlation matrix $\mathbf{C}$, using the three steps suggested in [1].

1. Estimate γ as the ML estimates, assuming independence between $x_{i,1}, .., x_{i,p}$
2. Transform each component of the vector $\mathbf{x}_i$ to $\mathbf{y}_i$ for $i = 1, ..., p$ using the mapping function
$$y_j = y_j(x_j; \gamma_j) = \Phi^{-1}(G_j(x_j; \gamma_j)) \tag{2.31}$$
3. Estimate $\mathbf{C}$ as the sample correlation matrix for $\mathbf{y}_1, ..., \mathbf{y}_p$.

2.3 Probability Distributions in Remote Sensing

Probability distributions in general are used to describe the world around us, and most of them arise from either a need to describe phenomena, or as a means to do so. If we assume fully developed speckle in a SAR band, and represent the complex variables by $Z = X + iY$, Z will follow a complex Gaussian distribution. Its amplitudes, $A = \sqrt{X^2 + Y^2}$ will be Rayleigh distributed [30], and the intensity $I = A^2$, will have the exponential distribution. If the intensities I are multi-looked, such as in an MLC image, they will no longer have the exponential distribution, but rather the Rayleigh. Optical images can often be assumed to follow a Gaussian distribution [31], or a non-central Gamma. The multilooked complex covariance matrix of a sar band, $\mathbf{C}$, has been shown to have a complex Wishart distribution [32,33].

These are a few theoretical distributions of different variables used in remote sensing, and many more exist. Accurately modelling of the marginal probability distribution functions is an important part when using the Meta-Gaussian classification scheme, and having some knowledge about which distribution to expect from data can make the initialization of the scheme easier, and may give better results.

2.3.1 Possible choices of the marginals

In general there are three possible choices when it comes to the marginals.

- Individually specified marginal probability densities. This means that we specify the PDF for each marginal. Either through some pre-existing knowledge about theoretical probability distribution of the data, or through a pre-classification test step. The test step could in that case be through estimation of different PDFs and choosing the one with the best data fit.
- A generalized probability density assumed to be adjustable for all possible marginals. Instead of having individually specified marginals, forcing the same on all makes for a more versatile classifier.
- Kernel probability density approximation for each of the marginals.

For the second case, having one pre-fixed distribution saves time in the pre-training part, but it may turn out to not fit all the underlying distributions as well as an individually specified probability density function would have. For the third case, a kernel estimation of the underlying probability density function will always fit the data as good as the choice of kernel and window size allows, but, at the same time, it may be more sensitive to outliers than the other two suggestions, which would in turn affect the transformation, and thus the end classification results. The second drawback is that it requires more time in the classification stage than the other two. Instead of simply saving the parameters from a known parametric distribution in the modelling stage, the entire marginal model has to be saved for each of the $(p \times K)$ number of classes.

So, to summarize, the hypothesis is that the case with individual marginals should give the most accurate results, given the right choice of marginals. This may also be a very suited approach when dealing with data where we have pre-knowledge, i.e. we already know which parametric distributions

the classes adhere to. If we do not have this, and need to do a test to find the distribution with the best fit, it may quickly turn out to be more time consuming than the two others. The second case, using a general parametric distribution, allows for a possible unsupervised extension of the method. It would then be a matter of finding and selecting a parametric distribution that can handle any and all possible underlying distributions. The third case, with the kernel density approximation of a non-parametric distribution, seems initially to perhaps be the best choice, but only in the supervised case.

Through a series of tests, using both simulated and real datasets, we will determine which of these methods give the best results, and whether there are any major differences in the classification results. A small decline in the performance of a classifier may be acceptable if we arrive at a more general procedure. This will be done initially for the supervised case, and then expanded to include the unsupervised case. It is not expected to arrive at the same conclusion for the unsupervised as for the supervised. Below follows a listing of some distributions that have been tested.

Beta Distribution

The Beta distribution is a continuous probability distribution that are generally characterized to be non-zero outside the $[0, 1]$ interval. This should exclude it from being considered as a choice for a general parametric density function. However, there exists a general Beta distribution, in which the valid range for x is not limited to $[0, 1]$ The Generalized Beta (GB) distribution was proposed in [34] , and is defined by the PDF

$$GB(y; a, b, c, p, q) = \frac{|a|y^{ap-1}(1-(1-c)(y/b)^a)^{q-1}}{b^{ap}B(p,q)(1+c(y/b)^a)^{p+q}} \text{ for } 0 < y^a < b^a/(1-c), \tag{2.32}$$

and zero otherwise. The parameters b, p, q are positive, $0 \leq c \leq 1$, a is limited by the constraints posed by $0 < y^a < b^a/(1-c)$. $B(p, q)$ is the beta function, given by

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt \tag{2.33}$$

Extreme Value Distribution

The extreme value distribution has a probability density function given by:

$$y = f(x|\mu, \sigma) = \sigma^{-1}\exp(\frac{x-\mu}{\sigma})\exp(-\exp(\frac{x-\mu}{\sigma})) \tag{2.34}$$

Normal Distribution

The normal distribution is perhaps the most famous probability distributions Its probability density function is given by

$$y = f(x|\mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.35}$$

And its cumulative density function by:

$$\frac{1}{2}[1 + \text{erf}(\frac{x-\mu}{\sigma\sqrt{2}})] \tag{2.36}$$

where erf$(\cdot)$ is the error function given by

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt \tag{2.37}$$

Gamma Distribution

The Gamma PDF is given by:

$$y = f(x|a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b} \tag{2.38}$$

where $\Gamma(\cdot)$ is the Gamma function, given by

$$\Gamma(a) = (a-1)! \tag{2.39}$$

when a is a positive integer. For complex numbers with a positive real part, it is defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \tag{2.40}$$

t Location-Scale Distribution

The t location-scale distribution is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[\frac{\nu + (\frac{x-\mu}{\sigma})^2}{\nu}\right]^{-(\frac{\nu+1}{2})} \tag{2.41}$$

where $\Gamma(\cdot)$ is the Gamma function, μ the location parameter, σ the scale parameter and ν the shape paramter. [35]. It has heavier tails, determined by ν than the normal distribution, but will approach it when $\nu \to \infty$.

Mixture models in Data

To fully understand many of the concepts that are being introduced, a certain insight into mixture models is required. Commonly written as:

$$p(\mathbf{x}) = \sum_{j=1}^{J} p(\mathbf{x}|j)P_j \tag{2.42}$$

where

$$\sum_{j=1}^{J} P_j = 1, \int_{\mathbf{x}} p(\mathbf{x}|j)d\mathbf{x} = 1 \tag{2.43}$$

Meaning that a combination of the J number of distributions, $p(\mathbf{x}|j)$, are required to form, or model $p(\mathbf{x})$. This allows us to model complex mixture distributions accurately, given the right parameters. [36]

2.3.2 Kernel Density Estimation

Usually there are two reasons for using a kernel density estimate of the underlying probability density function of a random variable. When a parametric distribution is not suited to describe the data, or when it is preferable to not make any assumptions about the underlying distribution, kernel estimates are well suited. Kernel density estimation, or Parzen-Rosenblatt window, after the two people who are acknowledge to have independently created the method [37, 38],is a non-parametric way of estimating the probability density of a random variable. If the data adheres to a known parametric model, using a kernel density approach to estimate a non-parametric model will not necessarily give better results, but will take significantly longer to compute. A simple approach to understanding kernel density estimation, is to think of it as a function describing the shape of an histogram. A common representation for the kernel density estimator $\hat{f}_\nu$ is given by

$$\hat{f}_\nu(x) = \frac{1}{n\nu} \sum_{i=1}^{n} K(\frac{x - x_i}{\nu}); -\infty < x < \infty \tag{2.44}$$

where $K(\cdot)$ is the kernel function, n is the number of samples in the data, and ν is the bandwidth of the kernel function. The bandwidth ν dictates the level of smoothing over the kernels. A high value for ν will give a high level of smoothing, whereas a small value for ν will adapt more to the fluctuations in the samples. The choice of the kernel will have an effect on the result. Popular kernels includes:

- Gaussian, using a standard normal distribution

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{2.45}$$

- Epanechnikov- using

$$K(x) = \frac{3}{4}(1 - x^2) \tag{2.46}$$

- Box-or uniform, using

$$K(x) = \frac{1}{2} \tag{2.47}$$

- Triangle- given by

$$K(x) = (1 - |x|) \tag{2.48}$$

The effects of these different kernels are shown in figure 2.1.

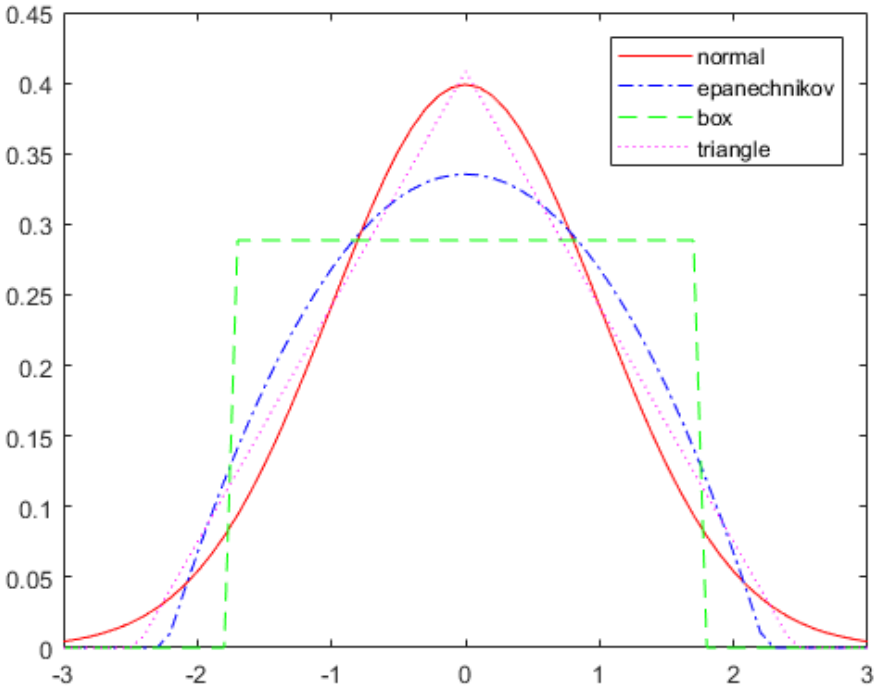


Figure 2.1: Effects of different kernels [39]

Chapter 3

Remote Sensing Data

As briefly mentioned in section 1.4, there exists a variety of remote sensing data. In this thesis, we will restrict us to using two of these, namely synthetic aperture radar(SAR) data, and multispectral data obtained from optical sensors. The main reasons for choosing these two are their availability, applicability, and complementarity. A SAR and an optical multispectral system operate very differently. Whilst a SAR will actively transmit an EM-wave, or pulse, and measure the backscattered signal, an optical system works passively, by measuring the intensity within certain bands of the electromagnetic spectrum, typically within the visible and infrared range, see figure 3.1.

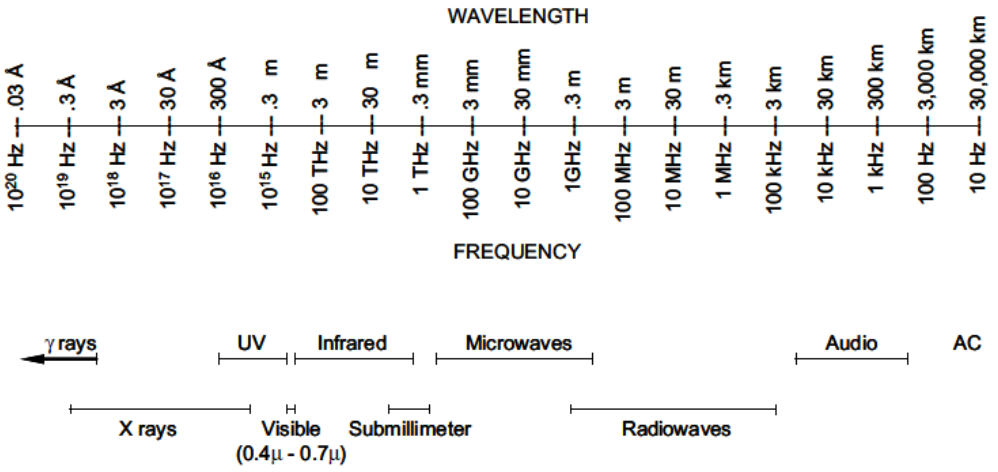


Figure 3.1: The electromagnetic spectrum, [40]

Table 3.1: Common Bands for Microwave Remote Sensing

Name	Width from(cm)	Width to(cm)
Ka	0.75	1.10
K	1.10	1.67
Ku	1.67	2.40
X	2.40	3.75
C	3.75	7.50
S	7.50	15.00
L	15.00	30.00
P	30.00	100.00

3.1 SAR Image Acqusition

The principle of SAR, is to simulate an antenna that has a much larger aperture size than what is physically possible. For a real aperture radar system, the size of the smallest object that can be detected is determined by the distance between the object and the radar, R, the wavelength λ and the diameter of the antenna, L_s through the following relation [41].

$$r_{azimuth} = \frac{R\lambda}{L_s} \tag{3.1}$$

If we want to find L_s for a representative example, such as what it would be in order to achieve the same resolution as the SENTINEL-1 C-band SAR, with a wavelength of 5.5cm, orbiting at 693km, and we take $r_{azimuth}$ to be 5metre,

$$L_s = \frac{\lambda R}{r_{azimuth}} L_s = \frac{0.0555\text{m} \cdot 693000\text{m}}{5\text{m}} \tag{3.2}$$

We would end up with an antenna diameter of 7692.3m, or, almost 8 kilometres. This would be a difficult structure to build down at Earth, let alone send it into space. This leads us on to the principles of SAR. It begins with the Doppler effect.

First suggested by Christian A. Doppler in 1842 who found that a constant sound had a higher pitch when moving towards the observer, and a lower pitch when moving away. Formulated mathematically, this translates to:

$$f = \frac{c + v_r}{c + v_s} f_0 \tag{3.3}$$

Where f is the observed frequency, f_0 is the emitted frequency, c is the velocity of the wave in the medium, v_r is the relative velocity of the receiver to

the medium, and v_s is the relative velocity of the source to the medium. v_s is negative when the source is moving away from the receiver, and positive if it's moving towards the receiver. v_r is negative if the receiver is moving away from the source, and positive if it's moving towards the source. This effect is found to be true for all waves. In SAR, this means that since we know the transmitted frequency, and we are able to observe the reflected frequency, and the time between transmission and detection, we can also know the relative movement that has taken place. This allows us to determine the position of the scattering object.

The resolution in ground range is not affected by the size of the aperture, nor does it utilize the Doppler effect, but rather the duration of the signal. In that case, the ground range resolution is given by

$$r_{range} = \frac{c\tau_p}{\sin\theta} \tag{3.4}$$

Where c is the velocity of the propagated wave, θ the angle of incidence, and τ_p the pulse length. If we treat the medium the wave propagates in as vacuum, the velocity of a radar signal is given by the speed of light, $c = 3.0 \cdot 10^8 \mathrm{m\,s^{-1}}$. Further, considering that the expression $\sin\theta$ is limited in the practical range between $[0°, 90°]$, the only thing we have major control over is the pulse length. For the resolution in ground range to be sufficiently high, this means that τ_p would have to be somewhere in the range of 10^{-8} to 10^{-7}. The power held in such a short signal is not enough for the backscatter to be above the signal to noise ratio (SNR) for an orbital system. An alternate method was therefore proposed, using pulse-compressed signals, or chirps. The resolution in range direction, given a pulse-compressed signal, r_{range} is given by

$$r_{range} = \frac{c}{2B\sin\theta}, \tag{3.5}$$

where c is the velocity of the propagated wave, and θ the incidence angle of the aperture. B is the spectral bandwidth, and also the main influence of the range resolution. In essence, the higher bandwidth, the higher resolution.

The resolution in azimuth direction for a conventional radar system is given by:

$$r_{azimuth} = R\theta_H = \frac{R\lambda}{L_a}, \tag{3.6}$$

where R is the slant range, θ_H is the angular spread of the beam, which can also be expressed in terms of L_a, the aperture length; and λ, the wavelength. Since this expression is dependent on the distance from aperture to target,

which for satellites usually is in excess of 600 kilometres, the azimuth resolution obtained is not sufficient for any conventional application. And this is where the equation for frequency shift through the Doppler effect is useful. Recall

$$f_{recived} = \frac{1 - \frac{v}{c}}{1 + \frac{v}{c}} f_{transmitted} \tag{3.7}$$

when the satellite moves away from the target, and

$$f_{recived} = \frac{1 + \frac{v}{c}}{1 - \frac{v}{c}} f_{transmitted} \tag{3.8}$$

when the satellite moves towards the target. For satellites, we can assume that $c >> v$, and we can simplify these expressions, thus:

$$f_{recived} = 1 - \frac{v}{c} f_{transmitted}, \tag{3.9}$$

when the satellite moves away from the target, and

$$f_{recived} = 1 + \frac{v}{c} f_{transmitted} \tag{3.10}$$

when the satellite moves towards the target. The frequency difference that we observe , $\Delta f = f_{recived} - f_{transmitted}$ is in terms of the backscattered signal. Assuming constant relative velocity throughout, this would then be

$$\Delta f = 2\frac{v}{c} f_{transmitted} \tag{3.11}$$

$$r_{azimuth} = \frac{\lambda R}{2v_{rel}} \delta f_d \tag{3.12}$$

δf_d is the

The resolution in azimuth direction is given by:

$$r_{azimuth} = \frac{h\lambda}{L_a} \tag{3.13}$$

Since we know the transmitted frequency, and the relative speed of the satellite to the ground, we can use the frequency shift, and the return time of the received signal to determine its location.

3.1.1 Distortions

Due to the nature of which a SAR image is acquired, preprocessing is a necessary step of practically any SAR image analysis. The raw data downloaded from the satellite is incomprehensible as it is, and needs to go through a long chain of processing steps before it can be used. For a normal user, most of these steps have already been performed by the data provider, and I will therefore not go into much detail regarding this, but some noteworthy artefacts that are typical in SAR images are important to know of when analysing these images. Because SAR views at an angle, some geometric distortions can occur in steep or very rough terrain. Shown in figure 3.2 are examples of these effects. Foreshortening takes place in sloped areas, when the distance between the radar is the same across the slope. This causes distances between points on the slope to appear shorter in the SAR image than the true ground distance. Layover is an extreme case of foreshortening, in which the top of a tall feature will be reflected before the base of the feature, and will therefore appear to be leaning over. Shadowing occurs when an area is blocked, or shadowed by another object.

Scattering Mechanisms in SAR

Scattering mechanisms are very important in polarimetric SAR, because they can tell us a lot about the observed areas. The scattering mechanisms present in an object will be dependent on the wavelength, and the look angle of the aperture. Objects of the same size as the wavelenght will give a strong return signal, which is part of the reason for C-band SAR being used for ocean monitoring, as the small capillary waves typical of open seas are of the 5 cm size. In figure 3.3 we see representative examples of scattering mechanisms in SAR. In a), we have reflection off a smooth surface. The angle of reflection is equal to the incident angle, and no backscatter is received at the aperture. This is often seen in calm, flat water, which will then appear black on the SAR image. in b), we have scattering off a rough surface, is here an arbitrary word, as the measure of roughness will be related to the wavelength of the signal. Such surfaces will usually give a good return signal. In c) and d), we have double bounce scattering, this is typical of urban areas, or any man made structures with sharp angles, but may also be found in forests and such, where the signal bounces off tree stems. This usually gives a strong return signal. Large ships on the ocean will typically show up as a bright dot, in an otherwise dark image. Volume scattering is usually due to the dielectric properties of the material, which causes absorbtion and re-emission, of the signal, or in inhomogeneous objects, such as tree canopies [40].

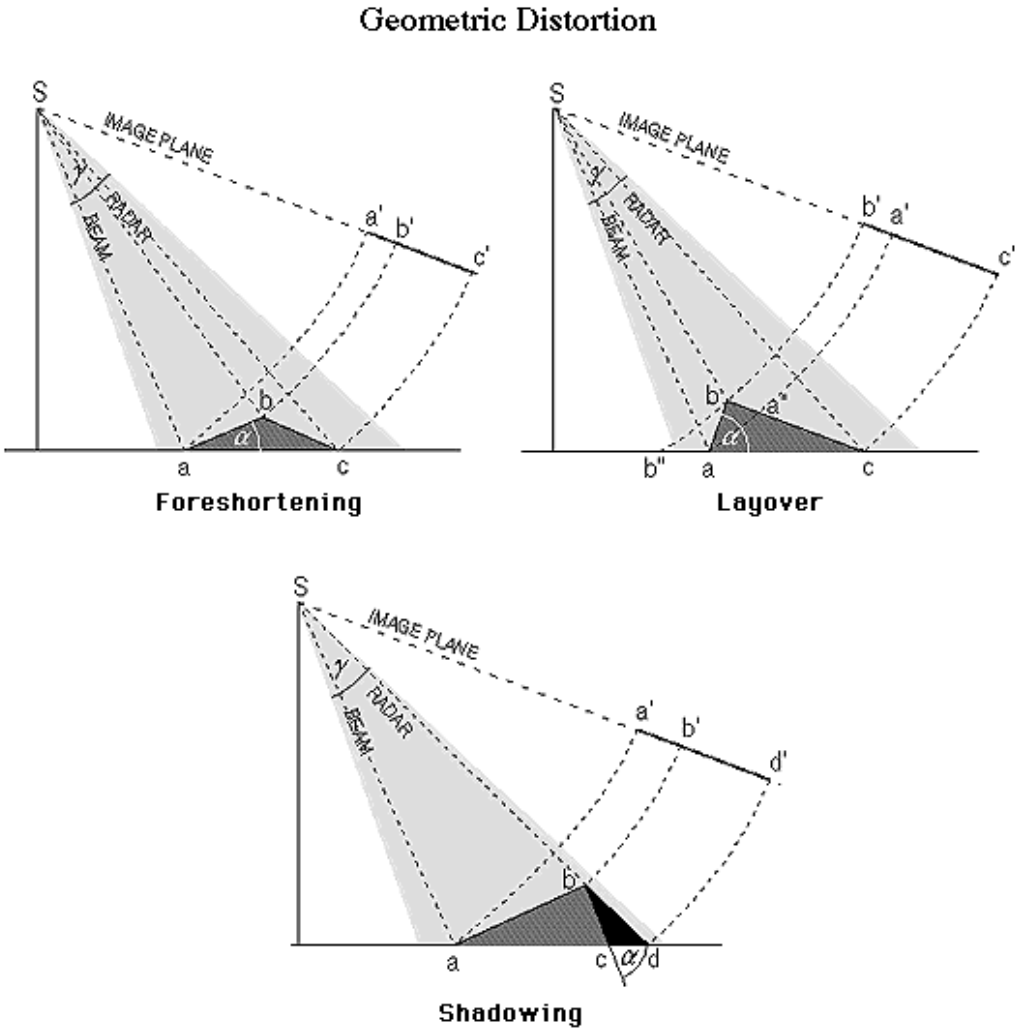


Figure 3.2: Terrain effects on radar images, Image credit:ESA [42]

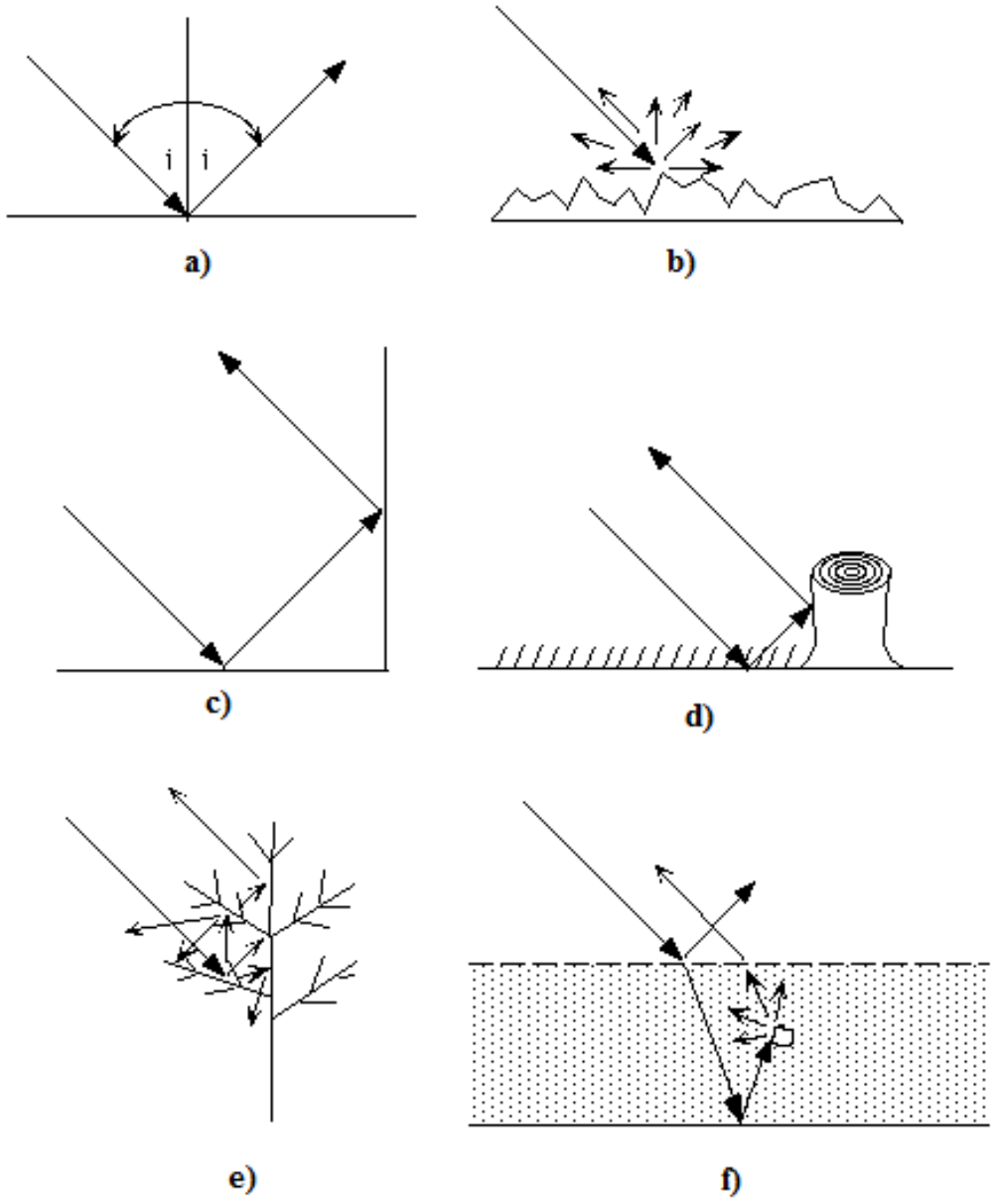


Figure 3.3: Various scattering mechanisms: a) Reflection on a smooth surface, b) Scattering off a rough surface, c) & d) Double bounce scattering, e) & f) Volume scattering. Image credit: ESA [42]

Speckle

Speckle is an important factor in SAR imagery. Visually, speckle in SAR data is a seemingly random effect, that presents itself as black and white pixels in the image. Sometimes wrongly dubbed speckle noise, or salt and pepper noise, speckle is caused in the formation of the image itself. Speckle is caused by different scattering mechanisms that may exist within a resolution cell. The multiple scattering waves that return to the sensor to form the pixel, will then interfere with each other. When waves have a displacement in the same direction, we will get a constructive interference, where the amplitudes of each respective wave, would add up, forming a new wave, with a higher amplitude. The opposite case, where waves have a displacement in opposite directions, creates a destructive interference, where the signal perceived is in fact lower than what should be. The extreme case of this occurs when two waves that share the exact same frequency and amplitude, but where the phases are shifted, a signal would effectively be cancelled out. A principle that is perhaps familiar in practice in the form of active noise cancelling headsets, where surrounding noise is analysed and then attempted cancelled using this simple principle. To reduce the appearance of speckle, a common approach is multi-looking- either as a part of the image generation itself- by splitting the radar beam into several beam-segments when acquiring the image, allowing for several independent "looks" of the area that is being targeted, and then averaging these independent looks. Multi-looking can also be done after the image has been generated, through a local averaging filter. There are many other filtering techniques that can be used to reduce the appearance of speckle, such as the Lee and the Frost filter [43].

3.1.2 Polarimetry

In recent years, SAR systems have been equipped to receive and transmit in different polarizations. Conventionally, a full-polarimetric SAR system is able to both transmit and receive linear horizontal(H) and vertical(V) polarised waves. This enables it to measure a total of four combinations in terms of a transmitted and received electromagnetic field,respectively HH, HV, VH, VV. To fully understand this concept, we need to understand basic principles of electromagnetism.

Electromagnetic waves propagate through space, and can be sufficiently described through a two dimensional complex vector. If this wave is incident on an object that can be considered to be a scatterer, we can also observe the scattered wave and describe this using a two-dimensional complex vector. We

can describe this relationship between the incident wave and the scattered wave through the following relation

$$\begin{bmatrix} E_h^s \\ E_v^s \end{bmatrix} = \frac{e^{-jkR}}{R} \begin{bmatrix} S_{hh} & S_{hv} \\ S_{vh} & S_{vv} \end{bmatrix} \begin{bmatrix} E_h^i \\ E_v^i \end{bmatrix} = \frac{e^{-jkR}}{R} [S] \begin{bmatrix} E_h^i \\ E_v^i \end{bmatrix} \tag{3.14}$$

Here, E denotes the electric field, with subscript h or v relating to the polarization (horizontal or vertical), and superscript i or s referring to an incident wave or a scattered wave, respectively. k is the wavenumber, and R denotes the radial distance between the antenna and the scatterer. The $\frac{e^{-jkR}}{R}$ term ensures that the propagation effects for amplitude and phase are included. The scattering coefficients S_{ij} are subscripted with the incident and the scattered wave, respectively. $[S]$ denotes the scattering matrix that contains the scattering coefficients S_{ij}. Fully polarimetric SAR data typically has a more narrow field than dual-polarization and single-polarization systems, and it also enables us to differentiate between the scattering mechanisms present in the scatterer.

Each pixel in a full-polarimetric SAR image is represented by a scattering matrix S. In the mono static case, where a single antenna is used for both transimitting and reciving the electromagnetic field, and if we can assume reciprocity we have that $S_{hv} = S_{vh}$, and we find the lexicographic scattering vector $\underline{\Omega}$ given by

$$\underline{\Omega} = [S_{hh}, \sqrt{2}S_{hv}, S_{vv}]^T \tag{3.15}$$

S and $\underline{\Omega}$ can be called single-look complex (SLC) representations of the scatterer. To reduce speckle in the images, a common approach is that of multilooking. The multilook complex (MLC) covariance image is given by

$$\boldsymbol{C} = \frac{1}{L} \sum_{i=1}^{L} \boldsymbol{\Omega_i} \boldsymbol{\Omega_i^H} \tag{3.16}$$

Where L is the number of looks that is used in the averaging, and $(\cdot)^H$ denotes the Hermetian transpose. This is equivalent of applying a local moving average filter. In doing this, we have a loss of spatial resolution that is proportional to the number of looks, L, but we gain a clearer image with less speckle.

Radar equation

The radar equation is a compact representation of the power received at the antenna, or the backscatter, in terms of the transmitted power. It is given

by

$$P_r = P_c A \tag{3.17}$$

Where P_r is the power received at the antenna, and A is the area of the antenna. P_c is the scattered power density, given by:

$$P_c = \frac{P_s}{4\pi R_2^2} \tag{3.18}$$

Where R_2 is the distance from the scattering object to the receiving antenna, and P_s is the scattered power at the object,

$$P_s = P_i s \sigma_0 \tag{3.19}$$

Where s is the size of the scattering object, σ_0 is the normalized backscattering cross section, defined as:

$$\sigma_0 = \lim_{R \to \infty} \frac{4\pi R^2}{A_0} \frac{|E_r|^2}{|E_i|^2} \tag{3.20}$$

Where A_0 is the illuminated surface area, E_r and E_i are respectively the reflected and the incident electric field. R is the distance from the antenna to the target. P_i is the incident power density,

$$P_i = \frac{P_t G_t}{4\pi R_1^2} \tag{3.21}$$

Where in turn R_1 is the distance from the transmitting antenna to the target, P_t is the power transmitted, and G_t is the antenna gain.

All this gives us the full radar equation for real aperture radar:

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^3 R^4} \frac{\lambda R}{L} \frac{c\tau_p}{2 \sin \theta_i} \sigma_0 \tag{3.22}$$

Where c is the speed of light, λ is the wavelength of the radiated signal, L is the length of the antenna, θ is the viewing angle, and τ_p is the pulse length.

3.2 Optical Image Acquisition

Optical satellite imagery goes back to August 14, 1959, when NASA's Explorer 6 took the first satellite picture of the Earth, shown in figure 3.4. Since then, a multitude of satellites have been launched, with improving spatial, temporal and radiometric resolution. Optical multispectral satellite sensors

Figure 3.4: First satellite image of the Earth, taken by the Explorer 6, image credit:NASA

such as that employed on the Quickbird-2, can currently image with a spatial resolution of 61 centimeters from an altitude of 450 kilometers [44].

The multispectral sensors such as employed by Landsat, etc, are generally passive sensors. This means that unlike the active SAR instruments, which emits electromagnetic waves, and measure the backscattered signal, they only measure the intensity of the electromagnetic waves that are naturally transmitted from the viewed object. And every object on Earth, or more specifically, every object with a temperature above absolute zero, will have some measurable transmission. For perfect blackbodies, this is expressed through Planck's radiation law. In most cases however, the temperature of an object is too low for the energy to be detected in space, and an external energy source is required- the sun.
Solar radiation is an incredible source of energy, and its interaction with elements or areas on Earth, allows us to image them by detecting the reflected or transmitted energy. Multispectral imaging sensors typically work by detecting the reflected or emitted energy within certain specified bands of the electromagnetic spectrum. Optical data has a wide range of applications, and is actively used to monitor sea surface temperatures, forest fires and land cover.

Chapter 4

Pattern Recognition

Pattern Recognition can be described as a field of science, where the aim is to classify or recognize, patterns or regularities in data. Allowing for items or objects to be categorized as belonging to a certain class. This process can either be done through a supervised classification, where we have, and utilize, some a priori knowledge of the data, or unsupervised classification, where we only consider the data itself.

In the following sections, the two distinct ways of performing classification are described, as well as the specific methods used in this thesis.

4.1 Supervised classification

In supervised classification, we have a priori knowledge about the data. This means that classification of unknown objects is done based on information we already have. Depending on the specific classifier, this information can be a number of things. One of the most common approaches in supervised classification is to use a training set, consisting of a set of data points in which the true class membership is known. These can then be used to train the classifier, or used directly in the classification, which is done in the case of the Nearest Neighbours classification rule.

Here, given N training points with known memberships, and an unknown point $\boldsymbol{x}$, $\boldsymbol{x}$ is assigned to the class that has the majority of the k closest points stemming from the training set [45]. More sophisticated methods, such as Neural Networks can use the training set to effectively learn to distinguish between different classes, through an iterative series of cost-function optimization. When using a supervised classifier based on Bayes desicion theory,

given by

$$P(\omega_i|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)P(\omega_i) \tag{4.1}$$

where

- $P(\omega_i)$ is the probability of class ω_i
- $P(\omega_i|\boldsymbol{x})$ is the probability of having class ω_i given the feature vector $\boldsymbol{x}$
- $p(\boldsymbol{x})$ is the probability distribution function of $\boldsymbol{x}$
- $p(\boldsymbol{x}|\omega_i)$ is the probability function of $\boldsymbol{x}$ given the class ω_i

If some knowledge is held about the PDFs and the class probabilities, these can be used directly, with no need for a training set. Or a training set can be used to estimate probability distribution functions for the different classes, and the class probability $P(\omega_i)$. This is the approach taken when implementing the supervised classification using Meta-Gaussian distributions, and is explained more thoroughly in Chapter 6 - "Implementation". Common for all supervised classification methods is that they will only segment the data into the classes that they are provided with. Any other underlying classes will not be considered. In addition to this, "poor supervision" will give corresponding results, and in many cases an unsupervised classification will be more suited. The choice of training data is an art in itself, and should be given as much thought as when selecting a classification algorithm. When performing any kind of supervised classification using a training set, a common approach is to omit parts of it from the training step, and use it to check the accuracy of the classifier in a later step.

4.2 Unsupervised classification

When the area of study is unknown, and when multiple classes is assumed to be contained within the study area, clustering is a method of separating the data into different classes, based on the available data. Due to the nature of the Meta-Gaussian approach to classification, there is a limited number of classification methods that can be used. For the transformation of the points into the Meta-Gaussian domain to be of any use, classification must take place in the Meta-Gaussian space. And in turn for the Meta-Gaussian transform to be accurate, it is vital that the marginal transformation is accurate. An unsupervised extension of the Meta-Gaussian classification method will consist of initial estimates of the marginals, The results of a clustering will not necessarily tell us what a particular class corresponds to. Consider

a simple example, using pictures of the numbers zero and one. When performing supervised classification we already know that class ω_0 corresponds to zeros, and ω_1 corresponds to ones. When classifying in an unsupervised setting, we get the same results, but, we do not know which class corresponds to the images of zeros, and which class corresponds to images of ones. In such a simple case, this is easy to verify manually, and you may find that class ω_0 corresponds to zeros, and class ω_1 to ones. The next time you run the same classification, this assignment could change, and you would need to recheck. Again, in this case it is merely bothersome, but for more complex problems, it could become challenging not only to compare different clustering results, but also to attach a meaningful description of what the classes represent. In many situations, we already have some expectations

Partly due to this, it is often more suitable to perform any type of classification on derived features which can be related to physical properties. These may not only be more descriptive in the clustering process itself, but can be a major asset when interpreting the clustering results.

4.2.1 Hard Clustering

Hard clustering is a sub-category of clustering in which a point is said to be belonging to only one class. No point is allowed to be assigned to more than one class at a time, and all points need to be included into a class. For pixels with large pixel sizes, this can be a problem in several ways. The first issue is that if the pixel is heterogeneous , it may be categorized as a hybrid between the true underlying classes. The second issue is that, for an heterogeneous pixel, it may be more beneficial to keep this information.

4.2.2 Soft Clustering

Soft clustering is the other sub-category of clustering. Here, a point may exist in several classes simultaneously, but the sum of the class membership must always be 1. A soft clustering can always be reduced to a hard clustering,

4.3 Expectation Maximization Approach

The Expectation Maximization (EM)- algorithm as known today, was introduced in 1977 by Arthur Dempster, Nan Laird and Donald Rubin in [46], although the concept had been around earlier, it was in their paper *"Maximum Likelihood from Incomplete Data via the EM Algorithm"* that an elaborate

and general procedure was introduced, along with the name. Although they claimed that it was only valid for exponential family distributions, this was later proved wrong [47], and it now stands as a versatile method with many applications.

When dealing with a problem in which the PDF of different classes are not linearly separable, a popular approach is that of the EM algorithm. Through successive iterations, it attempts to find the best division between classes. The EM-algorithm is designed to maximize the expectation of the log-likelihood function, conditioned on the observed samples and the current iteration estimate of $\boldsymbol{\theta}$. In the first step, the E-step, the estimation of the expectation of the log-likelihood function takes place.

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}(t)) \equiv E\left[\sum_k \ln(p_{\boldsymbol{y}}(\boldsymbol{y}_k; \boldsymbol{\theta}|X; \boldsymbol{\theta}(t)))\right] \tag{4.2}$$

In the second step, the M-step, estimates of θ is found through the maximization of the expectation of the log-likelihood function, $Q(\boldsymbol{\theta}; \boldsymbol{\theta t}))$

$$\boldsymbol{\theta}(t+1) : \frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{4.3}$$

For the first iteration, the unknown $\boldsymbol{\theta}$ needs to be initialized. This can either be done randomly, or, for a likely faster convergence, through a more sophisticated estimate. Different stopping criteria can be used. Typically, this is when the change between two subsequent iterations is below a predefined threshold. Change can for instance be measured over the parameters, or the clusterings. In many cases a maximum number of iterations allowed is also included in the stopping criteria. The EM-algorithm is not necessarily finding the global maximum. Depending on the stopping criteria, it may settle on a local maximum. A momentum term can be included in the stopping criteria, to try to avoid such local minima.

4.3.1 Generalized Mixture Decomposition Algorithmic Scheme

The Generalized Mixture Decomposition Algorithmic Scheme (GMDAS) is as the name indicates, an algorithmic scheme of the EM-algorithm that is used for unsupervised clustering. Assuming that the number of clusters, K is known, initial estimates for the unknown parameters $\boldsymbol{\theta}$ and the unknown class probabilities $\boldsymbol{P}$ are generated. Iteration indicator, t is set to zero. The

following is then repeated until a convergence criteria is reached. In this thesis, we used a change indicator

$$\delta = |\boldsymbol{\mu}_{old} - \boldsymbol{\mu}| + |\boldsymbol{\sigma}_{old} - \boldsymbol{\sigma}| + |\boldsymbol{C}_{old} - \boldsymbol{C}| \tag{4.4}$$

Where $\boldsymbol{C}$ are the correlation matrices, such that $\boldsymbol{C} = \boldsymbol{C}_1, \boldsymbol{C}_2 ... \boldsymbol{C}_K$, and $\boldsymbol{\mu} = [\mu_1, ..., \mu_p]^T$ and $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_p]^T$ are the parameters of the marginal estimates. The conditional probability of a class given the current estimates for the parameters Θ,

$$P(C_k|\boldsymbol{x_i}; \boldsymbol{\Theta}(t)) = \frac{p(\boldsymbol{x_i}|C_k; \boldsymbol{\theta_k}(t))P_k(t)}{\sum_{k=1}^{m} p(x_i|C_k, \boldsymbol{\theta_k}(t))P_k(t)} i = 1, .., N, k = 1, ..., K \tag{4.5}$$

$\boldsymbol{\theta_k}(t+1)$ is then found by solving the equation

$$\sum_{i=1}^{N} \sum_{k=1}^{K} P(C_k|\boldsymbol{x_i}; \boldsymbol{\Theta}(t)) \frac{\partial}{\partial \boldsymbol{\theta_k}} \ln p(\boldsymbol{x_i}|C_k; \boldsymbol{\theta_k}) = \mathbf{0} \tag{4.6}$$

with respect to $\boldsymbol{\theta_k}$ The class probabilities P_k are then updated through the following

$$P_k(t+1) = \frac{1}{N} \sum_{i=1}^{N} P(C_k|\boldsymbol{x_i}; \boldsymbol{\Theta}(t)), k = 1, ..., K \tag{4.7}$$

Class indicator t is set to $t+1$. Repeat previous steps. When convergence or maximum number of iteration is reached, the points $\boldsymbol{x_i}$ can then be classified using the Bayesian approach,

$$\hat{z}_i = \arg\max_k p(\boldsymbol{x_i}|C_k)P_k \tag{4.8}$$

Where $\hat{z}_i$ is a cluster indicator.

Chapter 5

Data and Features

5.1 Feature Selection

In classification using data obtained by satellite remote sensing, it is often helpful to use derived features rather than working on raw values. This is especially true when it comes to SAR data, where direct human interpretation is difficult due to the nature of the signal. The features of interest will often depend on the application, as well as the type of data that is available, although some are rather universal. Many features will have some physical interpretation, which allows for direct interpretation, such as the scattering mechanisms. For the scope of this thesis, discussion of features are limited to those that are of particular interest for the different real data that is used, which is forest classification and distinguishing ice.

As discussed previously, optical imagery has its limits when it comes to operational use in the Arctic, largely due to weather conditions and sunlight, or lack of thereof. But when it is available, it can yield valuable information about the surface properties. Colour for instance, which in many cases can be enough for a coarse segmentation. The normalised snow difference index (NDSI) [48] is an example of such a feature. Although it may not be as famous as it's vegetation equivalent, the NDVI, it builds upon the same principles:

$$NDSI = \frac{\rho_{VIS} - \rho_{SWIR}}{\rho_{VIS} + \rho_{SWIR}} \tag{5.1}$$

Where it takes advantage of the high reflective properties of snow in the visible part of the EM-spectrum, and its low reflectance in the shortwave infrared. This makes the NDSI well suited for distinguishing between snow and clouds, as clouds typically will have a high values in both regions. The NDSI on its own is not necessarily a robust measure of snow or no snow, and

other thresholds are usually included to make a more trustworthy assessment. Due to its inability to penetrate the surface, optical imagery has limitations, especially when it comes to the classification of sea ice. Although capable of distinguishing between sea ice and open water, it is not a valid measure of any internal structure or thickness. The Melt area detection index (MADI) is a simple method to determine surface melt. It uses the difference in the reflectance properties in dry and melting (wet) snow, to distinguish between them. It was originally developed for MODIS reflectance data, and is given by

$$MADI = \frac{R_{0.67}}{R_{2.1}} \tag{5.2}$$

Where $R_{0.67}$ is the reflectance values of MODIS band 1, with a wavelength of $620 - 670$ nm and $R_{2.1}$ is the reflectance values of MODIS band 7, with wavelengths between $2105 - 2155$ nm. The MADI is then used as a threshold value to segregate melting areas from areas without melting. This thesis uses different datasets, and each of them have their own set of applicable features.

5.2 Features derived from PolSAR

Recall the following representations of polarimetric SAR data, SLC and MLC data:
SLC vector data given by $\boldsymbol{s} = [S_{HH}, \sqrt{2}S_{HV}, S_{VV}]^T$, or, if we can assume reciprocity, $\boldsymbol{s} = [S_{HH}, S_{HV}, S_{VH}, S_{VV}]^T$.
And MLC matrix data given $\boldsymbol{C} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{s_i}\boldsymbol{s_i}^H$.
Below are some possible features derived from PolSAR [49] [50]. The somewhat constrained availability of full polarimetric data will typically exclude some of these from being used in many cases, such as in this thesis, where only dual-pol is available for ESAs SENTINEL 1 satellites . Nevertheless, they are mentioned here.

Mean radar backscatter

$$\boldsymbol{\mu} = \det(\boldsymbol{C})^{\frac{1}{d}} \tag{5.3}$$

Relative kurtosis

$$RK = \frac{1}{Ld(d+1)} \sum_{i=1}^{L} [\boldsymbol{s}_i^H \boldsymbol{C}^{-1} \boldsymbol{s}_i]^2 \tag{5.4}$$

Co-polarization ratio

$$R_{VV/HH} = \frac{\langle S_{VV} S_{VV}^* \rangle}{\langle S_{HH} S_{HH}^* \rangle} \tag{5.5}$$

Cross-polarization ratio

$$R_{HV/B} = \frac{\langle S_{HV} S_{HV}^* \rangle}{B} \tag{5.6}$$

Co-polarization correlation magnitude

$$|\rho| = |\frac{\langle S_{HH} S_{VV}^* \rangle}{\langle S_{HH} S_{HH}^* \rangle \langle S_{VV} S_{VV}^* \rangle} \tag{5.7}$$

Co-polarization correlation angle

$$\angle\rho = \angle(\langle S_{HH} S_{VV}^* \rangle) \tag{5.8}$$

5.3 Simulated bivariate Gaussian data

To test the various algorithms and variables, it is favourable to have a dataset in which the underlying distributions are known, to actually be able to estimate the performance of the classification. So, the various methods that are described in this thesis will be tested on both a simulated dataset, as well as on real data. To assess whether or not an unsupervised classification scheme based on the Meta-Gaussian distribution would work, we wanted to perform an initial test using simulated data. The data was simulated separately from four different bivariate Gaussian distributions, and concatenated to form the dataset. 100 bivariate samples were generated for each class, resulting in a test set of 400 two-dimensional vectors.

The classes were generated using the parameters
$\mu_1 = [0, 0]$, $\mu_2 = [8, 6]$, $\mu_3 = [20, 3]$, $\mu_4 = [4, 20]$ and

$$\sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}, \sigma_4 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.2 \end{bmatrix}$$

where $\mu_i, i = 1, ...4$ are the mean values of each class $\omega_1, ..., \omega_4$, and $\sigma_i, i = 1, ...4$ are the covariances. In figure 5.1, a scatterplot of the points is shown.

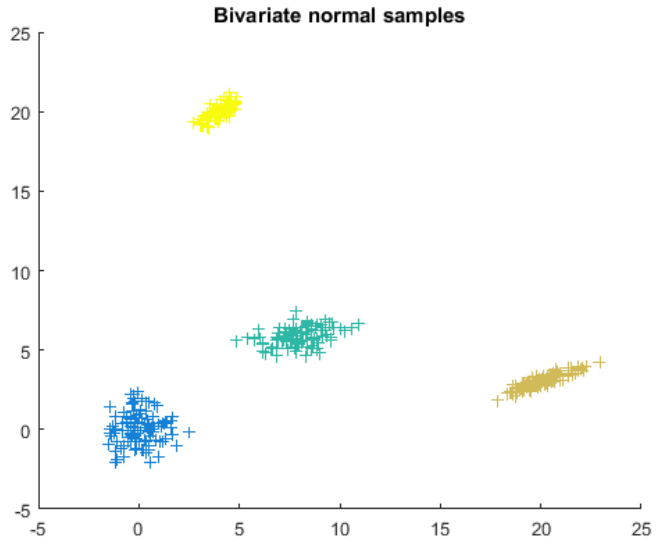


Figure 5.1: Scatterplot of simulated bivariate Gaussian samples. Different color indicates different class, blue is class ω_1, green ω_2, brown ω_3 and yellow ω_4

Band	Central Wavelength	Bandwidth	Spatial Resolution	Purpose
1	443 nm	20 nm	60 m	Aerosol detection
2	490 nm	65 nm	10 m	Blue
3	560 nm	35 nm	10 m	Green
4	665 nm	30 nm	10 m	Red
5	705 nm	15 nm	20 m	Vegetation classification
6	740 nm	15 nm	20 m	Vegetation classification
7	783 nm	20 nm	20 m	Vegetation classification
8	842 nm	115 nm	10 m	Near-infrared
8a	865 nm	20 nm	20 m	Vegetation classification
9	945 nm	20 nm	60 m	Water vapour
10	1380 nm	30 nm	60 m	Cirrus
11	1610 nm	90 nm	20 m	Snow/ice/cloud discrimination
12	2190 nm	180 nm	20 m	Snow/ice/cloud discrimination

Table 5.1: SENTINEL-2 Optical Bands, an overview of some important characteristics and description of the different bands.

5.4 SENTINEL SAR and Optical data

The European Space Agency(ESA) began its SENTINEL project in April 2014 with the launch of SENTINEL-1A, and a year later, an identical twin satellite, SENTINEL-1B followed. The two satellites forming this constellation operate 180° apart, providing global coverage. [51] These SAR satellites operate in C-band, and is available in dual-pol, HH/HV,VV/VH,HH,VV. In June 2015, the SENTINEL-2A optical satellite was launched, and two years later, in March 2017 its twin satellite SENTINEL-2B completed this constellation. The Sentinel 2 satellites also orbit 180° apart. The SENTINEL-2 satellites have a total of 13 bands in the visible and near-infrared(VNIR) and shortwave infrared (SWIR) domain of the electromagnetic spectrum. In table 5.1 an overview of the bands and their intended purpose is shown.

Two scenes were selected to be used in the analysis. SENTINEL-1 and SENTINEL-2 data from an area in Svalbard which contained land, open ocean and sea-ice. The SAR image from SENTINEL 1 was preprocessed in SNAP(SeNtinel Application Platform). The processing steps consisted of geocoding using the range-Doppler equations, and terrain correction. The optical image from SENTINEL-2 was resampled and geocoded. Finally, the two images were overlayed. There was no available ground truth data for the Svalbard scene, and analysis of segmentation results is done visually.

5.5 Nezer Forest SAR and Optical

The Nezer forest is an area in France, in which the planting of one single species of maritime pine has been performed in sections over many years. The data that is used from this area stem from the Landsat TM sensor of Landsat-4, as well as polarimetric SAR data from the NASA/JPL AIRSAR in C,L, and P-band. The AIRSAR data was aquired on August 16,1989 [52], and the Landsat-4 data on July 22, 1991.
The Landsat-TM sensor of Landsat-4, is a multispectral sensor, and operates in the visible and near infrared, a list of the seven bands is shown in 5.3. The data had already been preprocessed prior to my use, by Temesgen Gebrie Yitayew in his thesis *"Multi-sensor Data Fusion and Feature Extraction for Forest Applications"* [53], and further used in [54]. The scenes have been downsampled, and consists of 111×246 pixels, making up a total of 27306 in each image band. A ground truth map, shown in figure 5.3, is available for this dataset, and eight classes are identified in the scene, six of which correspond to maritime pine in different age groups, one is classified as bare soil, and the last is unknown data. The different classes, and their sample size are shown in table 5.2. The "unknown" class, with class-label 0, will mainly not be used in the classification. This is due to the undetermined nature of the points belonging to the class, which can consist of a multitude of different objects, and thus a very mixed distribution. Instead, when applicable it will be removed from the computations, and for graphical purposes, replaced when displaying the classified images. This reduces the number of available samples to 15683 per band.

Classification on this dataset will be performed using both the SAR data from the NASA/JPL AIRSAR, and the multispectral data from the Landsat 4 TM sensor. For the multispectral data, we will use the intensity values of the seven Landsat 4 TM bands, and the following six features derived from the TM bands:

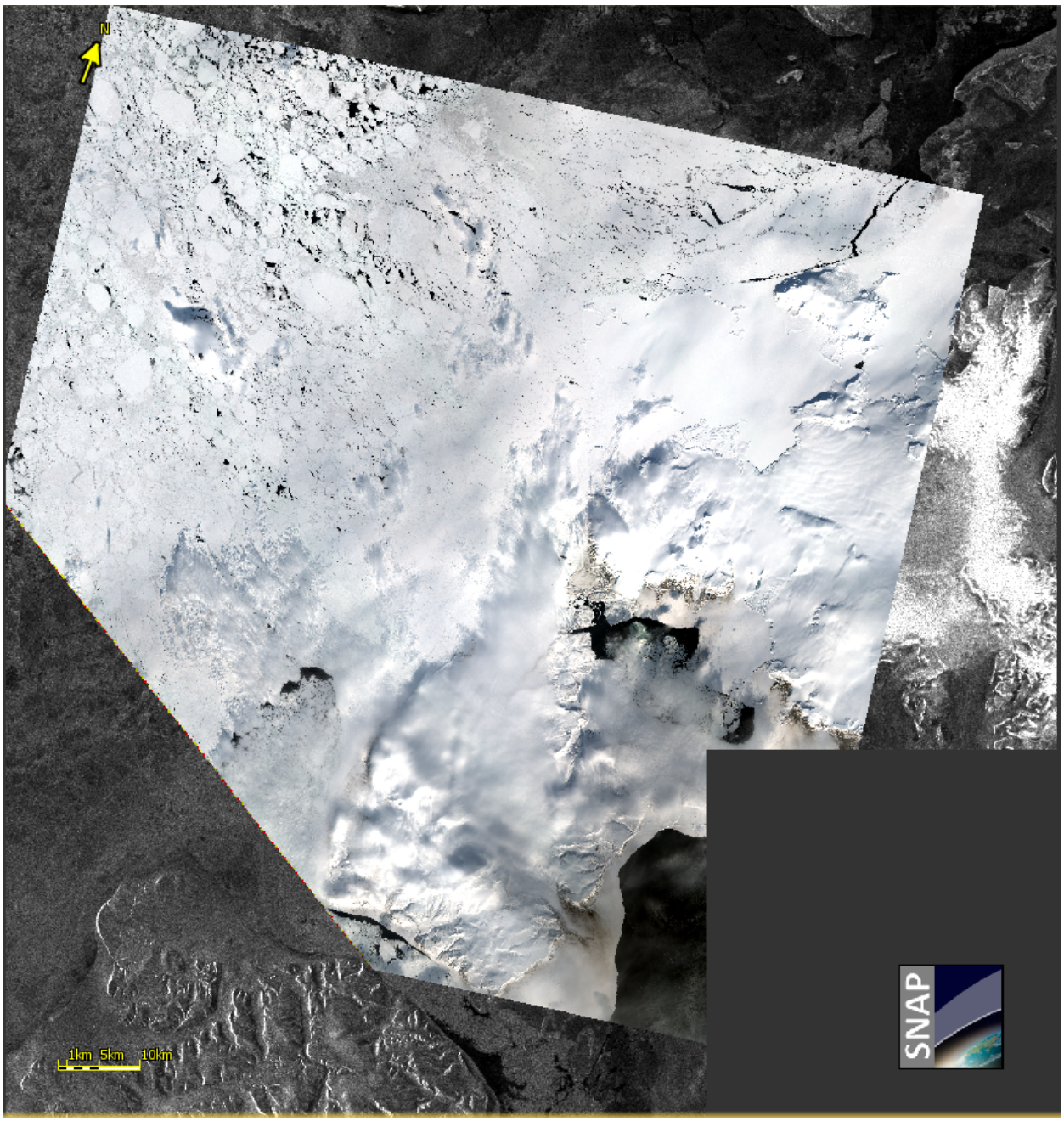


Figure 5.2: Amplitude of HH channel, truecolor RGB overlay.

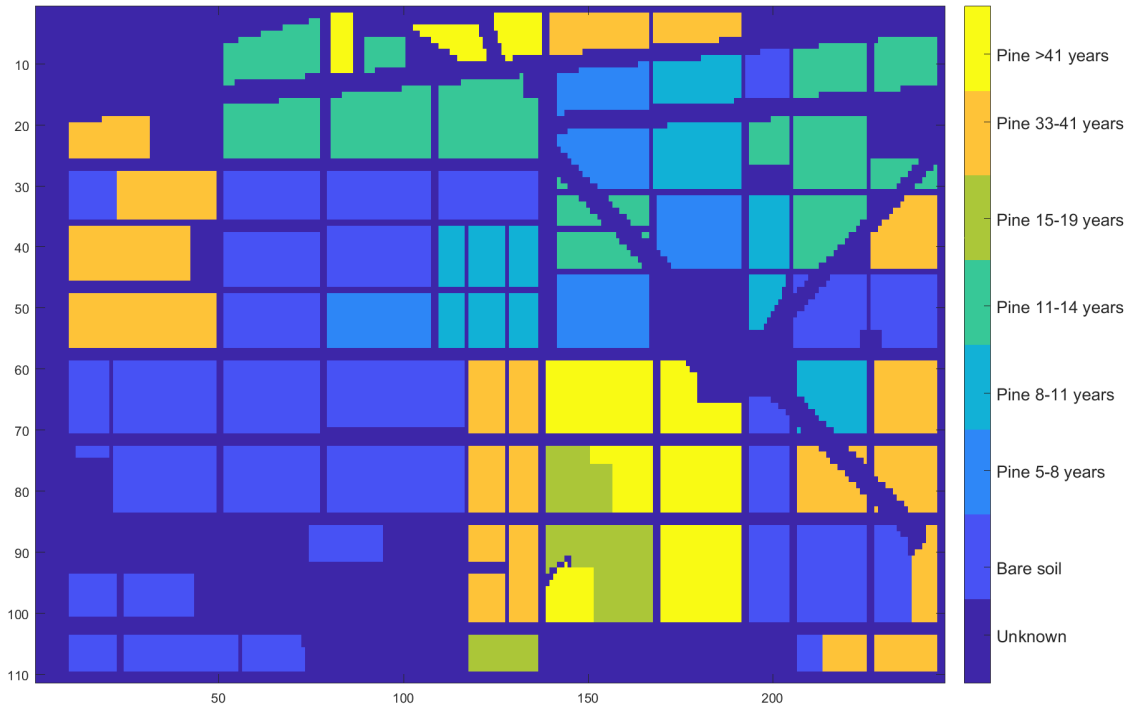


Figure 5.3: Ground truth map for Nezer forest data

Label	Class	Number of samples	Proportion
0	Unknown	11623	-
1	Bare soil	5765	36.76%
2	Maritime pine aged 5-8 years	1200	7.65%
3	Maritime pine aged 8-11 years	1307	8.33%
4	Maritime pine aged 11-14 years	2224	14.18%
5	Maritime pine aged 15-19 years	636	4.06%
6	Maritime pine aged 33-41 years	2958	18.86%
7	Maritime pine aged >41 years	1593	10.16%

Table 5.2: Classes and labels for Nezer forest data. Proportion in percent of the total classified area is taken not considering the "unknown" class

Band	Wavelength interval	Name
1	0.45-0.52 μ m	Blue
2	0.52-0.6 μ m	Green
3	0.63-0.69 μ m	Red
4	0.76-0.9 μ m	Near infrared (NIR)
5	1.55-1.75 μ m	Short wave infrared 1 (SWIR1)
6	10.4-11.5 μ m	Thermal infrared (TIR)
7	2.08-2.35 μ m	Short wave infrared 2 (SWIR2)

Table 5.3: Landsat 4 TM instrument bands, their bandwidth and common name.

- NDVI- given by
$$NDVI = \frac{B_4 - B_3}{B_4 + B_3} \tag{5.9}$$
- Brightness given by
$$B = \alpha_1 B_1 + \alpha_2 B_2 + \alpha_3 B_3 + \alpha_4 B_4 + \alpha_5 B_5 + \alpha_6 B_7 \tag{5.10}$$
- Greenness given by
$$G = \beta_1 B_1 + \beta_2 B_2 + \beta_3 B_3 + \beta_4 B_4 + \beta_5 B_5 + \beta_6 B_7 \tag{5.11}$$
- Wetness given by
$$W = \gamma_1 B_1 + \gamma_2 B_2 + \gamma_3 B_3 + \gamma_4 B_4 + \gamma_5 B_5 + \gamma_6 B_7 \tag{5.12}$$
- Perpendicular Vegetation index, PVI given by:
$$PVI = \sqrt{(0.355B_4 - 0.149B_2)^2 + (0.355B_2 - 0.852B_4)^2} \tag{5.13}$$

here $B_1, B_2, ..., B_7$ represents intensity values from band 1, band 2,..., band 7 of the TM sensor, as in table 5.3, and $\alpha_1, ...\alpha_6$, $\beta_1, ..., \beta_6$, $\gamma_1, ..., \gamma_6$ are constant coefficients, and can be found in [55]. For the NASA/JPL AIRSAR, the intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ corresponding to the elements $\mathbf{C}_{jj}, j = 1, 2, 3$ of the multilooked covariance matrices of the P-band, the L-band and the C-band were used. We also use the polarimetric features, described in section 5.2.

5.6 Data Transformation and Dimensionality Reduction

When working with any kind of high dimensionality data, there is always a chance of redundancy. In our case, with multispectral and multisensor data, correlation between different image bands is not uncommon, and reducing the number of bands can often be done without any real loss of information. The principle component analysis can be used both as a tool to investigate the redundancy in data, and also to reduce the number of dimensions whilst retaining the information. PCA works by applying a linear transform to our samples, $\boldsymbol{X}$ of dimensionality $[N \times p]$. N is here the number of samples, or observations, and p is the number of different observations, ie. the number of image bands.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}^T \tag{5.14}$$

Here, $\boldsymbol{Y}$ is our transformed data, $\boldsymbol{X}$ is our original data, which must have zero mean. If it does not, which is usually the case for real data, the sample mean is found and subtracted. $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = [\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{m-1}] \tag{5.15}$$

where $\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{m-1}$ are the column eigenvectors corresponding to the eigenvalues $\lambda_0, \lambda_1, \lambda_2, ..., \lambda_{m-1}$ that are found by taking the eigendecomposition of the sample covariance matrix, $\boldsymbol{R}$, of $\boldsymbol{X}$, where $\boldsymbol{R}$ is given by:

$$\boldsymbol{R} \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{5.16}$$

The eigenvalues resulting from the eigendecomposition of $\boldsymbol{R}$ are then arranged in a descending order, and we denote them such that $\lambda_0 \geq \lambda_1 \geq \lambda_2, ..., \geq \lambda_{m-1}$. The eigenvectors follow the order of their corresponding eigenvalue. If we choose $m = p$, the transformed data in $\boldsymbol{Y}$ will have the same dimensionality of $\boldsymbol{X}$, but $\boldsymbol{Y}$ will be mutually uncorrelated, with zero mean. If we choose $m < p$, we also reduce the dimensionality in $\boldsymbol{Y}$ through the transform. Applying this to the SENTINEL 1 & 2 data, with a total of 18 dimensions, we find that after the transform, the three most significant eigenvalues represent 99.99% of the variance.

Chapter 6

Implementation and Results

In this chapter, we start by presenting the general implementation of the supervised classification scheme, and continue with the experiments that are based on the supervised classification. First, we look at the generalization of the marginal probability densities, and compare supervised classification results obtained from specified parametric marginals, as well as kernel estimates. We compare the overall accuracy of these different implementations using the supervised Meta-Gaussian classification scheme, with that obtained in [53], where a maximum likelihood classification assuming multivariate Gaussian distributions was used.
A supervised implementation of the Nezer forest data using only Gamma marginals, only Gaussian, and a combination of marginals is also presented. This was done to assess how well the measured values behaves relative to the ground truth, as well as how the choice of marginals effect the classification results. This should allow us to better evaluate the classification results of the unsupervised implementations.

When the supervised part is concluded, we continue with the unsupervised part of the experiments, and present the general scheme for the unsupervised classification. We then move on to the individual unsupervised experiments, and run various test using simulated data, and real remote sensing data from SAR and optical sensors. We describe the parameters and the data that was used for each case, and present the results. A discussion of each case concludes the experiments.

6.1 Supervised Classification

For the case of supervised classification, a procedure in two steps, estimation and classification is used. All marginal probability densities are assumed to be the same unless otherwise specified in the following examples, i.e.. only Gaussian, or only Gamma. The supervised classification scheme requires as input the data $\boldsymbol{X}$ consisting of vectors $\boldsymbol{x}_i, i = 1, .., N$, where N is the number of samples, arranged in rows. $\boldsymbol{x}_i = (x_1, ..., x_p)$ is the vector containing pixel observation from p different image bands, or features, a corresponding label vector $\boldsymbol{Y}$ where the K classes to be used in the classification are represented by unique integers. And lastly, for this implementation, the number of samples, n to be used in the estimation of parameters needs to be specified.

The data is divided into a training set, to be used for estimation of parameters, and a testing set on which the classification is performed. $X_{Training}$ consists of n randomly sampled data vectors $\boldsymbol{x}$, with n_k number of points for class k. If $n_k = 0$ for some class, the random sampling is repeated until all $n_k \neq 0, k = 1, ...K$. The label vector $Y_{Training}$ follows the $X_{Training}$ index, and the remaining $(N - n)$ vectors and labels form X_{Test} and Y_{Test}.

The estimation of parameters is done separately for each class, and consists of a function that takes as input the training samples available, $\boldsymbol{X}_{Training}$ their corresponding labels, or ground truth, $\boldsymbol{Y}_{Training}$ and a class identifier. The class identifier must be equal to one of the label values.

For each of the $j = 1, .., p$ bands, estimation is done using the training points in the current class, denoting the marginal probability density for class $k = 1, ..., K$, band j by $g_{k,j}$. We then find the corresponding cumulative distribution function values using $G^{-1}_{k,j}(\boldsymbol{x}_{i,j})$ for i in class k. Using the CDF values, we then transform them into standard Gaussian values through the $\Phi(\cdot)^{-1}$ transform. After the normal quantile transform (NQT) has been computed for each band, we find the correlation coefficients, $\boldsymbol{C}_k$ for the class. This is then repeated for each of the classes in the training set. When the training, or estimation is completed, we move on to the classification.

The classification function takes as input the unknown vectors $\boldsymbol{x}_i$, the parameter estimates γ_k, and the correlation matrices $\boldsymbol{C}_k$
The classification scheme is then as follows. Starting with class k in K, we find $g_{k,j}(\boldsymbol{x_i})$ and $G_{k,j}(\boldsymbol{x_i})$ for all bands $j = 1, .., p$. We take the NQT, $\Phi(G_{k,j}(\boldsymbol{x_i}))$ Using For each unknown vector $\boldsymbol{x_i}$ we first compute $g(x_i)$ and

$G(x_i)$ given the current class and the current band. Inserting our values into

$$f_k(\mathbf{x};\gamma_k) = \frac{e^{-\frac{1}{2}\mathbf{y}(\mathbf{x};\gamma)^T(\mathbf{C}_k^{-1}-\mathbf{I})\mathbf{y}(\mathbf{x};\gamma)}}{|\mathbf{C}|^{1/2}} \prod_{j=1}^{p} g_j(x_j;\gamma_j) \tag{6.1}$$

Classification is then performed using the classical Bayes rule.

$$\hat{z}_i = \arg\max_k \pi_k f_k(\mathbf{x}) \tag{6.2}$$

where π_k is the probability of class k, based on the training set. In figure 6.1 a flowchart describing the supervised classification process, the "Classify" block is shown in detail in figure 6.2. The final output of the supervised classification scheme is an $(N-n)\times 1$ vector, that we denote $\boldsymbol{Z}$, that has as elements the classified class labels for each point in the test set, $\boldsymbol{Z} = [\hat{z}_1, ..., \hat{z}_{(N-n)}]^T$.

6.1.1 Kernel effects on supervised classifier

One of the goals for this thesis was to determine whether kernel approximations of the probability distribution functions was:

- a valid approach, and
- whether the choice of kernel had any major influence on the classification results.

Implementation

In this experiment, we ran the supervised classification scheme on some selected parametric probability distributions and kernel distribution functions. The parametric PDFs that were used was the Beta, Extreme Value, Gamma, Normal, t-Location Scale, Logistic, Rician, and the Nakagami. They were chosen due to their support. The kernels used were the Normal, Box, Triangular and Epanechnikov. In all of the below kernel estimates, the default bandwidth value, as determined to be optimal by MATLAB was used.
Three different datasets were used. The intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ of the MLC NASA/JPL AIRSAR P-band, the polarimetric features derived from the NASA/JPL AIRSAR P-band, and the reflectance values from the seven bands of the Landsat 4 TM sensor.
The classification was run 10 separate times with different random initializations of the training points. 1500 training points were selected at random from the entire dataset. The classification accuracy of each of the 10 independent runs was averaged, and the overall accuracy of each PDF along with the standard deviation of the accuracy was computed.

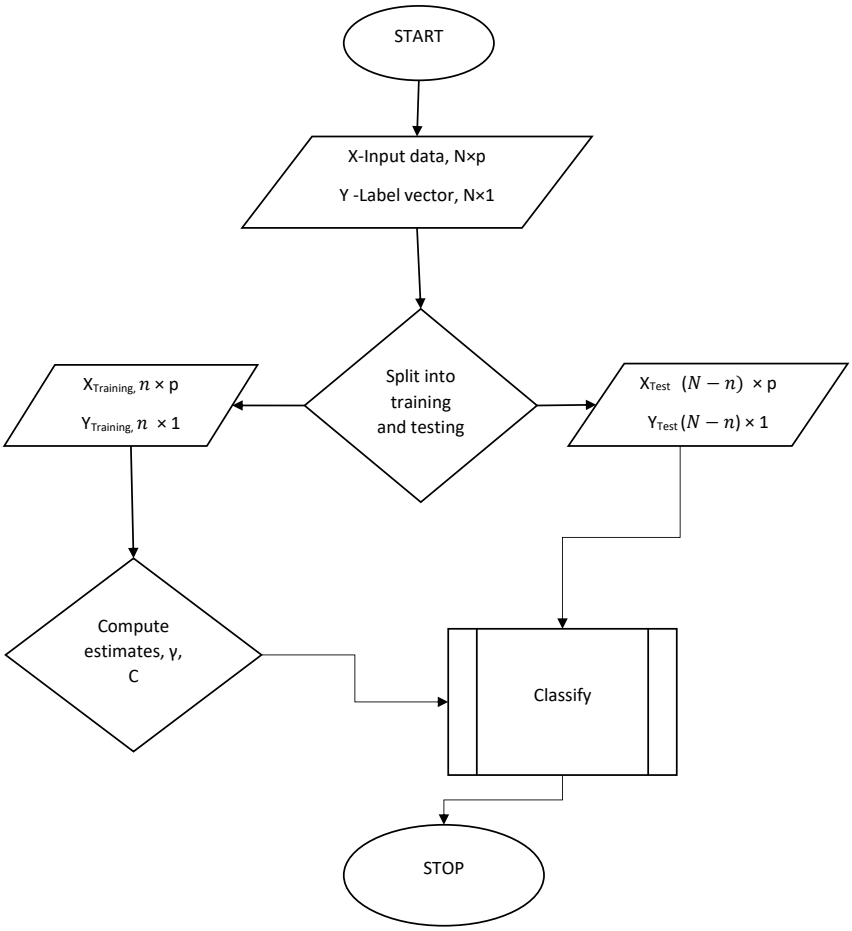


Figure 6.1: Flowchart for supervised classification. Process block "Classify" is shown in figure 6.2

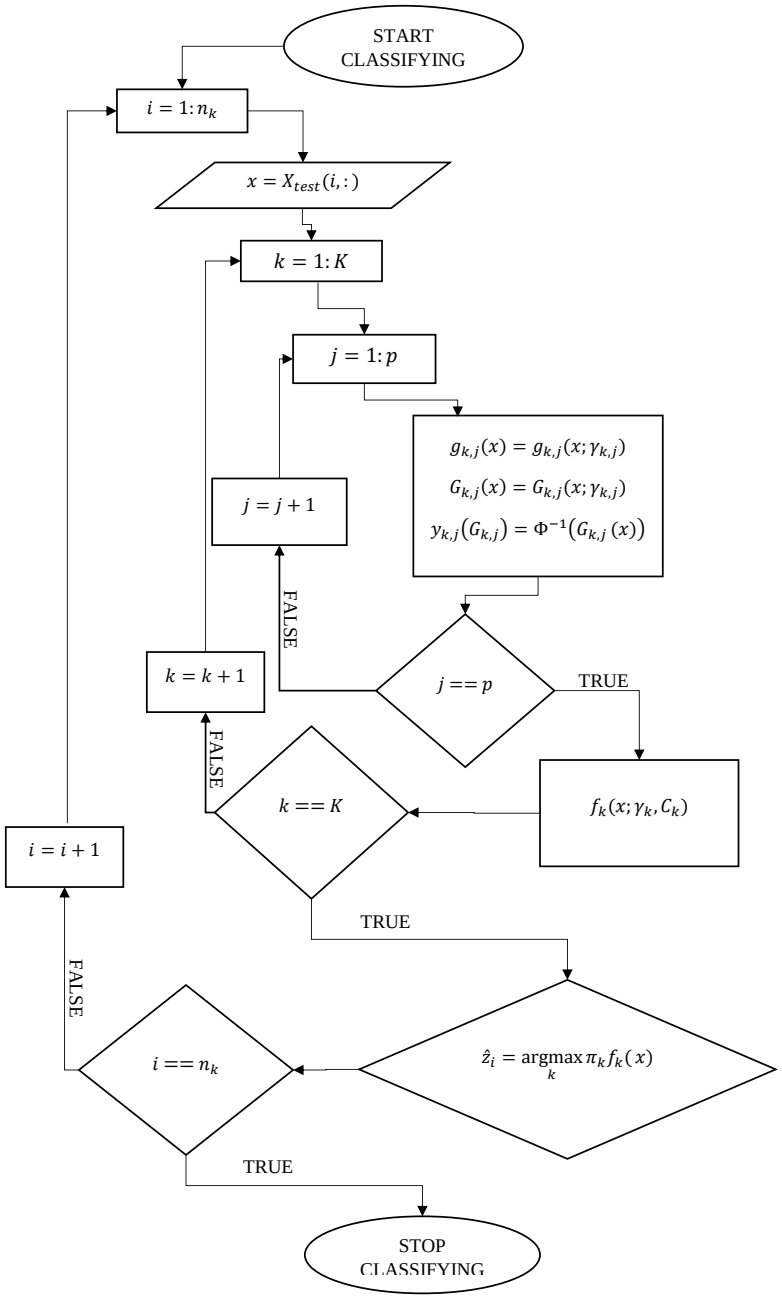


Figure 6.2: Flowchart for process block "Classify", that is used in figure 6.1 and 6.12.

Distribution	Overall accuracy in percent	Standard deviation in percent
Beta	70,76	0,38
Extreme Value	66,29	1,35
Gamma	70,53	0,37
Normal	70,59	0,28
t-Location Scale	69,24	0,32
Logistic	70,15	0,25
Rician	69,78	0,43
Nakagami	69,66	0,55
Kernel-Normal	70,27	0,37
Kernel-Box	69,83	0,49
Kernel-Triangular	70,41	0,32
Kernel-Epanechnikov	70,4	0,28

Table 6.1: Kernel classification results for P-band SAR- MLC Values

Results of kernel effects

In table 6.1,we have the classification results for the P-band MLC values, in table 6.2 we have the classification results of the polarimetric features derived from the P-band, and in table 6.3 we have the classification results of the Landsat 4 TM reflectance values.

Discussion of kernel effects

We see that overall, the kernel functions that were used generally do not perform significantly worse, and that they certainly perform better than those specified probability distribution functions that are not corresponding to the true underlying distribution. One of the main attributes of kernel functions is their ability to adapt, so therefore it is only natural that they would adapt to the data they are given. In this experiment, the accuracy was measured over 10 independent runs of the classification algorithm, and we see that the standard deviation of these accuracies were usually below 1%, with a few deviates.
This indicates that the results are consistent throughout, and that the different initializations doesn't affect the classification results in a large degree. In the case of features derived from P-band SAR data, we see a slight improvement in using the Epanechinikov Kernel with fitted bandwidth with a 71.44% accuracy, compared to the most accurate parametric probability distribution function, the Beta distribution which had a 70.75% accuracy. But, the standard deviation of the Epanechinikov was almost double that

Distribution	Overall accuracy	Standard deviation
Beta	70,75	0,25
Extreme Value	69,3	1,38
Gamma	70,27	0,66
Normal	69,76	0,67
t-Location Scale	70,38	0,43
Logistic	70,6	0,54
Rician	70,43	0,24
Nakagami	70,6	0,65
Kernel-Normal	71,32	0,46
Kernel-Box	71,25	0,37
Kernel-Triangular	70,66	0,38
Kernel-Epanechnikov	71,44	0,41

Table 6.2: Classification results for P-band SAR- Feature Values

Distribution	Classification results	Standard Deviation
Extreme Value	48,75	1,42
Generalized Extreme Value	56,89	0,59
Gamma	55,23	0,24
Normal	55,5	1,02
t-Location Scale	55,12	1,45
Logistic	57,28	0,27
Kernel-Normal	56,39	0,46
Kernel-Box	55,48	1,03
Kernel-Triangular	56,24	1,01
Kernel-Epanechnikov	56,51	0,61

Table 6.3: Kernel classification results for Landsat-reflectance values

of the Beta distribution, with 0.41% compared to 0.25%. So that if we take the extremes of each case, they would be likely to meet in the middle at 71%.

What we can also see from this, is that classification based on features derived from the NASA/JPL AIRSAR P-band MLC data, is only performing slightly better than just the three intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ of the P-band MLC, and only when the marginals are kernel estimates, which is interesting. This could indicate several things, namely that

- The specified parametric densities do not fit the data.
- The calculation of features were unnecessary. Which is not entirely impossible,

Another interesting result, is that if we look at classification results obtained in [53] and [54], shown in figure 6.3, where the author used a supervised maximum likelihood classifier, assuming multivariate normal distributions, on feature vectors that were formed by concatenating image bands(or features), we see that our results are proving better. It should be noted that in our case we used 1500 training samples selected at random from the entire set, whereas in [53] the author selected 150 training points at random, from each class. This means that although we use 450 more training points, ours are due to the laws of probability, representative of the dataset as a whole, but may be less representative on a single class, due to the low number of points that are likely to stem from some of the smaller classes.

This is a real indicator of the advantage in classification using the Meta-Gaussian distributions. It could suggest that the extra information held in the correlation matrix $\mathbf{C}$ that we obtain in the transform, outweighs, or reduces, the need of feature generation, and that it is, generally performing better than a standard maximum classification on multivariate normal densities.

And, it was shown that there is not a large difference in the overall accuracy, regardless of the choice of marginals.

Based on the test results from this experiment, we cannot conclude that a specified probability distribution function of each marginal, that would have to either be based on some pre-knowledge about the specific feature, complete guesswork, or through a iterative run over known and probable distributions such as done here would perform any better than a kernel approximation of the data. At the same time, we can conclude that a kernel approximation does not perform any worse. So, it is my recommendation that in further use

	The different feature vectors and their different combinations	**Average % classification accuracies with their corresponding standard deviation values**
1	SAR, P-band (six features)	67.89 ± 0.30
2	SAR, L-band (six features)	54.76 ± 0.44
3	SAR, C-band (six features)	43.97 ± 0.94
4	Vegetation indices (all eight features)	53.34 ± 0.47
5	Six TM bands (taken as six features)	52.07 ± 0.72
6	SAR, P-and L-band (12 features)	68.35 ± 0.48
7	SAR, P-and C-band (12 features)	68.67 ± 0.43
8	SAR, L-and C-band (12 features)	56.63 ± 0.48
9	SAR, P-, L-and C-band (18 features)	68.90 ± 0.46
10	SAR, P-band and eight VI (14 features)	75.34 ± 0.38
11	SAR, L-band and eight VI (14 features)	64.82 ± 0.47
12	SAR, C-band and eight VI (14 features)	58.19 ± 0.41
13	SAR, P-,L-bands and eight VI (20 features)	75.60 ± 0.39
14	SAR, P-,C-bands and eight VI (20 features)	75.91 ± 0.41
15	SAR, L-,C-bands and eight VI (20 features)	66.13 ± 0.53
16	SAR, P-,L-, C-bands and eight VI (26 features)	75.92 ± 0.27

Figure 6.3: Average classification results for different feature vectors and combinations, using supervised maximum likelihood classification, assuming multivariate normal distributions, obtained in [54].

of supervised classification using a Meta-Gaussian distribution, that kernel functions are used for the training and subsequent classification.

It is therefore shown that for a supervised setting of the classification based on the Meta-Gaussian distribution, using real data, a general initialization of the marginals can be done without loss of success. This can have some basis in the noise that is introduced during the acquisition and processing, and the randomness of the area itself, and the overlapping of classes, and mixing within classes. For the simulated data, in theory using the correct marginal should result in improved classification result.

6.1.2 Classification results for supervised classification

When evaluating the performance of the unsupervised classifier, it is good to have a reference to measure against. In the Nezer forest data, we have an available ground truth map for the area, but that doesn't necessarily mean that it is possible, or viable for a classifier to find exactly what the ground truth implies. We therefore ran some supervised tests, to see what the "best case" that we can achieve in an unsupervised setting is really like. Additionally, this should tell us if there are some discrepancies either in the

data, or in the ground truth map.

Implementation

The supervised classification scheme was performed once using Gamma PDFs as marginals, and once using Gaussian PDFs as marginals. Taking as input data the intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ of the MLC NASA/JPL AIR-SAR P-band,C-band and L-band, and the reflectance values from 6 bands of the Landsat 4 TM sensor, bands 1-5 and 7. A total of 15 features. 1500 training samples and their corresponding class label were selected at random from the entire set, and used in the estimation of parameters.

Results

Classification result using a supervised scheme, separately using only Gaussian and only Gamma PDFs as marginals is shown in figure 6.4, and corresponding confusion matrices in figure 6.5 for the Gamma PDF case, and figure 6.6 for the Gaussian PDF case.

Discussion

Here we see that the only distinct clear class is that of the bare soil, where 99.3% of the points where correctly classified when using Gamma marginals, and 99.7% when using Gaussian marginals. For the other classes, the commission error varies between 29.2% to 70.8%, and the omission error is in the ranges of 28.66% to 71.4% in the Gamma case, resulting in an overall classification of 73.9%.

In the Gaussian case, the commission error is slightly lower, in the range 27.7% to 61.0%, and an omission error between 30.3% to 56.5%. Here the overall classification accuracy is 75.5%.

The two most distinct, and dominating classes in the classified output, apart from the bare soil, are that of pine aged $33-41$ years, and pine $11-14$ years. What we can also tell, both by looking at the classified images in figure 6.4 and the confusion matrix in figure 6.6 and 6.5, is that the error is mainly caused by pine points wrongly being classified to one of its neighbouring age groups. This suggests that we should not expect to be able to distinguish that well between neighbouring tree age groups in the unsupervised classification, but that the bare soil should be well defined. This could imply that using fewer classes may be a better approach than to assume six different age groups in addition to the bare soil class.

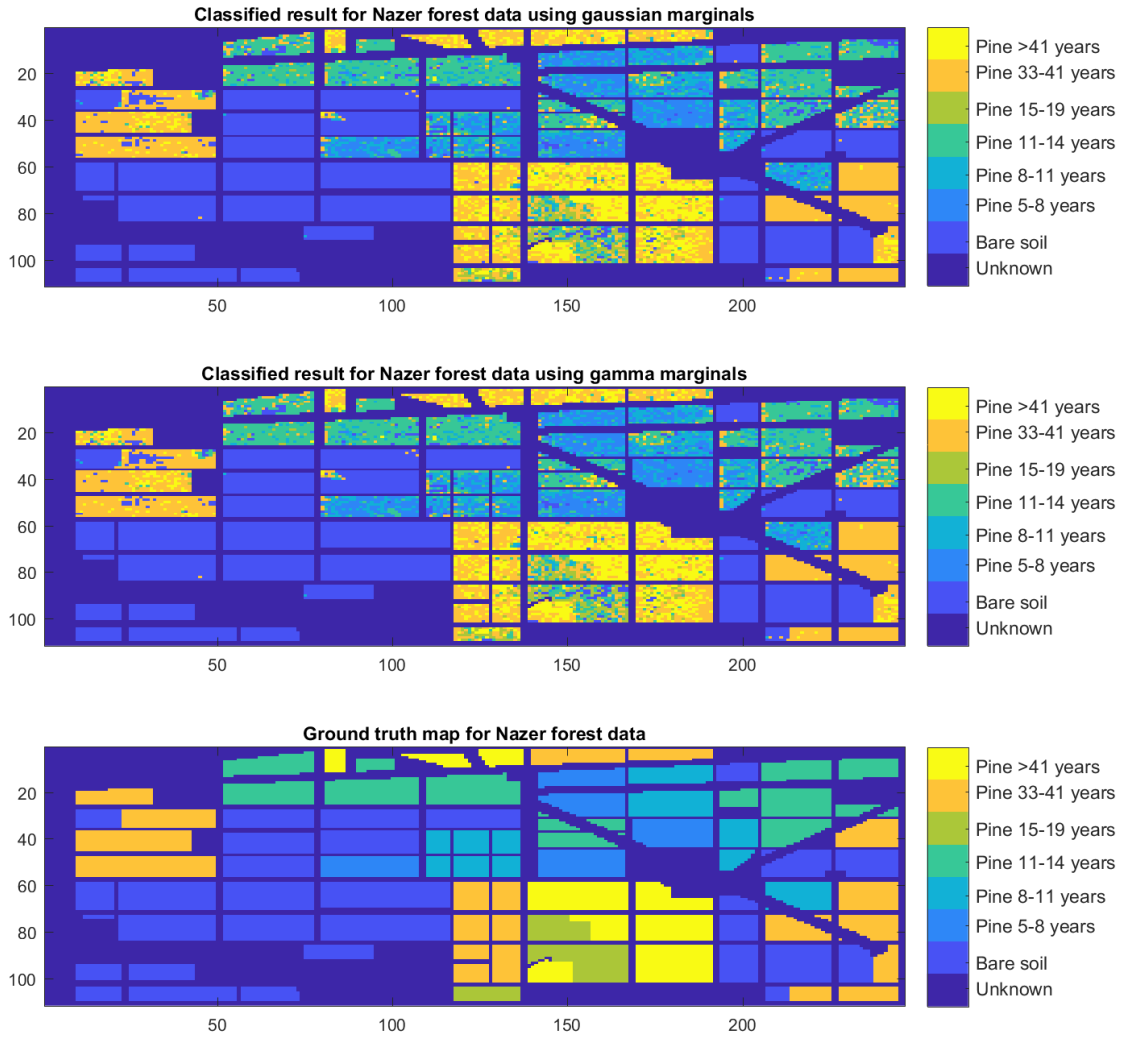


Figure 6.4: Comparison between classified maps and ground truth map for Nezer forest data.
Top: Gaussian marginals
Middle: Gamma marginals
Bottom: Ground truth
Ground truth labels for training points have been reinserted into the classified result for viewing purposes. Data with "unknown" label was not used in the classification.

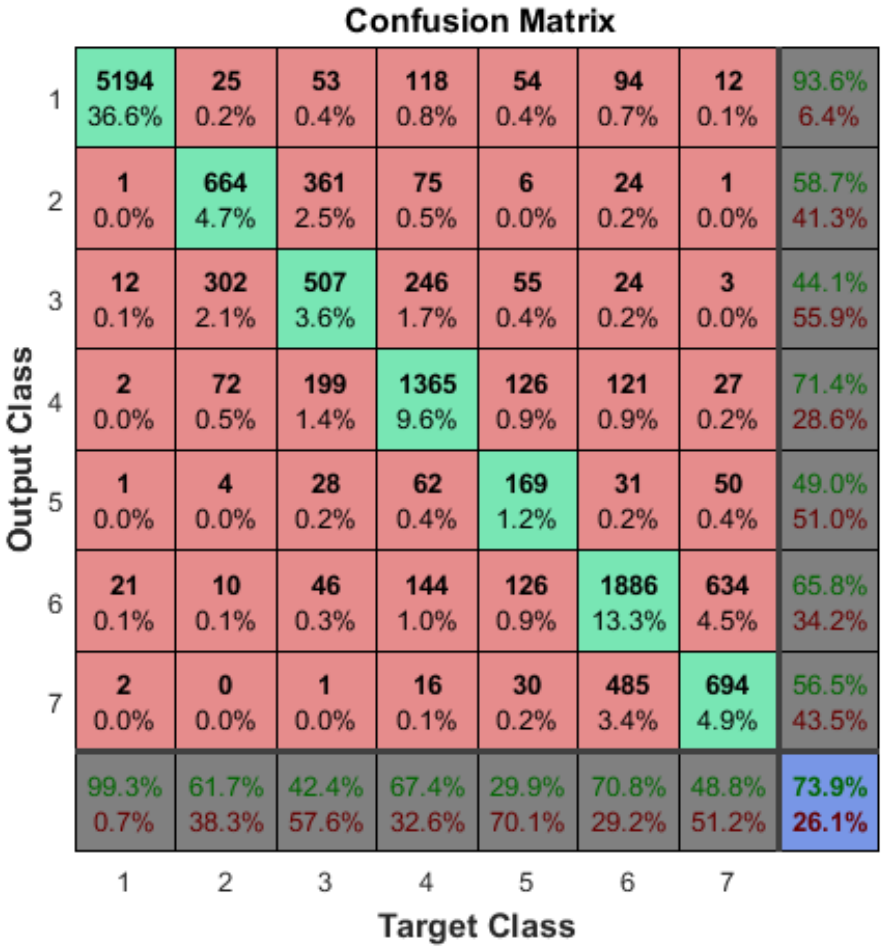


Figure 6.5: Confusion matrix for supervised classification of Nezer forest data using Gamma marginals. Diagonal elements are points which are correctly classified. Bottom row represents omission error, far right column is the commission error.

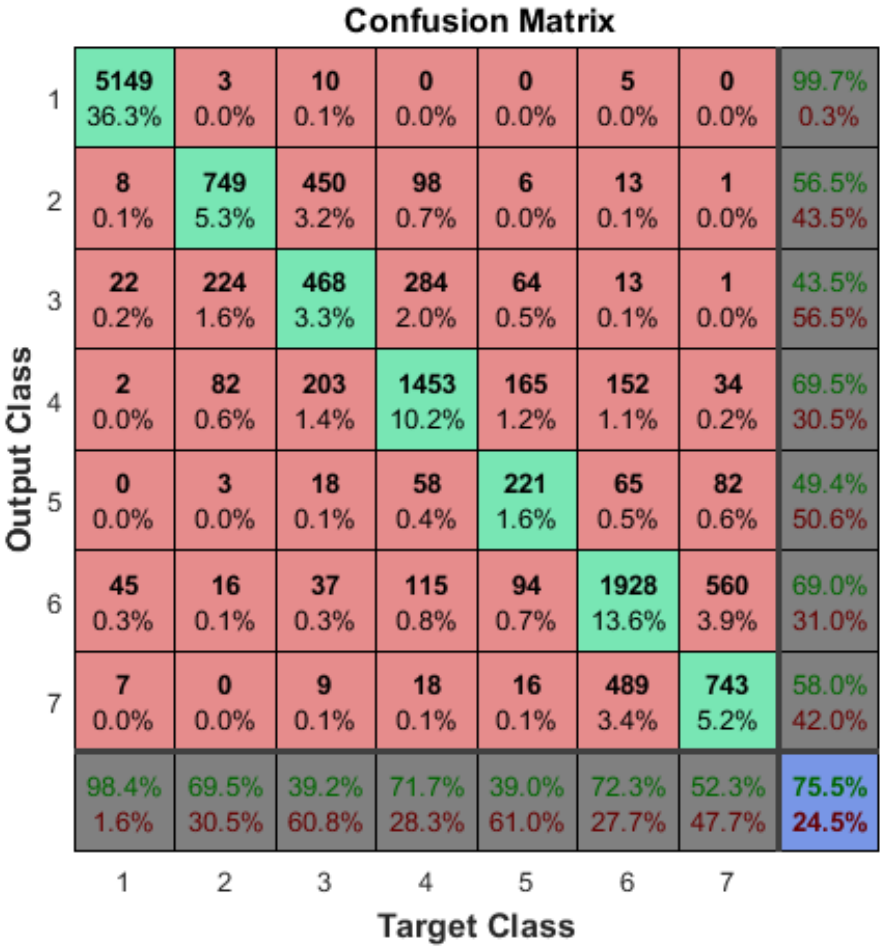


Figure 6.6: Confusion matrix for supervised classification of Nezer forest data using Gaussian marginals. Diagonal elements are points which are correctly classified. Bottom row represents omission error, far right column is the commission error.

6.1.3 Supervised classification using a combination of marginals

One of the strengths of the Meta-Gaussian classification scheme, is its ability to combine features with different marginal probability distribution functions, and represent it in a unified Meta-Gaussian multivariate function. To investigate the effect of using individually specified marginals, we repeat the above experiment, but now using Gamma marginals for the SAR features, and Gaussian Marginals for the optical features. In theory, this could improve the classification results.

Implementation

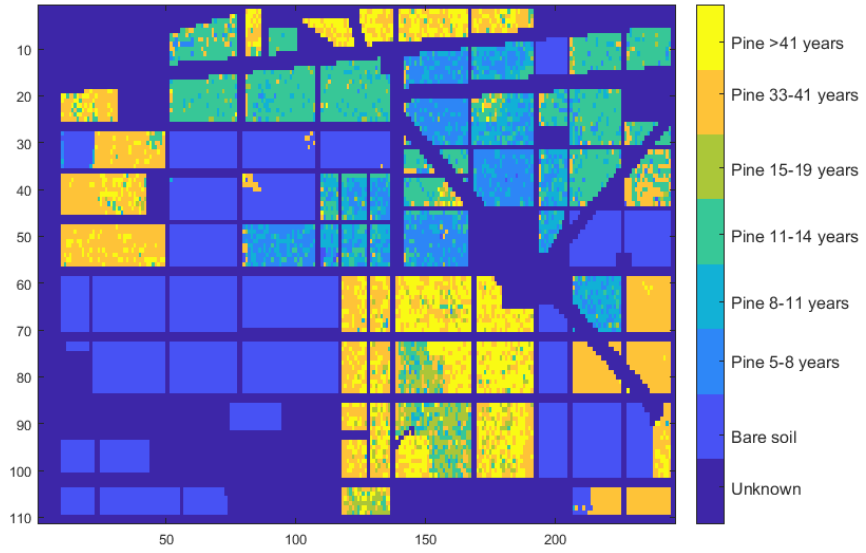
Classification was performed on nine MLC values, namely the $|S_{hh}|^2, |S_{hv}|^2$ and $|S_{vv}|^2$ from the C-band, L-band and P-band NASA/JPL AIRSAR, as well as six of the Landsat 4 TM bands, bands 1-5 and 7. A total of 15 features. Supervised classification using the Meta-Gaussian GMDAS scheme with Gamma marginals for the 9 SAR bands, and Gaussian marginals for the 6 optical bands was was tested. 1500 training samples and their corresponding class label were selected at random from the entire set, and used in the estimation of parameters.

Results

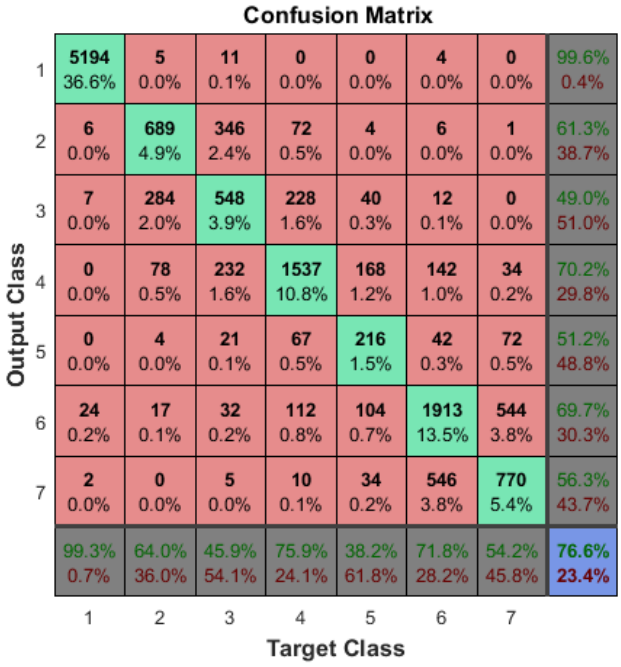
Classification result using a supervised scheme, with a combination Gamma and Gaussian marginals is shown in figure 6.7(a), and the confusion matrix of the classification result in 6.7(b).

Discussion

In figure 6.7 we see again that the only distinct clear class is that of the bare soil, and overall we have the same issue with adjacent classes being mixed as we had in the previous example, using only Gaussian and only Gamma marginals. What we do find is that the overall classification accuracy is slightly better than in the two previous cases. In this case, using the two theoretically correct marginals, we have 76.6% of the pixels correctly classified, whereas when only using Gamma we had 73.9%, and in the Gaussian case 75.5% correctly classified pixels. The same training points were used for all three cases. If we look at the histograms of the different features and classes, we may understand why this is happening. In figure 6.8 we show the histograms of the SAR features, and in figure 6.9 the histograms of the optical bands used in this experiment. And we can see that for the SAR



(a) Classified map, ground truth labels for training points have been reinserted into the classified result for viewing purposes



(b) Confusion matrix. Diagonal elements are points which are correctly classified. Bottom row represents omission error, far right column is the commission error.

Figure 6.7: Results for the supervised classification using a combination of Gamma and Gaussian marginals

features, the Gamma has a seemingly better fit, for all of the channels apart from the HH and VV of the C-band MLC, where we have a uniform distribution, and the Gaussian with its heavier tails captures it better. Although it should be noted that the x-axis of the histogram plots in figure6.8 were cropped for displaying reasons, which means that we do no see the heavy tails in the negative range. For the optical bands, we generally have a worse fit for both Gamma and Gaussian densities than we did in the SAR features, but the Gamma density seems to be the lesser of two evils in this case as well. One might wonder why the classification using Gamma marginals in figure 6.5 was the worst of the three cases we tested then. To hopefully answer this, we take a look at the histograms of some of the classes, for a few of the features, since these are the constituents of the marginal estimates that we are using in the transforms. In figure 6.10 we show histograms for the optical, class wise values, and similarly in figure 6.11 for the SAR features. And we note that for the SAR, the fitted Gamma densities seem to better fit, this is perhaps especially apparent in the class 1, HV channel P-band combination, shown in the top right of figure 6.11.

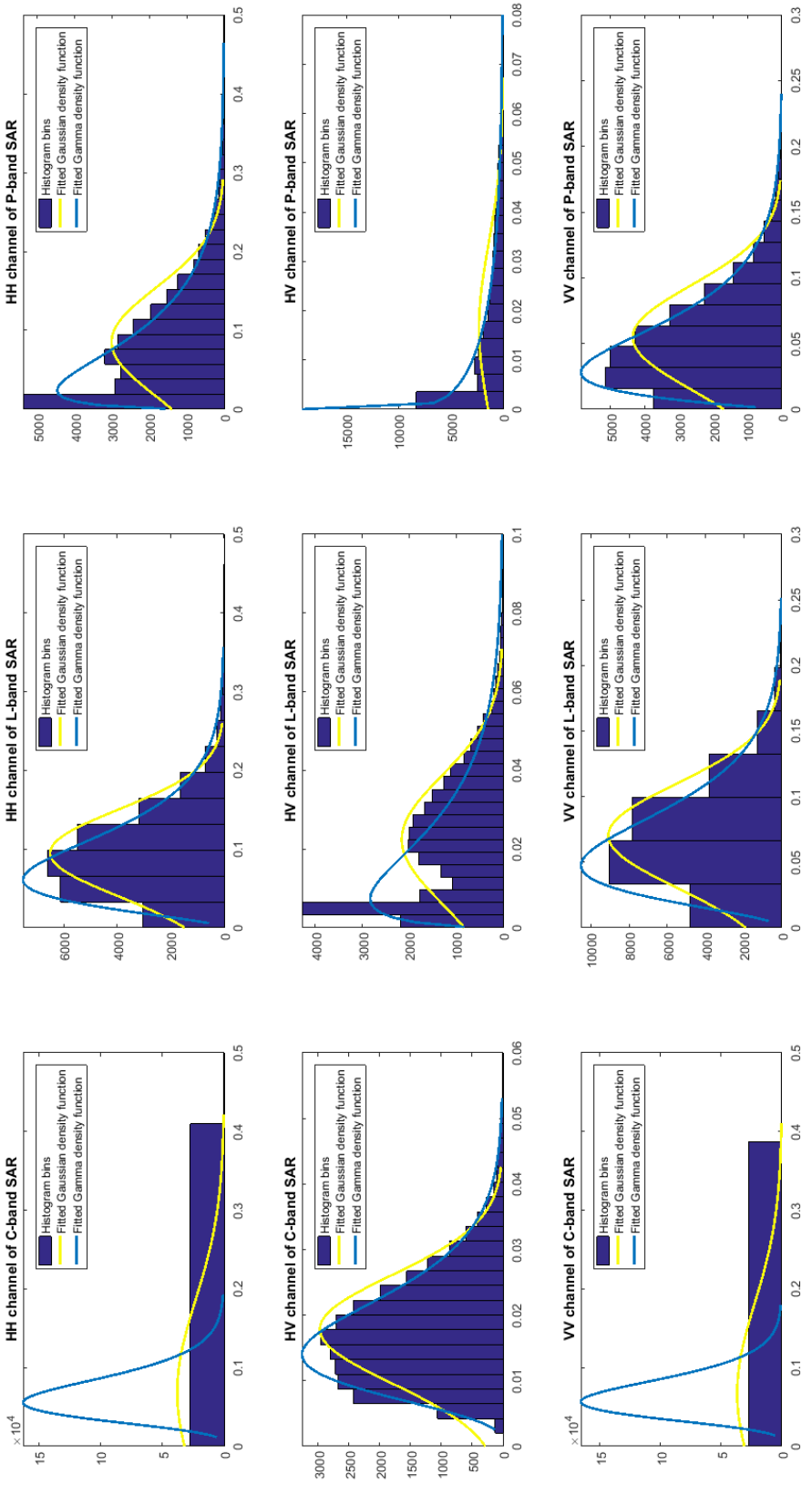


Figure 6.8: Histogram plots of SAR features $|S_{hh}|^2, |S_{hv}|^2$ and $|S_{vv}|^2$, of the MLC for bands C,P and L with fitted Gamma and Gaussian density curves. Yellow curves are the fitted Gaussian curves, and blue curves are the fitted Gamma. Dark blue bars are the histogram bins.

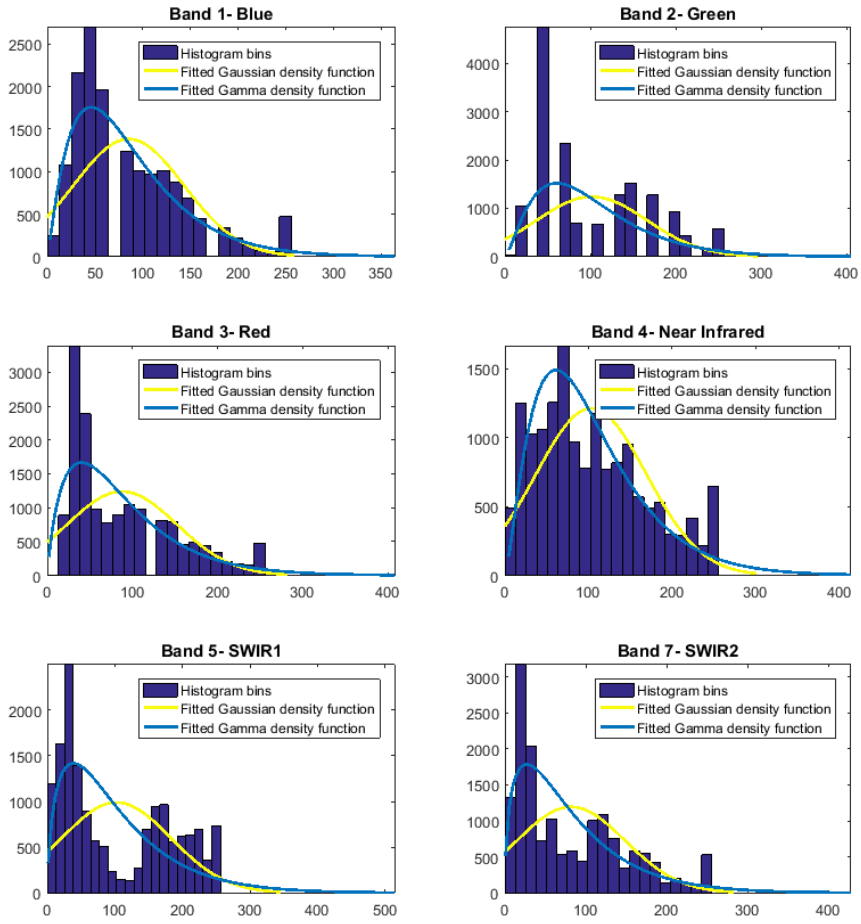


Figure 6.9: Histogram plots of the Landsat TM 4 bands $1-5$ and 7 with fitted Gamma and Gaussian density curves. Yellow curves are the fitted Gaussian curves, and blue curves are the fitted Gamma. Dark blue bars are the histogram bins.

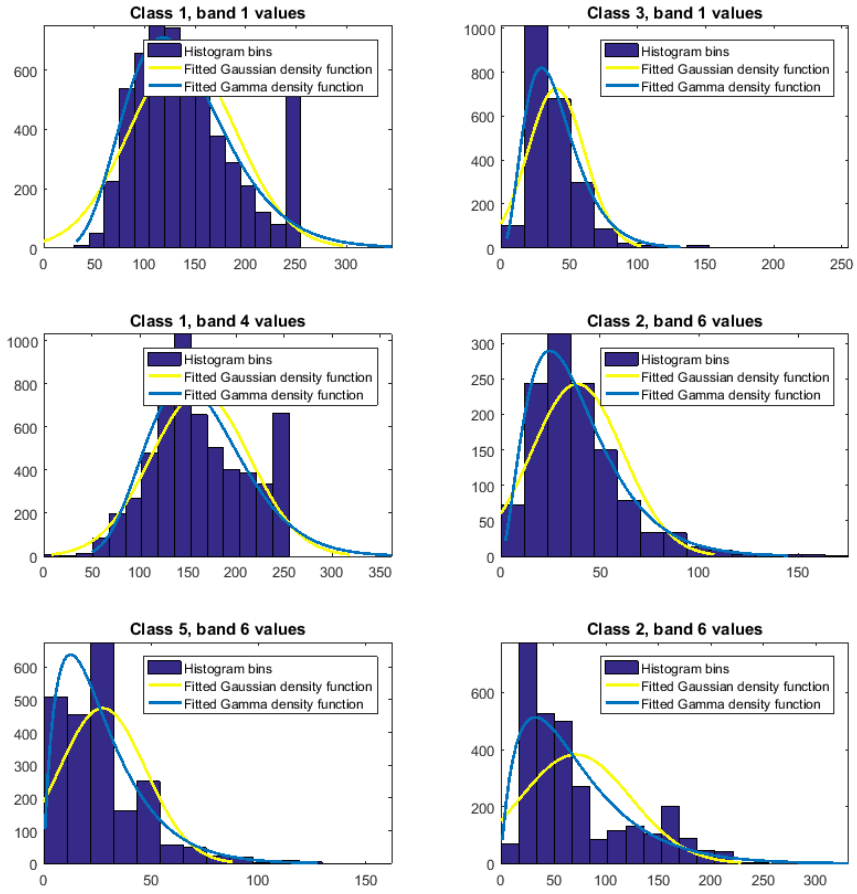


Figure 6.10: Histogram plots of selected classes of the Nezer forest, values from the Landsat TM 4 bands $1-5$ and 7 with fitted Gamma and Gaussian density curves. Yellow curves are the fitted Gaussian curves, and blue curves are the fitted Gamma. Dark blue bars are the histogram bins.

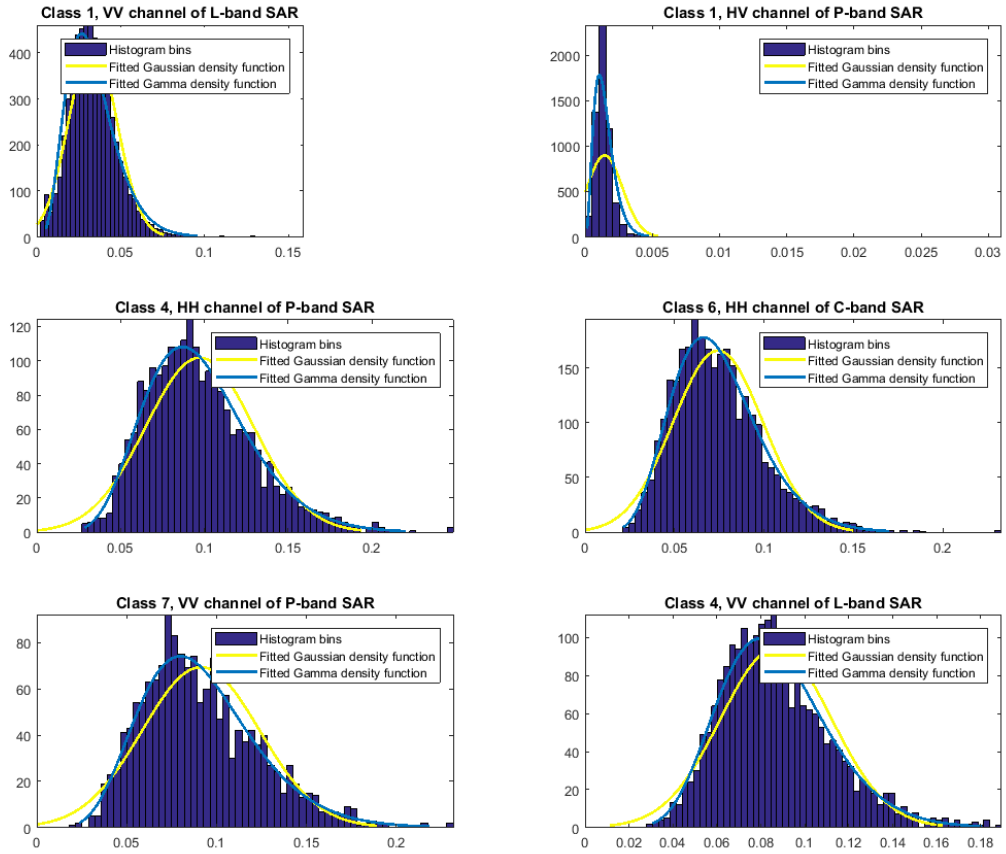


Figure 6.11: Histogram plots of selected classes of the Nezer forest, values from the SAR features $|S_{hh}|^2, |S_{hv}|^2$ and $|S_{vv}|^2$, of the MLC for bands C,P and L with fitted Gamma and Gaussian density curves. Yellow curves are the fitted Gaussian curves, and blue curves are the fitted Gamma. Dark blue bars are the histogram bins.

6.2 Unsupervised Classification using Meta-Gaussian Distributions

Using a Generalized Mixture Decomposition Algorithmic Scheme as a baseline, a classification scheme was implemented in Matlab. In this current implementation all marginal probability density functions are assumed to be Gaussian unless otherwise specified.

The unsupervised classification scheme requires as input the data $\boldsymbol{X}$ consisting of vectors $\boldsymbol{x}_i, i = 1, .., N$, where N is the number of samples, arranged in rows. $\boldsymbol{x}_i = (x_1, ..., x_p)$ is the vector containing pixel observation from p different image bands, or features, and the number of classes to be used in the clustering, K.

Given this, it would start by randomly assigning points to classes, for now assuming equal probabilities for all classes $k = 1, ..., K$. After this initialization, the iterative clustering scheme begins.

Based on the current cluster assignments, the estimation of parameters is done separately for each class. For each of the $j = 1, .., p$ bands, estimation is done using the points in the current class, denoting the marginal probability density for class $k = 1, ..., K$, band j by $g_{k,j}$. We then find the corresponding cumulative distribution function values using $G^{-1}_{k,j}(\boldsymbol{x}_{i,j})$ for i in class k. We transform the CDF values, into standard Gaussian values through the $\Phi(\cdot)^{-1}$ transform. After the normal quantile transform (NQT) has been computed for each band, we find the correlation coefficients, $\boldsymbol{C}_k$ for the class. This is then repeated for each of the classes in the current clustering. During the estimation process, we have also included a rule to ensure that no class is left empty. If an empty cluster occurs, points from the other classes that have the lowest probability of membership is relocated. After the estimation we compute the change indicator,

$$\delta_i = |\boldsymbol{\gamma}_{i-1} - \boldsymbol{\gamma}_i| + |\boldsymbol{\pi}_{i-1} - \boldsymbol{\pi}_i| \tag{6.3}$$

Where $\boldsymbol{\pi}_i$ is the vector containing the a priori probabilities of iteration i, and γ_i denotes the parameters of the ith iteration. When the estimation is completed, we move on to the classification.

The classification part uses the unknown vectors $\boldsymbol{x}_i$, the parameter estimates γ_k, and the correlation matrices $\boldsymbol{C}_k$ obtained in the estimation part. The classification scheme is then as follows. Starting with class k in K, we

find $g_{k,j}(x_{i,j})$ and $G_{k,j}(x_{i,j})$ for all bands $j = 1, .., p$. We take the NQT, $\Phi(G_{k,j}(\boldsymbol{x_i}))$ Using For each unknown vector $\boldsymbol{x_i}$ we first compute $g(x_i)$ and $G(x_i)$ given the current class and the current band. Inserting our values into

$$f_k(\mathbf{x}; \gamma_k) = \frac{e^{-\frac{1}{2}\mathbf{y}(\mathbf{x};\gamma)^T(\mathbf{C}_k^{-1}-\mathbf{I})\mathbf{y}(\mathbf{x};\gamma)}}{|\mathbf{C}|^{1/2}} \prod_{j=1}^{p} g_j(x_j; \gamma_j) \tag{6.4}$$

Classification is then performed using the classical Bayes rule.

$$\hat{z}_i = \arg\max_k \pi_k f_k(\mathbf{x}) \tag{6.5}$$

Clusters are updated according to the results of the classification. The procedure is repeated until we achieve a change in parameters, between two subsequent iterations, δ_i that is smaller than the specified error, e, or we reach the maximum number of iterations, denoted *maxiter* in figure 6.12. The classified output of the unsupervised classification scheme is an $N \times 1$ vector, denoted $\boldsymbol{Z}$, that has as elements the classified class labels corresponding to each point in the unknown data, $\boldsymbol{Z} = [\hat{z}_i, ..., \hat{z}_N]^T$.

In figure 6.12 we show a flowchart describing the process. Note that the "Classify" block, shown in detail in figure 6.2, is the same as in the supervised case, but that the loop indicator $i = n_k$ is changed to $i = N$, and the assignment $\boldsymbol{x} = \boldsymbol{X}_{Test}(i, :)$ will be changed to $\boldsymbol{x} = \boldsymbol{X}(i, :)$, since we are now classifying using all points in the data set, not on what we have denoted the test set, $\boldsymbol{X}_{Test}$. As we see in figure 6.12, after the initialization of the clusters, we repeat the estimation of parameters based on the current class membership, the computation of the change indicator, δ_i, and classify based on the current parameter estimates until one of the stopping criteria is reached.

6.2.1 Unsupervised classification of bivariate test data

The initial step of the unsupervised classification was after implementation to test it using very simple test data. In this case, it was using two dimensional Gaussian data, with four separate, distinct clusters, as described in section 5.3. Starting out with a data set in which we expect a 100% correct classification, is primarily done to assess whether the method, and the implementation works, not so much to evaluate any other aspects.

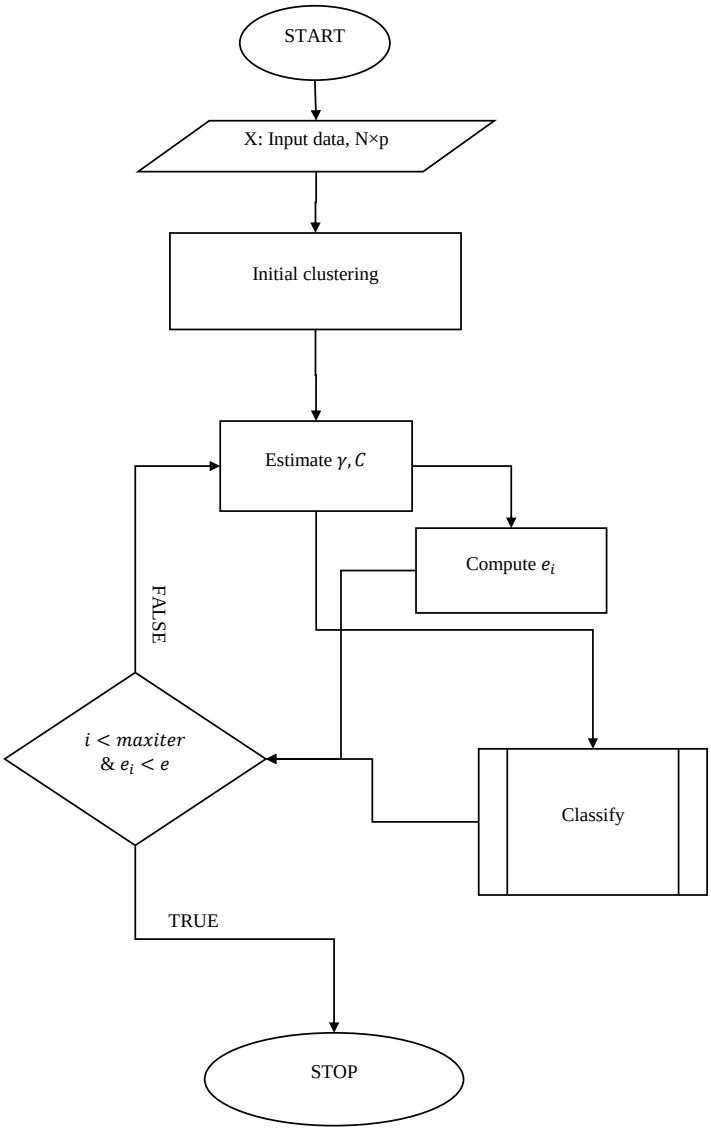


Figure 6.12: Flowchart for unsupervised classification

Implementation

The unsupervised Meta-Gaussian classification scheme was used on the simulated Gaussian dataset. The number of classes, K, was set as 4, which was also equal to the true number of classes. Maximum number of iterations was set as 100, and the error criteria e was set to 0.00001. Random assignment of the clusters was used to initialize the algorithm.

Results

Classification results are shown in figure 6.13. The algorithm converged to the error criteria in 14 iterations.

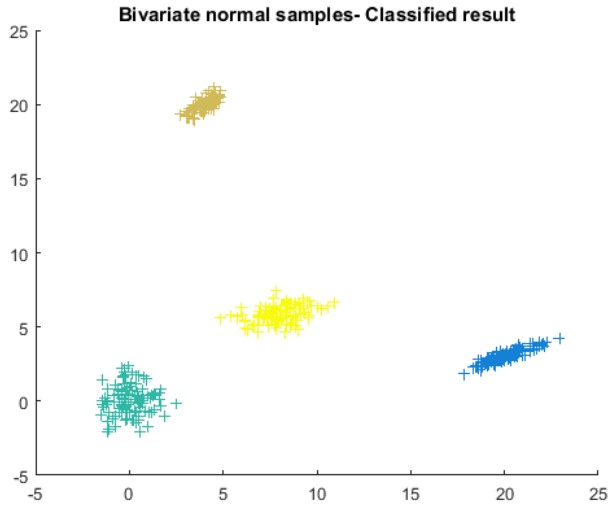


Figure 6.13: Classification result for simulated data

Discussion

After some inconclusive results, which failed to converge, an error was discovered in the computation of the matrix inverse in the Meta-Gaussian function. After this was corrected, we achieved the expected correct classification accuracy of 100%.

6.2.2 Unsupervised classification using Nezer Forest dataset

Using the Nezer forest dataset, an initial implementation using the P-band MLC data, and considering a two class case, where the intention was separation of the forest and the unknown terrain. Additionally, we wanted to compare the convergence and results when using prior probabilities obtained from the current clustering, to an implementation without using prior probabilities, which effectively means that we force $P_1 = P_2 = 0.5$.

Implementation

The data that was used was the intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ of the MLC NASA/JPL AIRSAR P-band. The number of classes, K was set to 2, the error e was set to 0.00001 and maximum number of iterations, *maxiter* was set to 50. The algorithm was initialized using random seeding of the data, and the classification was run separately for case 1- including current class probabilities, and case 2- no class probabilities were included.

Results

The classification results for the two cases, at iteration number 5, 10 and 50 are shown in figure 6.14

Discussion

We see that the algorithm is now able to go from a completely random starting point in the first iteration, to recognizable shapes in the fifth iteration. In subsequent iterations, we see that by comparing the results of the classification to the ground truth map, the the two classes that are found are apparently:

- ω_1= Bare soil
- ω_2= Forest and unknown pixels

If we compare the two cases iteration for iteration, it becomes clear that that of case 1, where priori probabilities are used in the calculation of the Meta-Gaussian, converges faster than that of case 2. We see that one of the problematic areas for the P-band SAR, is to pick up the vertical lines between the fields. Considering the contextual information, it is not unlikely that the lines are in fact dirt roads, which would likely have the same backscatter properties as bare ground, or it could be caused by the viewing angle.

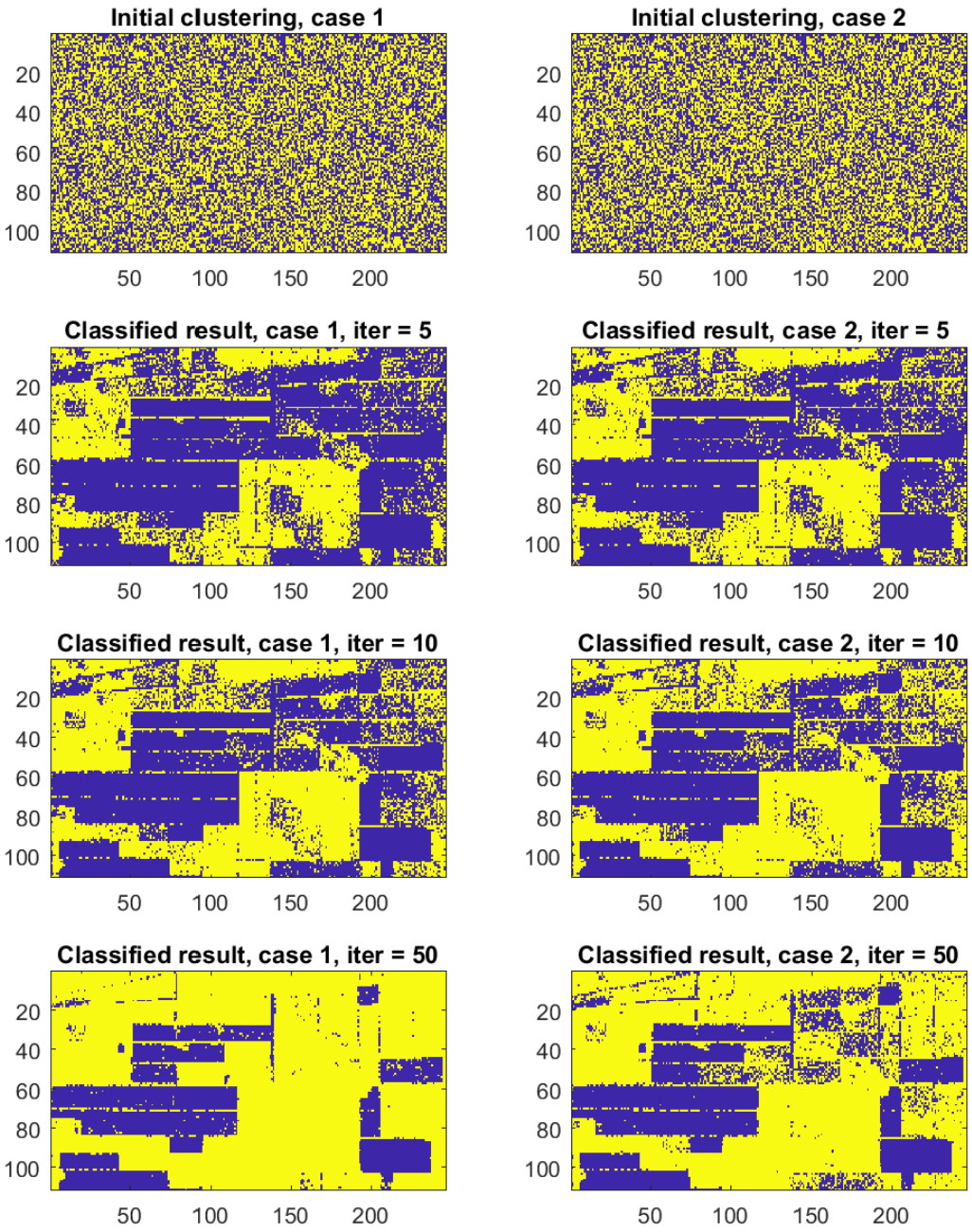


Figure 6.14: Results for a two class classification: comparison between classification using current class probabilities, denoted case 1, and without considering current class probabilities, denoted case 2.

6.2.3 Two class case- removing unknown values

We now tried with the same two class scenario, but where the unknown class have been removed. The expected class separation is now vegetation and bare soil. We also expect a good class separation.

Implementation

The data that was used was the intensity values $|S_{hh}|^2, |S_{hv}|^2, |S_{vv}|^2$ of the MLC NASA/JPL AIRSAR P-band. The number of classes, K was set to 2, the error e was set to $1 \cdot 10^{-6}$ and maximum number of iterations, *maxiter* was set to 45. The algorithm was initialized using random seeding of the data. Data having a ground truth label of 0 was removed before classification.

Results

In figure 6.15, we can see the results from this clustering. The values corresponding to the unknown values have been masked out, and is shown in dark blue in the ground truth and the classified maps.

Discussion

After 20 iterations, the absolute difference in parameters between subsequent iterations begins a steady drop towards zero. In figure 6.15, we can see the results from this clustering. The values corresponding to the unknown values have been masked out, and is shown in dark blue. We see that in only two iterations, it has already converged well towards a good separation, and in 20 iterations, it has almost correctly separated the two classes. If we segment our reduced ground truth into two classes, those containing trees, and those containing bare soil, and compare against the classified output, we find that the final clustering correctly classified 97.88% of the pixels.

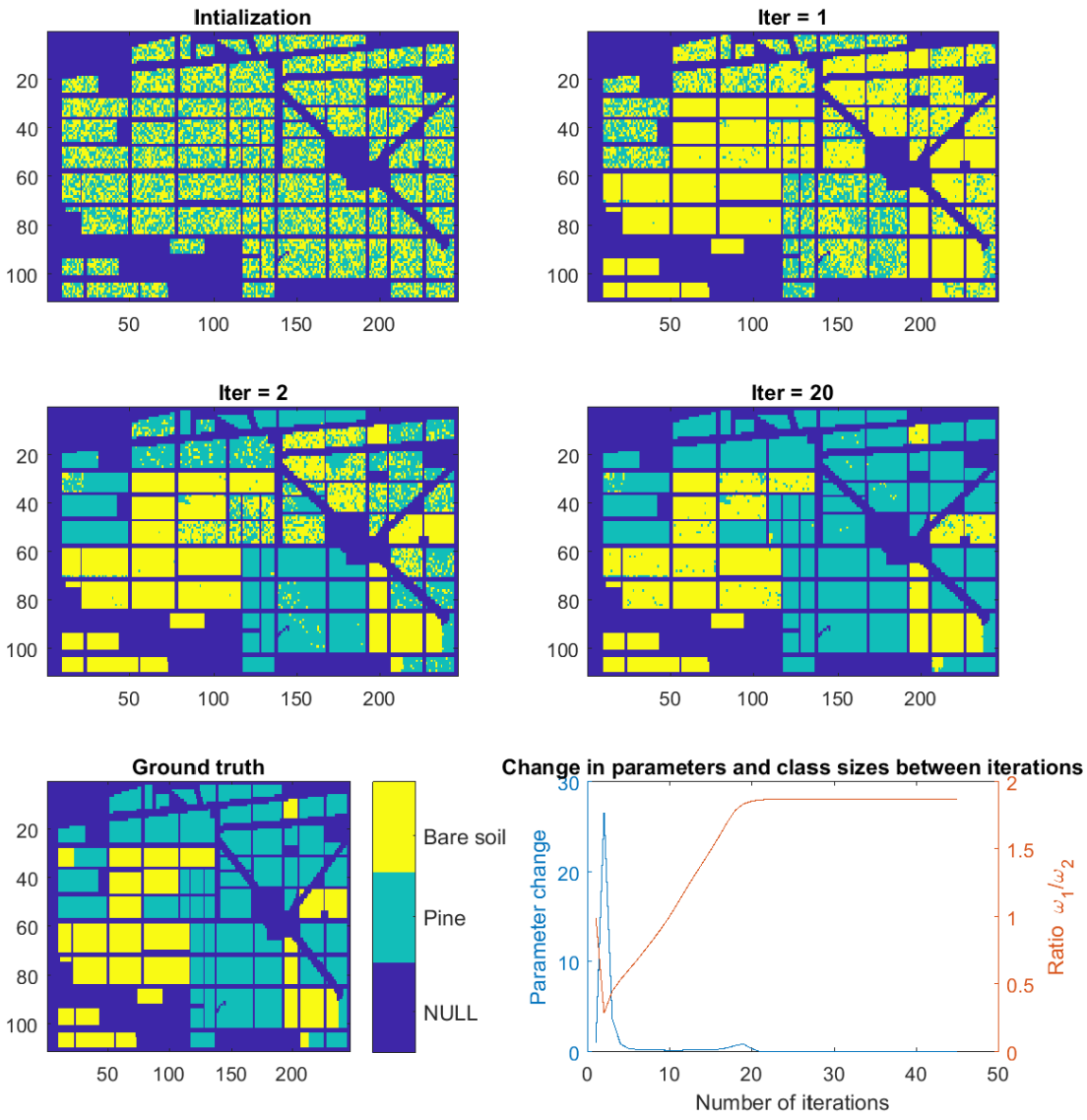


Figure 6.15: Results for a two class classification

6.2.4 Seven-class case

Seven is the number of classes that we have in our ground truth map for the Nezer forest if we ignore the "unknown" class, and we therefore assume that this should be the correct, and true number of classes. In this experiment, we attempt a clustering based on these seven classes, and we also wish to compare the Meta-Gaussian method with the more conventional Maximum Likelihood method using multivariate Gaussians.

Implementation

Classification was performed on three MLC values from the P-band NASA/JPL AIRSAR, namely $|S_{hh}|^2$, $|S_{hv}|^2$ and $|S_{vv}|^2$. Unsupervised classification using a standard GMDAS implementation, where a multivariate normal distribution was used, and the Meta-Gaussian GMDAS scheme with Gaussian marginals was was tested. The number of classes, K was set to 7, the error e was set to $1 \cdot 10^{-6}$ and maximum number of iterations, *maxiter* was set to 45. The algorithm was initialized using random seeding of the data. Data having a ground truth label of 0 was removed before classification.

Results

Classification results are shown in figure 6.16. Note that the class assignment will be different between initializations, and will rarely, if at all follow the same "numbering" as in the ground truth labels.

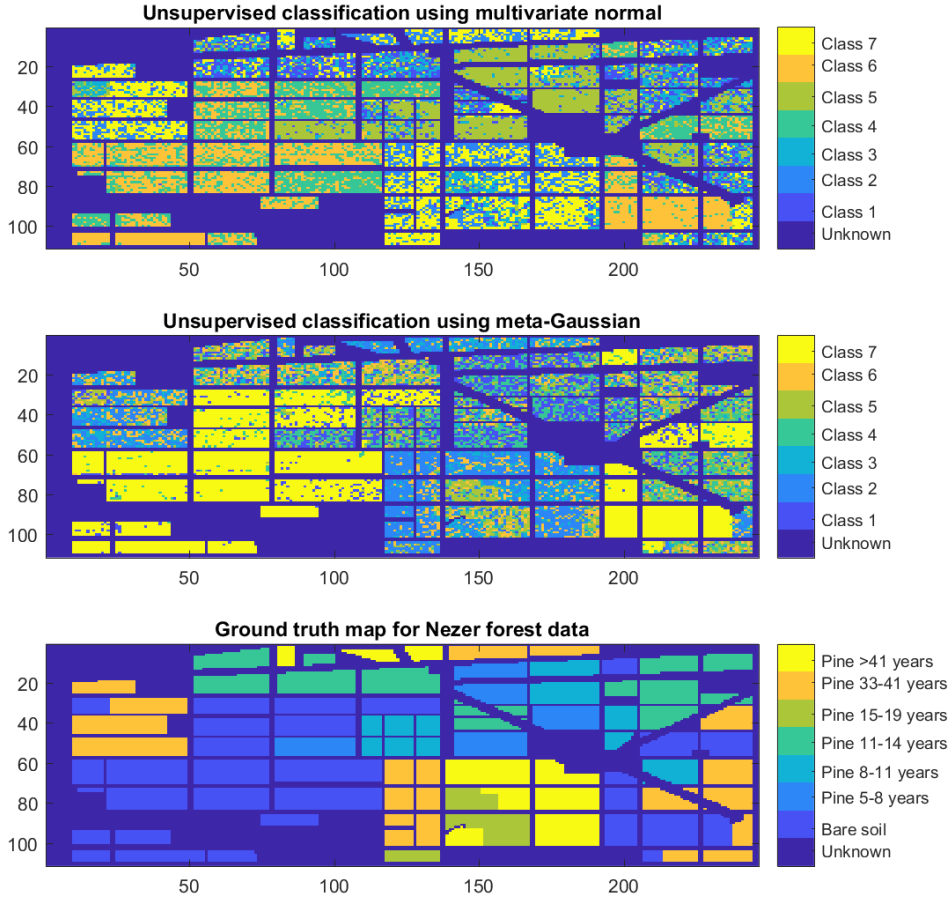


Figure 6.16: Comparison between classified map for GMDAS using Meta-Gaussian distribution, GMDAS using multivariate normal distribution, and ground truth map for Nezer forest data. Data with "unknown" label was not used in the classification.

Discussion

What we see is that in the classification using the multivariate normal distribution, the bare soil class is not limited to one class indicator, but is split amongst class 6 and 4 in the classified map. When using the Meta-Gaussian, this has correctly been grouped into one class. We also see the same tendency of an inability to correctly separate between adjacent age groups, in both cases the pine> 41 and $33 - 41$ years age groups are seemingly mixed.

We saw the same tendency in the supervised classifications, shown in figure 6.5,6.6 and 6.7.

6.2.5 Seven-class case using a combination of marginals

In the prior unsupervised experiments, we have used the same marginal probability density function on all features. We now wish to test the performance when using Gamma marginals for the SAR bands, and Gaussian marginals for the optical bands.

Implementation

Classification was performed on nine MLC values, namely the $|S_{hh}|^2, |S_{hv}|^2$ and $|S_{vv}|^2$ from the C-band, L-band and P-band NASA/JPL AIRSAR, as well as six of the Landsat 4 TM bands, that is, bands 1-5 and 7. A total of 15 features. Unsupervised classification using the Meta-Gaussian GMDAS scheme with Gamma marginals for the 9 SAR bands, and Gaussian marginals for the 6 optical bands was was tested. The number of classes, K was set to 7, the error e was set to $1 \cdot 10^{-4}$ and maximum number of iterations, *maxiter* was set to 60. The algorithm was initialized using random seeding of the data. Data having a ground truth label of 0 was removed before classification.

Results

Classification results are shown in figure 6.17. Note that the class assignment will be different between initializations, and will rarely, if at all follow the same "numbering" as in the ground truth labels.

Discussion

What we see in the classified map in figure 6.17 is that we have fairly clear groupings. What we also see is that the bare soil class is not one classified as one distinct class, but rather the four classes $\omega_1, \omega_2, \omega_4, \omega_6$ using the labelling in figure 6.17. It then follows that three classes that are left in the classified map, represent the six forest classes. We find that class ω_3 and ω_7 seem to represent the three oldest pine age groups, $15-19$, $33-41$ and >41 years. Class ω_5 is fairly clearly the three youngest pine age groups, $5-8$, $8-11$ and $11-14$ years.

If we now do a merging of classes according to these findings, we have three classes: bare soil, young pine trees and old pine trees. In figure 6.18 we show the merged classified map, along with a ground truth map using the

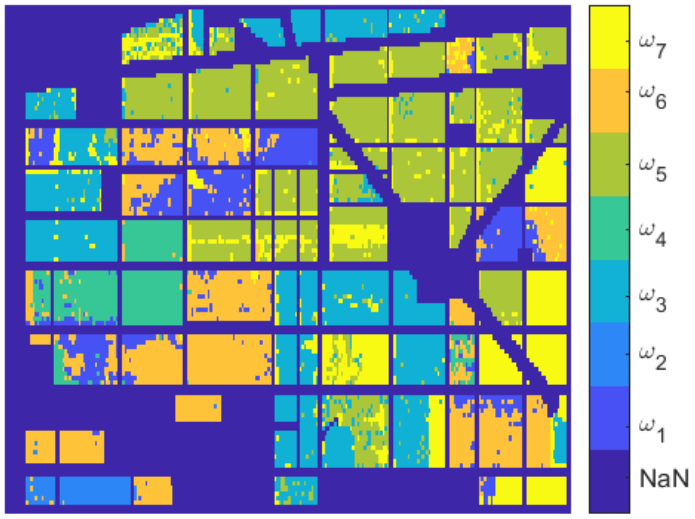
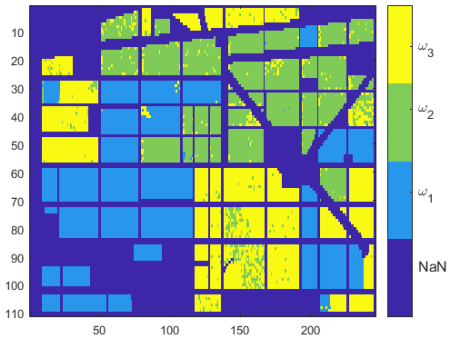
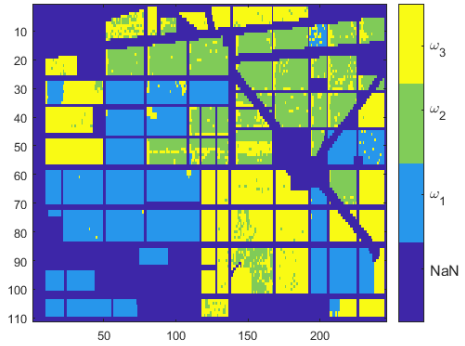


Figure 6.17: Results for a 7 class unsupervised classification using Gamma marginals for the SAR features, and Gaussian for the optical bands. NaN class was not used in the classification.

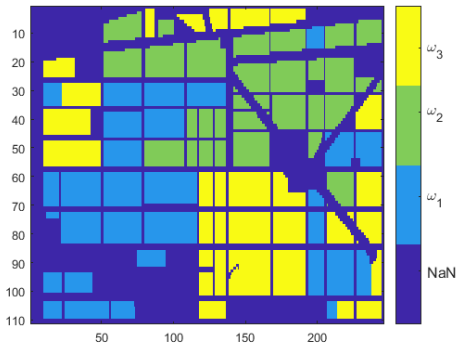
same grouping, and a merged classified map from the supervised classification using the same features and the same marginals. We now denote ω_1 as bare soil, ω_2 as young trees (age $5-14$ years), and ω_3 as old trees, (age $15- > 41$ years). In this particular case, we also included a confusion matrix, since we were able to obtain a clearly defined segmentation with the three classes, and find that the overall classification accuracy was 91.4% when using the ground truth map as reference.



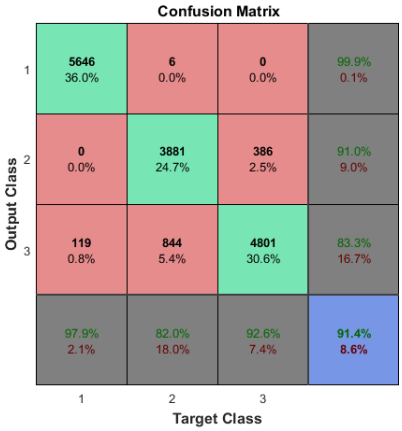
(a) Merged result of supervised classification



(b) Merged result of unsupervised classification



(c) Merged ground truth

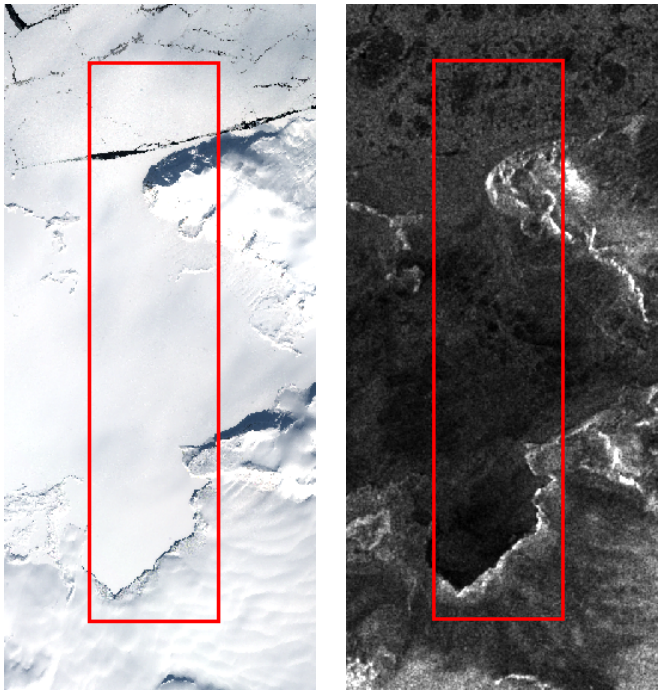


(d) Confusion matrix of merged results of unsupervised classification

Figure 6.18: Comparison of merging classes for supervised and unsupervised classification using a combination of marginals.

6.2.6 Unsupervised classification using SENTINEL data

We now move on to an Arctic setting. Since the original scene was very large, and the unsupervised classification not that fast, a small subset was selected, and can be seen in figure 6.19. The subset contains land, landfast sea ice and a lead that separates the landfast ice and a segment of more fragmented ice. Segmentation was performed on different sets of features. The first was using the amplitude and intensity from the HH and the HV band, green and blue and infrared, a total of seven features. Then, using the three transformed bands obtained from a PCA data transformation and dimensionality reduction (DTDR) using all 18 bands as input, that accounted for 99.99% of the variance. And finally, classification results using the SAR bands and the optical bands were tested separately. Individual results and summary are shown below. And note that we do not have any validated data from this area, so this is more a visual comparison.



(a) RGB image, subset enclosed in red rectangle

(b) HH Amplitiude band, subset enclosed in red rectangle

Figure 6.19: Images of the areas used in the SENTINEL segmentation

6.2.7 Comparison between 7 band raw features and 3 band PCA transformed features

Here, we wanted to check the difference between classifications using seven bands, and three feature bands resulting from a PCA. Ideally, we would wish to test using all 18 bands, but, due to the time consume anticipated with each iteration it was not deemed feasible at this time.

Implementation

The data that was used for the first case was the three transformed bands obtained from a PCA data transformation using the combination of the 13 bands of the SENTINEL 2, and the 4 channels of the SENTINEL 1. For the second case we used the 4 channels of the SENTINEL 1, and the green, blue and infrared bands of SENTINEL 2. The number of classes, K was $3, 4, 5, 6$ and 7. The error e was set to $1 \cdot 10^{-6}$ and maximum number of iterations, *maxiter* was set to 45. The algorithm was initialized using random seeding of the data.

Results

Classification results for the two cases, and for the different number of classes are shown in figure 6.20. Note that the class assignment will be different between initializations, and will rarely, if at all follow the same "numbering" as in the ground truth labels.

Discussion

What we see is that the PCA results are significantly more homogeneous in appearance. It also fails to find some of the ice structures that we saw in the amplitude image of the HH channel, shown in figure 6.21.

6.2.8 Classification using optical bands versus classification using SAR bands

Using the same subset as before, unsupervised classification was performed using Gaussian marginals. Four SAR bands was used in one trial, and the three optical bands in another.

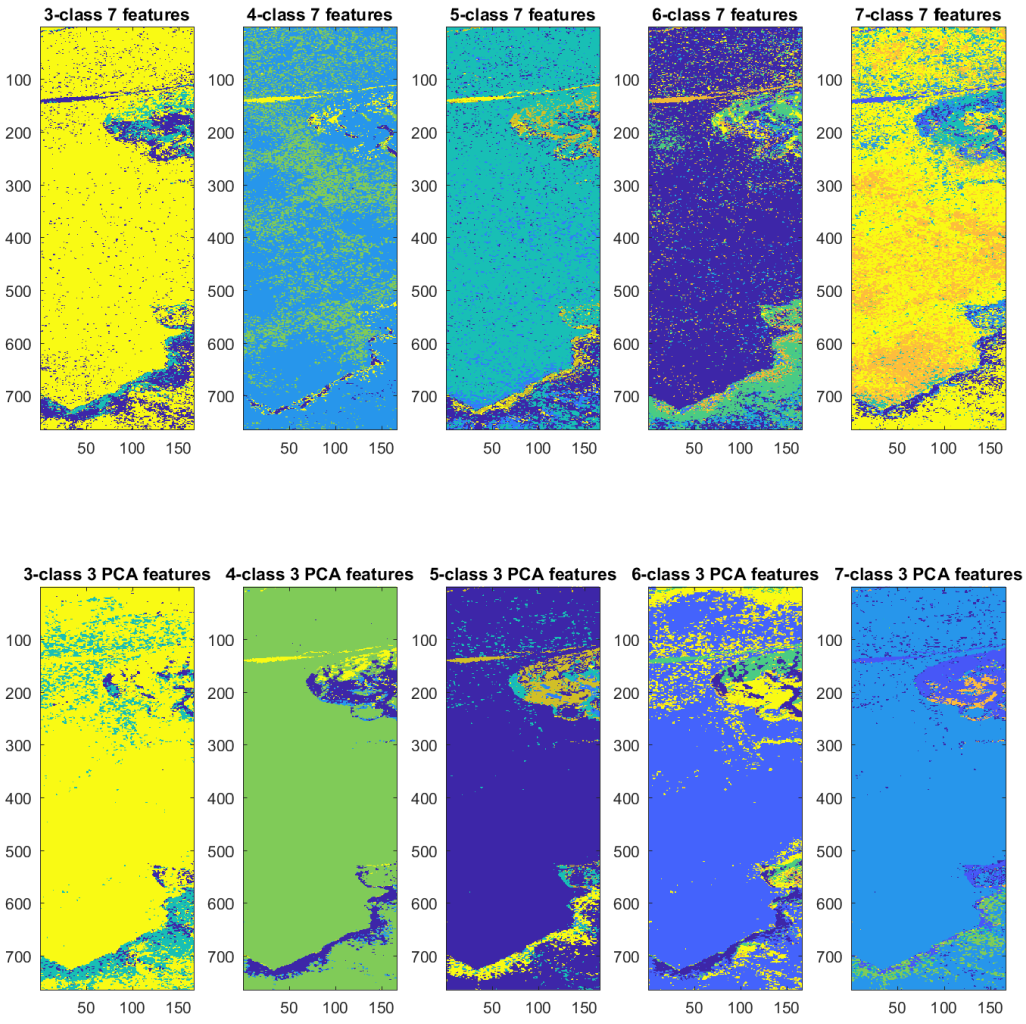


Figure 6.20: Comparison between classification using PCA and raw features. Top row shows classifications for the seven raw bands used, and bottom row classification using three PCA features

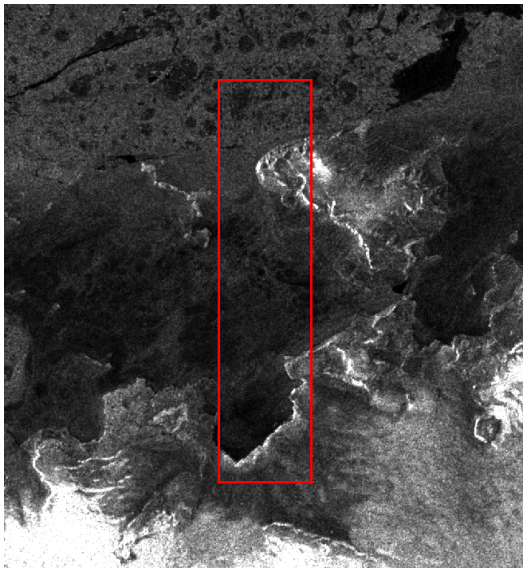


Figure 6.21: Amplitude image of the HH channel of SENTINEL 1, red rectangle indicates the area that is used for classification.

Implementation

The data that was for the first case the 4 channels of the SENTINEL 1, and for the second case the green, blue and infrared bands of SENTINEL 2. The number of classes, K was 3 The error e was set to $1 \cdot 10^{-6}$ and maximum number of iterations, *maxiter* was set to 45. The algorithm was initialized using random seeding of the data.

Results

Results of these implementations is shown in figure 6.22.

Discussion

What we first note is that using only three classes was in this case probably not enough, or it could have been if we had used a land mask before classifying. We see that especially in the segmentation based on the SAR backscatter values, the pixels corresponding to land dominate the segmentation. But, perhaps the only noteworthy feature in this comparison, is that of the open lead. It shows up clearly in the optical segmentation, but not in the SAR.

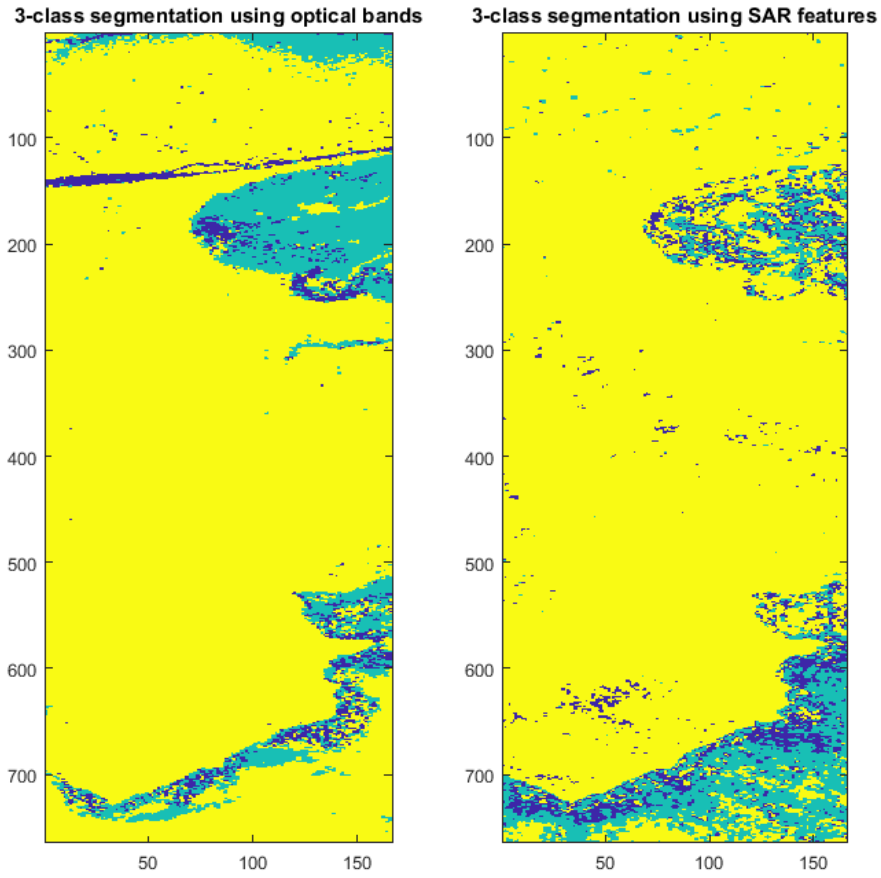


Figure 6.22: Comparison between 3 class segmentation using optical bands, and SAR bands. Segmentation based on optical bands is on the left, and SAR features on the right.

Chapter 7

Conclusion

This thesis started out by introducing some of the challenges and possibilities of remote sensing in terms of combining multi-sensor data. We presented the concept of data fusion, and introduced the Meta-Gaussian distribution.

The aim of this thesis was to build upon the previous work, [1], by Storvik et al, which was also recreated in the pilot project "Classification strategy for multi-sensor data using Meta-Gaussian distribution" [13] and extend the method to include:

- A clustering step to support unsupervised classification.
- Generalization of the marginal probability distribution functions(PDF). In many cases real data may not be well described by parametric models. In these cases, non-parametric, kernel based approximations of the PDF may prove to be better alternatives.
- Extensive testing on a multitude of data, both real and simulated.

Results from single-sensor classification and multi-sensor classification will be compared. The end goal would be to develop a multi-class classification algorithm based on the Meta-Gaussian data fusion method.

7.1 Conclusion

This thesis has been a test in both implementation and theory. Many hours have been spent troubleshooting code, and waiting for iterations to end. While's and if's and for's have accumulated, but in the end it all came through.

It is at the current stage not possible to say that this method of classification through a Meta-Gaussian transform is a contender to other established methods of data fusion and segmentation, but, it is definitely a working method, and one that is adaptable to any kind of data. It has so far proved to be versatile in terms of the marginals, and functional in an unsupervised setting as well. This has not to our knowledge been tested before in terms of using the Meta-Gaussian transform.

Segmentation using unlabelled data was also found to be possible, and functional, and a general unsupervised classifier was implemented successfully. Thus, it would not be unreasonable to say that we achieved our goal, which was to develop a multi-class classification algorithm based on the Meta-Gaussian data fusion method. Whether it can improve on current classification methods on general basis, will still need to be further validated. The small set of data used in this thesis can not be said to be versatile enough, and for true validation we would require more data, preferably with a corresponding ground truth, or other independent measurements. That being said, the results so far have been promising, and the method could prove to be a very suitable approach when combining and classifying multi-sensor data in the future.

We have also compared supervised maximum likelihood Bayesian classification, assuming multivariate normal distributions, with that of using the Meta-Gaussian, and found that the Meta-Gaussian approach is overall performing better. Kernel approximation of the marginals were found to produce slightly better results than parametric models when classifying on derived features. Overall the general parametric models that were tested performed within the same range as the kernel approximation.

7.2 Future Work

In this study it has been shown that supervised classification on multi-sensor data that has been fused using a Meta-Gaussian distribution can improve the accuracy when compared to single-sensor data, or data that has been fused on a rudimentary method. The current implementation did not parallelize the calculation- as a result the time spent for each iteration was directly related to the number of pixels, the number of bands used, and the number of classes. It was not within the scope of the thesis to look at optimization, and time has not allowed for it either, but the clustering method is highly

parallizable.
Other initialization methods of the unsupervised clustering should be looked into. The current method uses a random assignment of all points, assuming equal probability for the classes. An alternate method, using a smaller subset in the initial assignment and estimation, could give us more specific, or focused, marginal estimates to be used in the first segmentation. This could in turn help speed up the convergence. Where applicable, initialization based on parameter estimates could also be used, although that would be borderline supervised classification, at least if the estimates are based on some known values. And, more testing, on different types of data should be done.

Bibliography

[1] Storvik, B., G. Storvik, and R. Fjortoft. "On the Combination of Multisensor Data Using Meta-Gaussian Distributions." IEEE Transactions on Geoscience and Remote Sensing 47.7 (2009): 2372-379. Web.

[2] Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56(18), 3825-3833.

[3] "Google on the Forbes World's Most Valuable Brands List." Forbes, Forbes MagazineRetrieved July 14, 2017, from `https://www.forbes.com/companies/google/`

[4] Esa. "How to access data." European Space Agency, Retrieved July 23, 2017, from `http://www.esa.int/Our_Activities/Observing_the_Earth/How_to_access_data`

[5] U.S. Geological Survey. (n.d.). Retrieved August 7, 2017, from `https://www.usgs.gov`

[6] Campbell, J. B., & Wynne, R. H. (2011). Introduction to remote sensing. Guilford Press.

[7] Hoeffding, W. (1940),Massstabinvariante Korrelationstheorie, Schriften des Mathematischen Seminars und des Instituts fur Angewandte Mathematik der Universitat Berlin, 5, 181–233.

[8] Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8, 229-231.

[9] Embrechts, P., Lindskog, F., & McNeil, A. (2001). Modelling dependence with copulas. Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich.

[10] Schoelzel, C., & Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. Nonlin. Processes Geophys., 15(5), 761-772.

[11] Serinaldi, F., Bonaccorso, B., Cancelliere, A., & Grimaldi, S. (2009). Probabilistic characterization of drought properties through copulas. Physics and Chemistry of the Earth, Parts A/B/C, 34(10), 596-605.

[12] Mercier, G., Bouchemakh, L., & Smara, Y. (2007, July). The use of multidimensional copulas to describe amplitude distribution of polarimetric SAR data. In Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International (pp. 2236-2239). IEEE.

[13] Kvamme, A.B. (2017). Classification strategy for multi-sensor data using meta-Gaussian distribution, Project paper, University of Tromsø, Tromsø.

[14] AMAP, 2017. Snow, Water, Ice and Permafrost. Summary for Policymakers. Arctic Monitoring and Assessment Programme (AMAP), Oslo, Norway. 20 pp

[15] Arctic Council. Retrived July 23 from `https://www.arctic-council.org/index.php/en/learn-more/maps`

[16] Ice Information Portal. Retrived August 08 2017 from http://polarview.met.no

[17] Bogdanov, A. V., Sandven, S., Johannessen, O. M., Alexandrov, V. Y., & Bobylev, L. P. (2005). Multisensor approach to automated classification of sea ice image data. IEEE Transactions on geoscience and remote sensing, 43(7), 1648-1664.

[18] Scheuchl, B., Caves, R., Cumming, I., & Staples, G. (2001). Automated sea ice classification using spaceborne polarimetric SAR data. In Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International (Vol. 7, pp. 3117-3119). IEEE.

[19] Solberg, A. H. S., Jain, A. K., & Taxt, T. (1994). Multisource classification of remotely sensed data: fusion of Landsat TM and SAR images. IEEE transactions on Geoscience and Remote Sensing, 32(4), 768-778.

[20] Rockinger, O., & Fechner, T. (1998, April). Pixel-level image fusion: the case of image sequences. In Proc. SPIE (Vol. 3374, pp. 378-388).

[21] Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. Ann. Acad. Sci. Fennicae, AI, 34.

[22] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

[23] Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis. In Principal component analysis (pp. 115-128). Springer New York.

[24] Pohl, C., & Van Genderen, J. L. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applications. International journal of remote sensing, 19(5), 823-854. ISO 690

[25] Storvik, B., Storvik, G., & Fjortoft, R. (2003, December). Joint distributions for correlated radar images. In INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (Vol. 3, pp. III-2011).

[26] Baxter, R. A. (1998). Synthetic aperture radar image coding (No. JA-7535). MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB.

[27] G. Mercier, L. Bouchemakh, & Y. Smara. The use of multidimensional copulas to describe amplitude distribution of polarimetric SAR data. Proc. IEEE IGARSS, 2007, pp. 2236–2239.

[28] K. V. Mardia, J. T. Kent,&. M. Bibby, Multivariate Analysis. (1979). London, U.K.: Academic

[29] Yadolah Dodge.(2003).The Oxford dictionary of statistical terms. Oxford University Press

[30] Kuruoglu, E. E., & Zerubia, J. (2004). Modeling SAR images with a generalization of the Rayleigh distribution. IEEE Transactions on Image Processing, 13(4), 527-533.

[31] Solberg, A. H. S., Jain, A. K., & Taxt, T. (1994). Multisource classification of remotely sensed data: fusion of Landsat TM and SAR images. IEEE transactions on Geoscience and Remote Sensing, 32(4), 768-778.

[32] Lee, J. S., Grunes, M. R.,& Kwok, R. (1994). Classification of multilook polarimetric SAR imagery based on complex Wishart distribution. International Journal of Remote Sensing, 15(11), 2299-2311.

[33] Conradsen, K., Nielsen, A. A., Schou, J., & Skriver, H. (2003). A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data. IEEE Transactions on Geoscience and Remote Sensing, 41(1), 4-19.

[34] McDonald, J. B., and Xu, Y. J. (1995). A generalization of the beta distribution with applications. Journal of Econometrics, 66(1), 133-152.

[35] se.mathworks.com/help/stats/t-location-scale-distribution.html

[36] Theodoridis, S. and Koutroumbas, K. (2009). Pattern recognition. 4th ed. Amsterdam [u.a.]: Elsevier/Acad. Press.

[37] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3), 1065-1076.

[38] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3), 832-837.

[39] "Ksdensity." Kernel Distribution - MATLAB & Simulink - MathWorks Nordic, Retrived July 07 2017 from https://se.mathworks.com/help/stats/kernel-distribution.html

[40] Elachi, C., & Van Zyl, J. J. (2006). Introduction to the physics and techniques of remote sensing (Vol. 28). John Wiley & Sons.

[41] Chuvieco, E. (2016). Fundamentals of Satellite Remote Sensing: An Environmental Approach. 2nd edition. CRC Press.

[42] Scientific Background. Retrived July 29 from https://earth.esa.int/handbooks/asar/CNTR1-1-2.html

[43] Gagnon, L., & Jouan, A. (1997, October). Speckle filtering of SAR images: a comparative study between complex-wavelet-based and standard filters. In Optical Science, Engineering and Instrumentation'97 (pp. 80-91). International Society for Optics and Photonics.

[44] "QuickBird-2 - eoPortal Directory - Satellite Missions." QuickBird-2 - eoPortal Directory - Satellite Missions, retrived July 09 2017 from https://directory.eoportal.org/web/eoportal/satellite-missions/q/quickbird-2

[45] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[46] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

[47] Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.

[48] Salomonson, V. V., & Appel, I. (2006). Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results. IEEE Transactions on geoscience and remote sensing, 44(7), 1747-1756

[49] Doulgeris, A. P. (2013). A simple and extendable segmentation method for multi-polarisation SAR images.

[50] Moen, M. A., Doulgeris, A. P., Anfinsen, S. N., Renner, A. H., Hughes, N., Gerland, S., & Eltoft, T. (2013). Comparison of feature based segmentation of full polarimetric SAR satellite sea ice images with manually drawn ice charts.

[51] Esa. "Sentinel-1B launched to complete radar pair." European Space Agency, retrived July 17 from `esa.int/Our_activities/Observing_the_Earth/Copercenicus/Sentinel-1/Sentinel-1B_launched_to_complete_radar_pair`

[52] Ferro-Famil, L., Pottier, E., & Lee, J. S. (2001). Unsupervised classification of multifrequency and fully polarimetric SAR images based on the H/A/Alpha-Wishart classifier. IEEE Transactions on Geoscience and Remote Sensing, 39(11), 2332-2342.

[53] Yitayew,T.G. (2012). Multi-sensor Data Fusion and Feature Extraction for Forest Application, Master Thesis, University of Tromsø, Tromsø.

[54] Yitayew, T. G., Brekke, C., & Doulgeris, A. P. (2012, July). Multi-sensor data fusion and feature extraction for forestry applications. In Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International (pp. 4982-4985). IEEE.

[55] Crist, E. P., & Cicone, R. C. (1984). A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap. IEEE Transactions on Geoscience and Remote sensing, (3), 256-263.