Towards Unsupervised Online Band Selection in Hyperspectral Imaging

by

Gautam Balasubramanian

B.E. (Computer-Systems),
University of Technology, Sydney, Australia, 2003.
MSc. (Elec. Eng),
Monash University, Melbourne, Australia, 2005.

Thesis submitted for the degree of

Doctor of Philosophy

in

School of Electrical and Electronic Engineering
The University of Adelaide, Australia

December 2013

Typeset in LaTeX 2ε
Gautam Balasubramanian

To my Beautiful Wife, Loving Parents and Omnisicent Teachers

Contents

Abstract

This thesis explores the problem of unsupervised selection of a set of spectral wavebands in a hyperspectral sensor for a surveillance task. Selecting a subset of wavebands for surveillance has the advantage of reducing data throughput and hence network bandwidth requirements, computational complexity for processing the data and storage requirements in a ground-station. For the sensor designer, Signal-To-Noise Ratio and other sensor-band improvements can be made on those bands deemed critical for the surveillance task. In chapters 3 and 4, we propose the use of locally correlated high-dimensional Gaussian Mixture models to account for band overlap where maximum likelihood estimates of the parameters of such a model are provided using the SAGE-EM (Space Alternating Generalised Expectation Maximisation) algorithm. In both these chapters convex-relaxation strategies are proposed to handle the combinatorial complexity of selecting a subset-of bands that are locally correlated and contain non-Gaussian measurements. However, in chapter 4, we select bands according to anomaly detection criteria as opposed to modelling estimation accuracy (likelihood) as done in chapter 3. We breakdown the problem such that any pixel contains band measurements that belong to either an outlier or partial background distribution, where the distributions diverge across band-subsets in a Kullback-Leibler (KL) divergence sense. A pixel is deemed as an anomaly if it contains a certain number of outliers. We identify the bands that contain the most number of contiguous outlier measurements and also subsequently reveal the presence of anomalies. Finally, in the last chapter we solve the problem of online band selection for sub-pixel compositional hyperspectral models using a Bayesian approach. Online band-selection enables spectral-band cueing and automation for adaptive focal plane arrays where not all bands are used to measure each pixel. We apply beta process models to provide a recursive strategy to select bands based on prior knowledge of their utility as well as bands used in neighbouring pixels. Band utility is measured through convex-relaxation as the subset of bands that provides the best abundance estimation accuracy of training data. The

combination of a Gaussian process prior for possible end-members (pure materials) as well as a Gamma distributions for the abundance, enables efficient posterior sampling from a joint Normal-Gamma distribution. Furthermore, natural spectral band variations are retained making the model suitable for band selection, where approximate sum-to-one constraints are enforced through an intelligent update of the Gamma hyperparameters, based on the Dirichelet-Gamma relation. Experiments are conducted on synthetic Gaussian Mixture data with additive noise (Chapters 3, 4), Rochester Institute of Technology (RIT) Target Detection Test using the HyMAP sensor, (Chapter 4), synthetic sub-pixel data created using USGS spectral database [1] (Chapter 4) and AVIRIS-Cuprite dataset used by Mittelman et.al. in [2] (Chapter 4).

Statement of Originality

This work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of the thesis, when deposited in the University Library, being available for loan, photocopying and dissemination through the library digital thesis collection.

The author of this thesis acknowledges that copyright of published work contained within this thesis (as listed in the publications page) resides with the copyright holder(s) of that work.

Signed

Date

Acknowledgments

It is said that success has many fathers and failures but one. I measure success by personal growth rather than gain and in this sense I consider the journey thus far, a success. There are many contributors that have enabled this growth, one not greater or lesser in importance than the other both personally and academically.

Personally, my gorgeous wife, Prabha had to put up with me rushing to submit a conference paper on the first night of our marriage, alas it was the first of many more nights where her husband was missing, lost in equations within the confines of the study. Her enthusiasm for my work, belief in my abilities and passion for her husband's success made me work harder, gave me the constancy of effort required, she was always an un-ending source of encouragement and inspiration. She also instilled a sense of discipline I previously did not possess, an attribute of hers which I always hope to emulate. Thank You for the Umpteen Sacrifices, Love You heaps gorgeous. Next, my parents, who always gave me the freedom to pursue my own dreams throughout my Masters and now my PhD, selflessly, even though they knew it would keep me away from home, it must have been very difficult. Their presence always gave me perspective and reminded me to appreciate all facets of life and to never forgo responsibilities to wife, family and society although I am sure I did plenty of times. I'll always be eternally grateful for their presence in my life. I hope I have made you proud Mum n Dad. Finally, my mother-in-law and grandma who gave me constant nourishment, never shy from cooking up a storm with or without my behest also showing interest and enthusiasm on the outcomes of my work, where would I be without your kindness. I'll be eternally grateful for the sacrifice and faith of all my family members, this PhD is as much theirs as it is mine.

Academically, my principal Supervisor Professor Lang White, he's truly my first academic teacher from whom I have learned the most. His level of intelligence, comfort,

enthusiasm and care with subjects of Maths and Signal Processing was a sight to behold, a memory to treasure and I hope to emulate that throughout my career. Through many informal chats, emails and presentation of his own work he also gave me a great deal of structure to my thinking and improved my ability to frame problems which opened up creative avenues to solve problems and translated to a higher standard of work. Finally, his general niceness, fairness as a human being and compassion to myself and fellow students always made him a pleasure to be around. Secondly, Dr Jason Williams who is a Senior Research Scientist at DSTO: Jason was a force to behold. I tried picking-up many tid-bits, from the way he expressed himself mathematically on the whiteboard and paper, the way he attacked problems, his vast knowledge-base about a variety of different fields and unending thirst to gain knowledge. Jason also introduced the Machine Learning and Bayesian aspects to my work bringing to my attention the latest and greatest techniques relevant to problems pursued. His humility and work habits is something I always hope to emulate. My secondary Supervisors Dr Vittala Shettigara and Dr Tim Payne, bestowed the faith in me to pursue a PhD in-spite of our research group being short-staffed, remained steadfast with their commitment for which I'll always be grateful. Vittal, introduced me to the Band Selection problem and brought to my attention the significance of almost all the sub-problems pursued in this thesis always imparting the notion of presentation and completeness. Tim's timely feedback during the initial stages of my PhD also made me think deeper about framing Maximum Likelihood Estimation as a Convex Optimisation problem. Dr Sanjeev Arulampalam a Senior Research Scientist at DSTO and Professor Ian Fuss gave me the self-belief and faith to pursue the PhD and beyond, with vigour and to receive their encouraging words from people of stature made a striking difference to my work. Gratitude must also be extended to my peers at Adelaide Uni mainly Jiang He and Yassir. You made the journey very pleasant with your cheerful demeanour and inquisitiveness in my work. To the staff in the HSI group at DSTO, thank you for your patience and sacrifice in taking on a greater workload due to my constant absence.

Finally, to my Teachers at Chinmaya Mission, you set me off on this journey in 2003 and this work is but my humble offering to You. I hope it is worthy.

Gautam Balasuramanian (August 2013)

List of Figures

List of Tables

Chapter 1

Introduction

THIS chapter is the introduction to the thesis. It provides the motivation for the thesis, the problems addressed and their significance as well as key contributions made. It also introduces some background information on Hyperspectral Imaging to provide sufficient context for the rest of the thesis.

1.1 Overview

A Hyperspectral Image (HSI) provides a rich description of an observed scene by capturing signals reflected at hundreds of narrow contiguous wavelengths or frequencies in the electo-magnetic spectrum, where a unique narrow range of wavelengths is referred to as a frequency band or spectral band. A hyperspectral sensor is made up of hundreds of sensor elements, where each element measures the incoming radiance signal reflected by materials on ground, at a unique frequency band. Strides made in sensor development technology have made hyperspectral sensing a realistic tool to characterise natural and man-made material in-terms of their spectral properties owing to the narrow frequency bandwidth. Increase in computer processing power and network bandwidth over the last 20 years has enabled airborne and space-borne collection of vast spatial regions and dissemination of data on-ground for processing and evaluation. Both these factors have opened up applications in land-cover mapping in the agriculture sector, mineral mapping in the mining sector and target detection in defence surveillance to name a few. All these communities have seen rich-payoffs [3] for their problem at-hand and continue to seek improvements in their respective hyperspectral systems. From a systems-engineering standpoint, these systems typically consist of the sensor payload, an airborne platform (satellite or aircraft), computer processors and memory, networking interface and bandwidth to disseminate data to a ground station consisting of image analysts and their respective workstations. Furthermore, the system may also include spectral analysts who collect spectral signatures using spectrometers on-ground, prior to data collection to verify their result or as part of their analysis.

The practical applicability of hyperspectral systems across these sectors as well as prospective new ones is therefore restricted by system cost, timeliness in task completion and robustness which also suffers due to system complexity. Further applications such as persistent surveillance used for National Security and Defence [4] face obstacles such as data bottlenecks which introduce latencies in processing over-burdening

image and spectral analysts and thus restricting their capacity to deliver a timely result. Since most hyperspectral sensors are made up of more than a hundred frequency bands. The requirements to capture, store and disseminate data is a resource intensive exercise. Naturally, storage requirements are greater when the number of bands to be transmitted is large. Spectral bandwidth required to disseminate data to ground stations is also great and processing of the data to address surveillance problems is a computationally demanding exercise. The computational demands also place time constraints on how quickly a problem can be solved which is crucial in time-critical situations which is the case especially in Defence and National Security. Modern systems, carry out a certain degree of on-board processing [5] before the data can be disseminated. Such systems thrive on unsupervised processing where user intervention is minimal. If the data can be reduced whilst preserving useful information before dissemination and further processing takes place, computational and monetary requirements on hyperspectral systems will be reduced a great deal. Furthermore, such automated paradigms enable analysts to focus on higher-level analysis and inference tasks which cannot be easily automated.

In this thesis, we propose a methodology that reduces, analyst overload, improves reliability and also potentially reduces system cost. We show that this can be achieved through the unsupervised selection of frequency bands according to maximum likelihood, anomaly detection and sub-pixel mixing criteria. The rationale behind intelligent reduction of bands is that it reduces computational burden, processing and network bandwidth requirements when it is carried out prior to collection. For the Defence community, unsupervised band selection according to anomaly detection criteria not only reduces the number of spectral measurements disseminated to the analyst but also the number of pixels. We propose a paradigm where only anomalies across a selected number of frequency bands need to be disseminated which reduces time constraints and data complexity burden of the analyst. Since this proposed paradigm is unsupervised it reduces human resources required to carry out this task. Finally, we extend our rationale to low Signal-To-Noise (SNR) scenarios, resulting from a coarse sensor resolution and/or high platform altitude, the use of compositional sub-pixel

models are beneficial under such conditions. We propose a system to cue a subset of sensor bands online as the sensor gathers data based on prior knowledge of the scene and bands used to capture signals across previous pixels. We believe that proposed methods offer an exciting alternative for the development of future hyperspectral systems and the thesis offers a proof-of-concept, through novel generative models, problem formulation and optimisation techniques.

1.2 Hyperspectral Imaging (HSI)

An airborne hyperspectral sensor measures the energy reflected or emitted off the earth's surface by decomposing the incoming radiance signal into hundreds of snapshots of a spatial region of the ground. In hyperspectral sensing, each snapshot refers to a frequency band which in turn represents a narrow wavelength range in the order of 5 to 100 nanometers. The frequency bands are unique, contiguous, and span the visible (VIS) to short-wave-infra-red (SWIR) regions of the electro-magnetic spectrum which in-terms of wavelength is a range of 0.4 - 2.5 μm. The fine spectral resolution reveals subtle changes in the chemistry of visually similar materials on the ground and acts as a means of material identification which is un-available in other commonly known remote sensing modalities. We briefly define some terms which are used in the remainder of the study:

Definition 1: The collection of snapshots acquired at unique, contiguous frequency bands across a spatial region form *a hyperspectral image.*

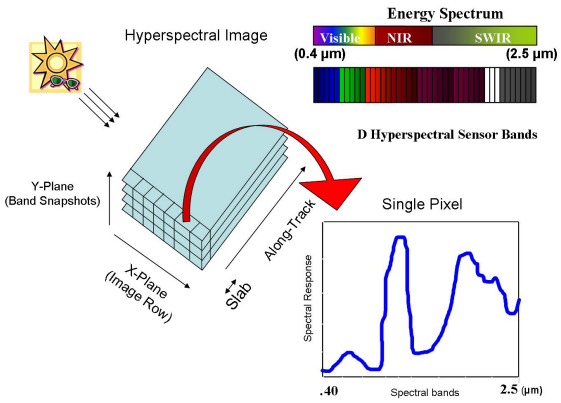
Definition 2: A *pixel* in the image represents an area on the ground and depending on the spatial resolution may contain one or many objects. A hyperspectral image consists of thousands of pixels.

Definition 3: A *Spectral measurement* refers to the at-sensor-radiance representing a single pixel. The spectral signal measured is a vector consisting of hundreds of co-ordinates, where each co-ordinate represents the energy reflected or emitted at a unique frequency band.

Definition 4: *Band measurements* refer to all measurements in a spatial region gathered by the specified frequency band.

We describe the image acquisition process briefly and subsequently describe the most popular method used for gathering hyperspectral measurements, pushbroom sensing [3]. Alternative modes of image acquisition such as electronically tunable focal plane arrays are discussed in the final chapter since they have an implication on the signal processing methods proposed in that chapter. The multi-dimensional signal that represents the pixel is a function of the instrument response of a 2-D detector array in the sensor, the sensor optics and the input irradiance (energy at the sensor). Measurements gathered across the X-Y co-ordinates of the 2-D detector array represents the detector's spectral response both in a spatial and spectral sense. Measurements across the X-direction correspond to a row of pixels whilst measurements in the Y-direction correspond to spectral response across contiguous frequency bands. Whilst the sensor optics split incoming signals across multiple frequency bands, the detector array integrates the at-sensor-irradiance in the order of micro-seconds converting them into voltages and subsequently digital numbers (a single number at each frequency band for each pixel). A new "slab" (refer to Fig.4.1a.) of measurements is captured by the detector array as the sensor-platform moves along-track forming a hyperspectral image, one row at a time. Some prominent aspects of hyperspectral sensing relevant to this thesis include: spectral measurement correlation, sub-pixel mixing and noise sources.

Spectral correlation refers to (1) correlation of neighbouring spectral responses in the detector array due to the narrow frequency bandwidth. The extent of overlap and correlation is specified by the sensor design and is known prior to measurement. (2)



(a)

Figure 1.1. a) A slab of measurements are recorded by multiple frequency band snapshots as the sensor platform moves along track. The X and Y axes correspond to the 2D detector array. The multi-dimensional radiance signal recorded across multiple bands, for a single pixel is also shown. The acronyms VIS refers to Visible, NIR refers to the near-infrared region of the electro-magnetic spectrum and SWIR refers to the short-wave-infrared region.

Unknown frequency band correlation is also present in each signal due to material chemistry [6] which varies according to the material present in each pixel. This can be seen in the smoothness in the spectral response across multiple bands, in the experiments conducted in Chapters 4 and 5. In this thesis, both sources of correlation are considered. In chapter 4, local correlation due to the sensor is built into a generative Gaussian mixture model. In chapter 5 we also capture correlation due to the material chemistry of a known class of materials using the prior covariance obtained from a Gaussian Process.

Sub-pixel Mixing refers to the nature in which pixels in a scene are distributed. The rationale behind hyperspectral sensing is to build spectral richness by sampling across narrow frequency bins as opposed to improving the spatial resolution (also referred to as the Ground Sampling Distance - GSD). The GSD can vary from a fraction of a meter to tens of meters. As a result, in a natural scene pixels may be a combination of multiple material classes such as tree, grass and soil, at varying proportions. The spectral

measurement in each pixel can be considered as a linear combination of the materials with some additive noise. The proportion or abundance of each material present, reflects the spatial extent to which it occupies the pixel. This behaviour is not captured by the Gaussian mixture model, where each measurement is assumed to be a representative from a single material class. A compositional model best describes the sub-pixel mixing that occurs where the spectral response received is a convex combination of up-welling radiance of multiple materials within a pixel. Typically, the constituent materials within a pixel are representative of unique classes of materials referred to as *endmembers*. Nonetheless, in chapter 4 a Gaussian Mixture model is used as a simplified representation of sub-pixel data due to advantages in analytic tractability in the inferring parameters of a Gaussian Mixture model with a fixed number of components. However, in chapter 5 we use a sub-pixel compositional model to represent the scene, thereby building a richer model but losing analytic tractability in the process.

Spectral variability and Noise Sources There are many sources of noise in the spectral responses of pixels which in combination with the coarse spatial and rich spectral resolution provide a wide gamut of spectral responses characterising a scene. Some sources of scene noise include atmospheric absorption and scattering of radiation, water vapour and aerosol conditions, adjacency effects of neighbouring sensor elements and nonlinear motion of sensor leading to sensor artifacts. Spectral variability refers to the variation in the spectral responses of pixels that belong to the same material class. This is an outcome of un-even terrain orientation, radiation scattering and reflection from nearby pixels, shadows in the scene and seasonal variations. Both spectral variability and noise sources motivate the inference of statistical probabilistic models to represent hyperspectral data as opposed to use of the raw measuements. Spectral variability and non-homogeneity of material classes in the scene are captured by Gaussian mixture and compositional models where spectral variability and scene noise are built into the statistical models.

1.3 Hyperspectral Band Selection

Traditionally, hyperspectral band selection is designed to reduce redundant measurements as well as noise gathered during the measurement process. When used, manually, automated or both, band selection is often used as a pre-processing step after the sensor gathers data across all bands. Typically, it is carried out by an analyst before they use an eigen-decomposition technique such as Principal Components Analysis (PCA) or Minimum Noise Fraction (MNF) [7] to reduce the data dimensionality and use the the K most significant eigenvectors. This has been found to improve both classification [7] as well as detection performance [8]. Other techniques which preserve physical band information vary according to performance criteria and extent to which they are automated. Most techniques also assume some prior knowledge about the materials within the scene using a spectral library that contains representative signatures of each material class. However, these methods are typically conducted post-collection by the image analyst which does not address network bandwidth, memory or time constraint issues stated earlier. A clever future hyperspectral sensor may collect data across a subset of bands that provides the most useful information in a scene, where the band subset varies from scene to scene [9]. This not only reduces power and energy requirements of the sensor but also allows electro-optics engineers to improve signal-to-noise ratio on those bands that are likely to be more useful. Hypothetically, another application maybe that, we cue certain bands to gather measurements for certain parts of the scene, the intention again here is to reduce the amount of data collected and improve that ability to produce a faster surveillance outcome. Such an adaptive system is already being proposed by DARPA [9]. Finally, consider a surveillance problem where the objective is to find man-made materials amongst natural background, but there are no prior spectral signature(s) available of the man-made material. In hyperspectral sensing, man-made materials are known to vary from natural background across a certain distinct wavelength range of the energy spectrum. If this knowledge is available for a geoscientist or spectral analyst and a set of useful bands identified fall within the distinct wavelength range, we can then confirm the presence of that man-made material. Standard benchmark band-reduction schemes such as PCA alone does

not provide such insights about the bands. PCA transforms the data such that maximum information content lies in an ordered set of principal components. This enables data compression, since the latter set of components which contain the least information can be discarded. However, the physical band-structure of data is altered, which means we do not have any information regarding the utility of each band, in terms of its contribution to a surveillance task.

1.4 Problems Addressed

We address three problems in this thesis:

1. Unsupervised Band Selection to improve Model Estimation Accuracy of a Scene
2. Inferring Appropriate Bands to find Anomalies in the Scene
3. Pixel-by-Pixel Online Band Selection for Band Cueing using Sub-Pixel Mixing Criteria

1.4.1 Unsupervised Band Selection to improve Model Estimation Accuracy of a Scene

For many surveillance scenarios where the objective is to find targets or anomalies, one of the key steps is to estimate a probabilistic model for the scene background. These include Gaussian [8], Gaussian Mixture [10], [11] and Elliptic-T distributions as used by Manolakis *et. al.* in [12] and Theiler *et. al.* in [13]. Band noise is known [12], [14] to have adverse effects on model estimation accuracy even when using just the principal eigenvectors. For Gaussian Mixture backgrounds, the unsupervised removal of noisy bands whilst statistically guaranteeing improvements in model estimation accuracy is an unsolved problem within HSI. This is because it requires clustering in a high-dimensional space, where the number of bands in many hyperspectral sensors exceed well over a hundred. We carry out clustering using the Space-Alternating Generalised Expectation (SAGE) condition in-conjunction with the standard Expectation-Maximisation (EM)

algorithm to guarantee maximum-likelihood convergence of model likelihood. Band selection is carried out using a relaxation of a binary band-utility indicator variable to overcome combinatorial complexity. The problem is formulated as a convex optimisation problem with convex inequality and linear equality constraints and solved using semi-definite programming [15]. The formulation guarantees the best possible model estimation accuracy in relation to the subset of bands selected. The method is restricted by the fact that the analyst needs to specify the number of bands he/she wishes to preserve. Nonetheless, the proposed method does not rely on any prior knowledge of the scene or materials within the scene and provides an unsupervised generic pre-processing step to reduce the number of bands. We conduct tests with synthetic Gaussian Mixture data and leave all testing with real data to subsequent chapters since the noise removal claims of the chapter are sufficiently validated with synthetic data and also that Gaussian Mixture models are well-known descriptors of HSI data.

1.4.2 Inferring Appropriate Bands to find Anomalies in the Scene

Unlike the previous problem, we seek bands that reveal the presence of anomalies in the HSI scene. Therefore, the problem addressed here is the simultaneous identification of these critical bands as well as anomalies in the scene to restrict the data throughput from the sensor to just the anomalies across these critical bands. This provides a drastic reduction in computational complexity and means that potential targets can be verified by tools such as spectral matching just using anomalies and a reference library. This is relevant for many surveillance scenarios, where the scene captured contains a large number of natural background pixels relative to a sparse number of man-made anomalies which are of interest. Identifying the necessary bands that reveal their presence also promotes an improvement in SNR across these bands in future sensor-design. Furthermore, properties of man-made materials across certain regions of the infra-red spectrum is well known but the capacity to utilise this knowledge in bands selection is largely unknown. We develop a method using EM and Convex Relaxation (as the previous problem), but formulate a model such that we find anomalies and critical bands

according to a distributed Kullback Leibler (KL) divergence measure across band subsets. Thus we carry out simultaneous band selection and detect anomalies whilst also enabling the analyst to impose prior knowledge about potentially useful bands. It must be noted that the procedure can be implemented on-board an aircraft or satellite before the data is transmitted to ground stations for dissemination and analysis. We believe the procedure significantly reduces the processing burden on the analyst. Unlike the previous method there is no need to specify the number of useful bands apriori but we do assume that data is collected across all bands as per convention. We conduct tests with both synthetic Gaussian Mixture data generated across 30 bands and real HSI data obtained from the RIT Target Detection blind Test dataset [16] [17] using a HyMap airborne sensor with 126 bands.

1.4.3 Pixel-by-Pixel Online Band Selection for Band Cueing using Sub-Pixel Mixing Criteria

The advent of electronically tunable focal plane arrays referred to as Adaptive Focal Plane Arrays (AFPAs) [9] have meant that bands can be adaptively cued to measure the content of each pixel improving both data storage and computational requirements of conventional HSI. Since materials exhibit different properties across frequency bands each pixel may be sufficiently described by a different set of bands. In many land-cover mapping scenarios, each pixel contains multiple materials and more complex sub-pixel compositional models are required to provide an accurate model. Nonetheless, the problem is alleviated by prior knowledge of possible materials in the scene as well as their spectral signatures in the form of stored spectral libraries. The problem addressed, is the ability to cue appropriate bands to gather measurements such that sub-pixel un-mixing performance is retained as per use of the full set of bands. In this methodology, the only output that is required from on-board processing in the aircraft is the abundance fraction of each material at each pixel, with band selection carried out online and implicitly as the data is received. This removes the need for end-member extraction procedures carried out post-processing and promotes the notion of online HSI. We provide a Bayesian treatment of the spectral un-mixing problem using convex

optimisation to estimate band utility on training data, Gaussian processes to capture the natural variations across endmembers and non-parametric Beta processes to select the bands online as the data is gathered by the sensor. We formulate the posterior probability of the endmembers, abundances and band selection variables in the test image and develop a Gibbs sampling algorithm to estimate these parameters. Tests and sub-pixel mixing performance comparisons are conducted with synthetic USGS [1] data and real HSI data acquired from the AVIRIS sensor with 226 bands and used by Mittelman *et. al.* in [2].

1.5 Contributions and Publications

Contributions made in this thesis are divided into practical and theoretical:

Practical

1. Unsupervised Band Selection algorithm for removing Band Noise: Chapter 3
2. Anomaly Detection algorithm using Convex Optimisation and EM: Chapter 4
3. Unsupervised Band Selection algorithm using Anomaly Detection criteria: Chapter 4
4. Online Spectral Unmixing algorithm using Gaussian Processes and Gamma Distributions: Chapter 5
5. Online Band Selection using Convex Relaxation and Beta Processes: Chapter 5

Theoretical

1. Application of SAGE-EM structure to HSI data without any modification to the measurement set.
2. Application of Convex Relaxation for Measurement Selection in Un-Categorised Gaussian Mixtures.
3. Formulation of Simultaneous Anomaly detection and Unsupervised Band Ranking problem using Convex Relaxation and distributed Likelihood Ratio Tests.

4. Proof that the solution to the problem is equivalent to maximising a KL divergence term between outliers and partial background distributions.

5. Application of a Gaussian Process to estimate endmember mean and covariance modelling spectral correlation of hyperspectral signals.

6. Use of Dirchelet-Gamma Relation for posterior hyperparameter update of abundances which improves sparsity as well as pixel likelihood.

7. Deriving an abundance sampling technique that is equivalent to sampling from each element of Dirichelet distribution using a non-linear transformation of random variables.

8. Derivation of a Gibbs-Sampler to estimate posterior probabilities of endmember and abundances using a sparse number of bands.

Publications

1. Inferring Appropriate Bands to Find True Anomalies: Submitted to IEEE Transactions in Signal Processing (Under Review)

2. Band Sparsity for Compositional Models in Hyperspectral Imaging: Journal Paper In Preparation

1.6 Thesis Structure

The thesis can be broadly split up into the following, (1) the assumed generative model of the spectral measurements (2) the band correlation structure (3) and the scoring criteria for selecting bands. We begin by imposing just the local sensor band-correlation and using model estimation accuracy as the band scoring method to select a known number of bands under a Gaussian mixture model in chapter 3. Using a similar locally correlated Gaussian mixture model, we develop a band scoring method that uses an anomaly detection criteria for selecting bands without any prior knowledge of the possible number of useful bands in chapter 4. In the final component of the study, we

assume sensor as well as material dependant correlation whilst assuming sub-pixel mixing and hence apply a compositional model rather than a Gaussian mixture model in chapter 5. We apply two independent band-scoring measures to carry out online band selection on training and test data, where the band scoring procedure using the training set provides forms discrete prior band utility for the test set. In a broad sense, the systems-engineering approach of gradually building model complexity highlights the role that band-structure plays in hyperspectral surveillance which is important in this thesis. Pre-requisite knowledge required for the thesis is provided in chapter 2. Final conclusions and some thoughts for future work are elaborated in the final chapter.

Chapter 2

Preliminaries

THIS chapter covers the technical background material required to follow the mathematical argument in the chapters to follow in terms of providing a background for the techniques introduced as well as introducing mathematical notation.

2.1 Introduction

In this chapter, we introduce the technical background that is necessary to understand and follow the remaining chapters. Our focus is on parameter estimation, where the probabilistic models relevant for hyperspectral measurements motivate a variety of parameter estimation techniques. This section can be divided into four parts - two for the statistical models used and two for the band selection techniques applied. First, we introduce the two statistical models to represent hyperspectral data in this thesis (1) Gaussian mixture models and (2) compositional or sub-pixel models. We introduce parameter estimation techniques that are used to infer parameters for both models. Since the rationale behind the thesis is to preserve the original frequency bands and avoid rotating the measurement axes, we introduce two distinct measurement selection approaches which reduce hyperspectral measurement complexity for the models used to describe them. These include (3) convex optimisation and (4) stochastic beta processes. The remaining chapters in the thesis describe the varying ways in which these statistical models and parameter estimation techniques are applied for a variety of band selection problems. In this chapter we establish the foundations of the parameter estimation techniques as well as how they are motivated by the proposed model. The proposed models to describe HSI data gradually increase in their complexity and techniques to infer parameters graduate from a non-Bayesian to a Bayesian form. This also applies to band selection, where the optimal number of frequency bands are treated as a random variable when using the Stochastic Beta Process as opposed to deterministic parameters which is the case when using convex optimisation techniques.

2.2 Gaussian Mixture Models

2.2.1 Parameter Estimation - Maximum Likelihood

Consider an $\mathbb{R}^D$ valued random variable, Y with probability distribution function $P_Y(.;\theta)$ parameterised by a vector $\theta \in \Theta \subset \mathbb{R}^q$. Let $y_1, \ldots, y_N$ denote independent and identically distributed samples of Y. Suppose that Y is a continuous random variable, so

that it has a probability density function (pdf) $p_Y(.;\theta)$. The likelihood of the parameter θ is the product of individual samples from the p.d.f, $p_Y(.;\theta)$,

$$L(y_1 \dots y_n;\theta) = \prod_{n=1}^{N} p(y_n;\theta), \tag{2.1}$$

where Y is assumed known and fixed. A standard procedure for estimating the parameters θ when they are regarded as deterministic and unknown is Maximum Likelihood estimation, and the resulting estimate is given by

$$\hat{\theta}^{ML} = \arg\max_{\theta\in\Theta} L(y_1 \dots y_N;\theta) \tag{2.2}$$

where, $\hat{\theta}^{ML}$ is considered to have good asymptotic properties when N is large [18]. An important class of probabilistic models we shall use to characterise HSI data is Gaussian Mixtures which have the p.d.f,

$$p_Y(;\theta) = \sum_{k=1}^{K} \mathcal{N}(\mu_k, \Sigma_k)\pi_k \tag{2.3}$$

where, Y represents N pixel measurements or observations, $y_1 \dots y_N$, π_k represents the proportion of measurements described by the kth Gaussian component, $\mu_k \in \mathbb{R}^D, \Sigma_k \in \mathbb{R}^{D\times D}$ are the kth Gaussian component mean and covariance. The likelihood for the Gaussian Mixture model is given by,

$$L(y_1 \dots y_N;\theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} \mathcal{N}(y_n;\mu_k, \Sigma_k)\pi_k \tag{2.4}$$

Maximising the likelihood (2.4) over the set of parameters $\theta = (\pi_1, ..., \pi_K, ...)$ is a constrained maximisation problem (the weights need to be non-negative and sum to one and the covariance matrices need to be positive semi-definite) for which an analytic solution can't be found and which is also numerically intractable [19]. The effective total number of parameters is (K-1) for the mixture weights, incorporating the sum-to-unity constraint, KD for the means and $KD(D+1)/2$ for the (symmetric) covariance matrices. Hence, the problem is intractable and an analytical solution cannot be derived.

2.2.2 Maximum Likelihood Estimation of Gaussian Mixtures via Expectation-Maximisation

A tractable solution for maximum likelihood estimation of the parameters of a Gaussian mixture from i.i.d. samples, however can be achieved through iterative means. Consider the log-likelihood for the Gaussian Mixture,

$$\ell(y_1 \dots y_N; \theta, V) = \log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{v_{n,k}} \mathcal{N}(y_n; \mu_k, \Sigma_k)^{v_{n,k}} \tag{2.5}$$

where, $v_{n,k} \in \{0,1\}$ is an indicator variable that indicates whether the *nth* measurement is a member of the *kth* Gaussian component. The problem is deemed *in-complete* or *un-categorised* if $v_{n,k} \forall n = 1 \dots N, k = 1 \dots K$ is unknown for all measurements. The *complete* set of random variables is denoted by $Z = \{Y, V\}$, where probability distribution,$P_Z(\theta) \equiv P_{Y,V}(\theta)$. Given that, V is unknown, and Z in-complete, analytic maximisation is carried out using the expected log-likelihood of the expression in (2.5), where

$$\begin{aligned} \ell_E(y_1 \dots y_N; \theta, W) &= \mathbf{E}_V\{L(y_1 \dots y_N; \theta, V)\} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} w_{n,k} \log(\pi_{n,k}) + \sum_{n=1}^{N} \sum_{k=1}^{K} w_{n,k} \log \mathcal{N}(y_n; \mu_k, \Sigma_k) \end{aligned} \tag{2.6}$$

where, $w_{n,k} = \mathbf{E}\{v_{n,k}\} \in [0,1] \forall n = 1 \dots N, \forall k = 1 \dots K$ is the component membership. The expected log-likelihood is the lower-bound [20] of the complete likelihood where a closed form expression to compute θ is derived using estimates of the expected value, W. Iterative calculations of both W, θ form the Expectation-Maximisation or EM algorithm involve solving the following optimisation problem,

$$\hat{\theta}^{(m+1)} = \arg\max_{\theta \in \Theta} \ell_E(y_1 \dots y_N; \hat{\theta}^{(m)}, W^{(m)}) \tag{2.7}$$

where, parameter estimates are computed each $m+1$th iteration using the mth approximation of each measurement's membership to that component. These two steps are respectively referred to as the respectively as the M-step and E-step, and are repeated until analytic convergence of likelihood is reached. Its relatively easy to show [20] that the E and M steps are given by,

E Step:

$$w_{n,k}^{(m)} = \frac{\mathcal{N}(y_n; \hat{\mu_k}^{(m-1)}, \hat{\Sigma}_k^{(m-1)})\hat{\pi}_k^{(m-1)}}{\sum_{k=1}^{K}\mathcal{N}(y_n; \hat{\mu}_k^{(m-1)}, \hat{\Sigma}_k^{(m-1)})\hat{\pi}_k^{(m-1)}} \tag{2.8}$$

M Step:

$$\hat{\theta}^{(m+1)} = \arg\max_{\theta\in\Theta} \ell_E(y_1 \dots y_N; \hat{\theta}, W^{(m)}) \tag{2.9}$$

where,

$$\begin{aligned} \hat{\pi}_k^{(m+1)} &= \frac{1}{N}\sum_{n=1}^{N} w_{n,k}^{(m)},\ \hat{\mu}_k^{(m+1)} = \frac{\sum_{n=1}^{N} w_{n,k}^{(m)} y_n}{\sum_{n=1}^{N} w_{n,k}^{(m-1)}} \\ \hat{\Sigma}_k^{(m+1)} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} w_{n,k}^{(m)} AA^T \end{aligned} \tag{2.10}$$

where, $A = (y_n - \mu_k^{(m+1)})$. The EM algorithm guarantees that the model likelihood increases during each iteration when we are maximising the expected log-likelihood. However, the algorithm suffers from convergence rate issues as the data dimensionality grows. It is shown in [19] both theoretically and via simulations that asymptotic convergence of parameter estimates suffers as data dimensionality D, grows. Barber et. al in [21] and Neal *et. al.* in [22] provide stochastic variants, which would provide faster convergence rates for high-dimensional datasets. However this comes at a cost of asymptotic properties offered by the standard EM. In this thesis, we wish to preserve the asymptotic property and hence continue to use the standard EM algorithm.

2.2.3 Space Alternating Generalised Expectation (SAGE)

Given that $Y \in \mathbb{R}^D$, when D is large, parameter complexity also grows, asymptotic properties are not preserved under such instances and also result in slow convergence rates which was found to be the case when $D > 10$ in simulations conducted in this thesis and previous work [23].

In HSI D corresponds to the number of bands. To alleviate this problem , Marden *et. al.* in [11] opt to reduce the number of bands by mapping the data to its principle eigenvectors before implementing EM. Alternatively, the HSI surveillance community describes

the data using heavy-tailed Elliptic T distributions rather than Gaussian Mixtures to avoid the computational challenges in accurately estimating mixture parameters [12]. Elliptic Contoured pdfs can be considered as heavy-tailed Gaussian pdfs, where an additional parameter is introduced to control the heaviness in the tail. This pdf captures the large measurement variability rather than an additional Gaussian component.

Whilst the standard EM algorithm requires that Y is fixed throughout the E and M steps, Fessler *et. al.* in [24] introduce the SAGE condition which enables independent maximisation of parameter subsets, fast convergence rates and asymptotic convergence of parameter estimates. Consider a sequence of R independent random variables, $Y_1 \dots Y_R$, where the rth random variable $Y(r) \in \mathbb{R}^{Q(r)}$, its probability distribution is denoted by $P_{Y(r)}(.;\tilde{\theta}(r))$ and $y_1(r) \dots y_N(r)$ denote the N iid samples of $Y(r)$. The probability distribution of Y is given by, $P(Y) = \prod_{r=1}^{R} P_{Y(r)}(.;\tilde{\theta}(r))$, and parameterised by $\tilde{\theta}(r) \subset \theta \in \mathbb{R}^m$. If $y_1(r) \dots y_N(r) \forall r = 1 \dots R$ represent the total Fessler *et. al.* show that asymptotic properties of the maximum likelihood estimator can be achieved through independent maximisation of $\tilde{\theta}(1) \dots \tilde{\theta}(R)$ as long as $Y(r)$ is independent $\forall r = 1 \dots R$ and $\forall n = 1 \dots N$.

Guaranteed asymptotic convergence improves our confidence in the predicted model and in further decisions made based on the derived model. Such guarantees exist only because of analytic convergence properties of EM and maximum likelihood estimation in a non-Gaussian setting. Furthermore, the computational complexity of the problem is reduced since we are maximising over parameter subsets rather than the entire set. Fessler. *et. al.* in [24] claim faster convergence and demonstrate it on large computerised tomography data. For HSI Surveillance problems, the SAGE property can be applied in a weak sense in two instances. Firstly, we can assume that pixels are more or less spatially independent, where adjacency effects are negligible as discussed in the previous chapter. We can also assume that R band-subsets are conditionally independent to each other, where the influence of material chemistry across band subsets evident in the smoothness of signatures is not explicitly modelled under this framework. We model the material chemistry explicitly in 5.3 using Gaussian Processes. We

apply the SAGE condition in chapters 3 and 4, both in a spectral and a spatial sense, where Y represents the total measurement set.

2.3 Compositional Models

The parameter estimates for a Gaussian mixture model describe the behaviour of N measurements and thus remain fixed across N measurements. Whereas, in a compositional model, the parameter estimates can vary for each nth measurement. Under this model each measurement is described as a convex combination of source signals with some additive noise. This signal model for N measurements is described by the following affine model with sum-to-one and inequality constraints,

$$\begin{aligned} y_n &\sim \textstyle\sum_k G_k X_k + W \\ \text{s.t. } x_{k,n} &\geq 0, \textstyle\sum_k x_{k,n} = 1 \forall n = 1 \dots N \text{ measurements,} \end{aligned} \tag{2.11}$$

where, $G_k \in \mathbb{R}^D$ is a random variable describing the kth source signal out of $k = 1 \dots K$ possible sources describing N measurements, $X_k \in \mathbb{R}$ describes the source signal weight and W corresponds to the random variable representing the additive noise. Alternatively, the model can be described in terms of samples of the random variables to describe each nth measurement, where $g_{k,1} \dots g_{k,N} \sim G_k$, $x_{k,1} \dots x_{k,N} \sim X_k$ and $w_n \sim W$. Thus, each measurement is described by a sample of each random variable,

$$\begin{aligned} y_n &= \sum_k g_{k,n} x_{k,n} + w_n \\ \text{s.t. } g_{k,n} &\geq 0, \sum_k x_{k,n} = 1 \end{aligned} \tag{2.12}$$

Typically, exact values of both $x_{k,n}, g_{k,n} \forall k = 1 \dots K \forall n = 1 \dots N$ are unknown and so is w_n. Given some prior knowledge on the K possible sources, G_k, the parameter estimation problem is to estimate sample values $x_{k,n}, g_{k,n}, w_n$ of respective random variables $X_k, G_k, W, \forall k = 1 \dots K \forall n = 1 \dots N$ that best describe y_n. Note that the formulae for the compositional model varies from the standard Gaussian Mixture model in that proportions X are a random variable compared to π which is a deterministic parameter. The source G is also treated as a separate random variable with a prior probability,

whereas the Gaussian Mixture model the sources are inferred from N measurements, where the assumption is that each measurement can only be a member of a single source. Furthermore, the sources themselves are not labelled unlike the compositional model where use of prior probabilities automatically impose a label.

The sample instance of the compositional model is referred to as a linear mixture model or linear subpixel model in Hyperspectral Imaging. The problem of inferring the sample values is known as spectral unmixing and is a parallel to the blind source separation problem found in speech processing. Although we have modelled he problem using a compositional model, with probability, many techniques deal with the raw measurements directly. We refer to such methods as deterministic methods, where unknown parameters are not random variables but are fixed values. Deterministic methods are typically iterative, due to the need to evaluate sum-to-one and positivity constraints. The problem is solved via a combination of dimensionality reduction, since $D >> K$, and norm-minimisation, where inclusion of the constraints is addressed through optimisation of the Lagrangian Dual. Additionally matrix factorisation methods [25], enable the parameters to be estimated simultaneously assuming the noise is fixed and known across N measurements. A Bayesian treatment of the problem enables parameter uncertainty to be captured through the form of variance, which is unavailable through matrix factorisation methods. Furthermore, it enables the utilisation of prior knowledge of $g_{k,n} \forall k = 1 \dots K$ which is available through the form of spectral libraries. Measurements from spectral libraries differ from the scene due to atmospheric, instrumentation and scene variation stated in the first chapter which further illustrates the advantage of introducing randomness and probability. In Chapter 5 we provide an elaborate treatment of varying approaches to this problem in the hyperspectral community. The purpose here is to introduce the Bayesian treatment of the problem.

2.3.1 Bayesian Parameter Estimation

In the Bayesian treatment of the parameter estimation problem, unknown parameters are treated as random variables and are described by probability distributions. Thus the parameter Θ introduced previously, is a random variable as opposed to a fixed parameter and $\theta_1, \dots \theta_N \sim \Theta \in \mathbb{R}^p$ are its samples. Thus, the nth measurement lies in the joint space described by the joint probability distribution P_{y_n,θ_n}, where $y_n \in \mathbb{R}^D$. The parameters for the sub-pixel model consist of $\Theta = \{G_k, X_k, \forall k = 1 \dots K, W\}$, where parameter estimates correspond to sample values of the parameters. The posterior probability $P(\Theta = \theta_n | y_n)$ given measurement y_n is derived according to the Bayes rule,

$$P_{\Theta=\theta_n|y_n} = \frac{P(y_n|\Theta = \theta_n)P(\Theta = \theta_n)}{p(Y = y_n)}, \tag{2.13}$$

where $p(Y = y_n)$ is the prior probability of the measurement expressed as,

$$P(y_n) = \int_{\theta_n \in \Theta} P(y_n|\theta_n)P(\theta_n)\partial\theta. \tag{2.14}$$

Computation of aposteriori joint distribution and in (2.13) is possible only when the joint probability between $P(G_k, X_k)\forall k = 1 \dots K$ can be computed. Although a prior distribution can be enforced on each of the K candidates in G, we do not know which of the K sources describe the measurement. Therefore, both parameter estimates are unknown along with their respective means and variances given y_n. We rely on iterative Monte-Carlo simulations such as the Gibbs Sampler to estimate the conditional aposteori distribution of each parameter and subsequently the joint aposteori distribution.

2.3.2 Inference using Gibbs Sampling

In chapter 5 of this thesis a Gibbs Sampler is applied to a problem where the generative model is Bayesian. We derive suitable conditional posterior probabilities to sample source signal and weight parameters. Source signal means and variances are estimated by a Gaussian Process, making the source signals Gaussian and the posterior probability given the weights also a Gaussian. The weights and sources are drawn

from a conditional Gamma Gaussian distribution, where the conditional probability is computed iteratively between each *kth* source and *weight*. Gibbs Sampling is an instance of Markov Chain Monte Carlo (MCMC) method that is used to approximate an intractable high-dimensional distribution where a closed form expression for the distribution cannot be obtained.

For this problem, although the Gamma Gaussian provides a closed form expression for the joint distribution, simultaneous estimate of the parameters are not tractable. Moreover, the sum to one and positivity constraints also make sampling from the conditional posterior a tricky exercise since a Gamma distribution or any other conjugate distribution to the Normal does not provide sum-to-one-constraints. The intractable distribution corresponds to the joint distribution of all parameters, where the term high-dimensional corresponds to multiple random variables. In a Gibbs sampler, any random variable, Z_k is sampled one-at-a-time conditional on the knowledge of remaining $K-1$ instances of the same parameter $Z_{k'}$ as well as other remaining random variables, Y/Z, whose values remain fixed at the time of sampling. The assumption here, is that it is tractable to sample from $p(Z_k^m|y_n, Z_{k'}^m, Y/Z)$ which is the conditional posterior distribution. If $m \to \infty$, then $z_k^{(m)}$ is a valid sample from the originally intractable distribution. There are polynomial bounds that describe how long it takes for the parameter estimates to converge to its true values [26], but it is difficult to guarantee. Hence, it is generally advised to run the Gibbs Sampler from several different initialisation conditions. If parameter estimates are slow to change and slow mixing is observed, Rao-Blackwellised Gibbs Sampling or Blocked Gibbs Samplers can alternatively be used [27].

2.4 Band Selection

Unsupervised band selection in this thesis is carried out using two different methods: convex optimisation (Chapters 3,4,5) and stochastic beta processes (Chapter 5). In each chapter the band selection/ranking cost function varies according to optimisation criteria and motivation. Optimality criteria for convex optimisation methods include maximum likelihood and Kull-back Leibler divergence, where band-selection

is carried out offline once the scene has been gathered by all frequency bands. We use stochastic beta processes in an online band selection framework in Chapter 5, where the optimality criteria is based on most likely bands in a probabilistic sense required for measurement of each pixel. In this section we explain how the underlying optimisation/inference is performed for both convex optimisation and stochastic beta processes. However, the exact form of each cost function is introduced separately in each chapter.

2.4.1 Convex Optimisation

Consider the following problem,

$$\begin{aligned} &\min_{t} -f_0(t,\theta) \\ &\text{s.t. } c_1 \leq f_1(t) \leq c_2, f_2(t) = c_3 \end{aligned} \tag{2.15}$$

where, $f_0(t,\theta)$ represents the objective function and $f_1(t), f_2(t)$ are constraint functions. The problem differs from the maximum likelihood problem stated in equation (2.2) in terms of the constraints and parameterisation of variable t. For example, in chapter 3,4, the band selection problem is described as a combinatorial measurement selection problem. The objective $f(0)$ corresponds to the likelihood of statistical model describing the data and is additionally parameterised to infer the measurements which maximise the likelihood given the model parameters are fixed. The in-equality constraints $f_1(t), f_2(t)$ placed on the problem correspond to minimum and maximum number of measurements determined apriori. Similarly in Chapter 3, a similar parametrisation is applied to maximise the Kullback-Leibler divergence between two statistical models.

A convex optimisation problem requires that the objective function is convex, in-equality constraint functions f_1 are also convex and equality constraints f_2 are affine. These objective and constraints can be manipulated to obey these conditions in many cases. Convex optimisation is suitable for problems without an analytic solution that are NP-Hard to solve iteratively, can be framed in terms of the Lagrangian dual with properties stated above and take the primal form stated in (2.15). The selling point for convex

optimisation lies in the fact that a solution to such a problem is equivalent to the solution to the original primal problem stated in equation (2.15). Firstly, a Lagrangian relaxation to the problem stated in equation (2.15) merges constraints into the objective, where each constraint represents a penalty that constricts the number of possible solutions. Thus, the relaxed objective is given by,

$$\ell(\lambda_1, \lambda_2, \lambda_3, t) = f_0(t) + \lambda_1(c_1 - f_1(t)) + \lambda_2(f_1(t) - c_2) + \lambda_3(f_2(t) - c_3) | t \in T \quad (2.16)$$

where, T represents a feasible set that includes all possible solutions and $\lambda_1, \lambda_2, \lambda_3$ are Lagrangian multipliers. The dual problem is given by,

$$\ell^* = \max_{\lambda_1, \lambda_2, \lambda_3} \ell(\lambda_1, \lambda_2, \lambda_3 | t_{\min}), \quad (2.17)$$

where, $t_{\min}$ is first obtained by minimising (2.16) with respect to t. Given that we have a feasible set, T, it is well established that the solution to the dual in (2.17) is equivalent (2.15) under the stated conditions. Please refer to [15] for further details on the proof. The problem is then solved using semi-definite programming or interior-point methods [15] depending on the type of cost function and constraints. These techniques are not described in this chapter since the focus is on formulating the problem and moreover we are not interested in the computational performance but are satisfied as long as the global maxima/minima can be reached and an optimal solution can be found. In chapters 3 and 4 we show why the problem requires a convex optimisation solution and frame the problem such that the necessary conditions for the objective and constraints are fulfilled.

2.4.2 Stochastic Beta Processes

Consider a random variable $Y \in \mathbb{R}^D$ which describes the possible measurement values for all $n = 1 \ldots N$ pixels across D bands. An unknown random variable, $X_n \in \mathbb{R}^{P_n}$ is a subset of Y, $X_n \subset Y$, contains $P_n <= D$ desired number of bands to be inferred from a maximum number D for each nth measurement. The value of P_n is unknown for all N pixels. Non-Parametric Bayesian techniques can be used to identify the exact number P_n as well as the exact co-ordinates that are unique to each nth pixel. This problem is

referred to in a general sense as a feature selection problem in machine learning, where each co-ordinate refers to a single band or feature. It is general because in many cases P_n is assumed to be constant across many features. However, in chapter 5, we aim to carry out online band selection for each nth pixel within a probabilistic framework. Stochastic Beta Processes are a non-parametric stochastic process that enables such inference to take place. We briefly define its mechanics.

Consider a measure space which represents the possible values of band measurements across D frequency bands across N pixels, represented by $\Theta \subseteq \Re$. For the nth pixel, the joint probability of measurement values across the co-ordinates are uniquely parameterised by $P(\theta_n(1), \dots \theta_n(P))$ where $\theta_n(1), \dots \theta_n(P) \in \Theta$ correspond to P_n unknown number of independent parameters in measure space for each nth pixel. The parameters themselves are drawn from a base distribution $G_0(\Theta)$. Therefore,

$$\begin{aligned} y_n(d) &\sim P(\theta_n(d)) \\ \theta_n(d) &\sim G_0(\Theta), \forall n = 1 \dots N, \forall d = 1 \dots P_n \end{aligned} \tag{2.18}$$

where $y_n(d)$ denotes the measurement value at the dth band of the nth pixel, $\theta_n(1) \dots \theta_n(P) \sim G_0(\Theta)$ corresponds to P_n draws from a base distribution G_0. Since the parameters are independent they form unique partitions. Non-Parametric Bayesian methods provide a framework for determining P_n possible candidates $\theta_n(1) \dots \theta_n(P)$ for each nth observation without specifying P_n. The stochastic process describing likelihood or utility of each co-ordinate across $n = 1 \dots N$ pixels is given by,

$$G_n = \sum_{d=1}^{P=D} \pi_n(d) \delta_{\theta_n}(d)(\Theta), \forall n = 1 \dots N \tag{2.19}$$

where, $\delta_{\theta_n}(d)$ is a dirac delta indexing the dth band location or co-ordinate for the nth pixel and $\pi_n(d)$ is the likelihood of the dth band parameter $\theta_n(d)$ being used to describe the nth pixel. The values of θ and π are drawn successively from unique distributions, $\theta_n(1) \dots \theta_n(P) \sim G_0, \pi_n(1) \dots \pi_n(P) \sim B$ making each θ value represent a unique partition along with the corresponding π. Note that the upper limit is set to $P = D$ and D may equal an infinitely large number and is often denoted as ∞. After many observations of θ, the posterior update shown below ensures that the number

of possible parameters reduce to a manageable number. In this thesis, B corresponds to a Beta distribution, which does not enforce sum to one constraints on π unlike the Dirichelet distribution. For $n = 1 \dots N$ pixels the sequence of random variables form the stochastic non-parametric process where the number of parameters grows with each pixel measured. The posterior probability of G given $N \times P_n$ draws of $\theta_n(d)$ is given by,

$$\mathrm{E}\{G(\Theta)|\theta_n(1)\dots\theta_n(P)\} = \frac{1}{c+N}(cG_0(\Theta) + \sum_{d=1}^{P} N(d)\delta_{\theta_n}(d)(\Theta))$$
$$\text{where, } N_d = \sum_{i=1}^{P} \delta(\tilde{\theta}_n(i), \theta_n(d)) \tag{2.20}$$

where N_d is the number of previous draws denoted $\tilde{\theta}_n(i)$ that equal $\theta_n(d)$, c is a Beta process hyperparameter that weights the influence of the prior distribution in relation to the previously chosen parameters. This leads to a predictive distribution of similar form where the $P(\tilde{\theta}_{N+1} = \theta_n(d)) \propto N_d$. Thus, the probability of rarely used features or its corresponding parameter estimates diminish as other estimates and features are chosen. Please refer to [27] and [28] for an introduction to Bayesian Non-Parametrics, Thibaux *et. al.* in [29] apply the problem to feature learning using Beta Processes. In chapter 4 of this thesis P_n corresponds to P_n bands out of D that are sufficient to represent the nth pixel, the non-parametric stochastic process is a Beta process and $\theta_n(1)\dots\theta_n(P)$ corresponds to the observation parameters across each band for the nth pixel, drawn from a discrete band utility prior, B_0, estimated using training data.

Chapter 3

Unsupervised Band Selection using Gaussian Mixtures and Maximum Likelihood Criteria

IN this chapter, we deal with the problem of unsupervised identification of a subset of frequency bands to improve estimation accuracy of the scene model. This is a measurement selection problem, where the aim is to eliminate noisy measurements from the set of signals describing a scene. Experiments conducted on synthetic Gaussian Mixture data demonstrate the effectiveness of the proposed method to choose an optimal subset of sensor frequency bands that improve the model estimation accuracy. The chapter also highlights the limitations of the proposed approach in the context of the thesis.

3.1 Introduction

It is shown in [3] that noise sources such as payload motion, measurement noise and atmospheric interference are uniform for all pixels in the scene. Some of these noise sources result in additive or multiplicative noise and are typically localised to certain frequency bands [3]. Not all noisy bands are easily picked up by atmospheric correction algorithms [30] which do not address estimation accuracy of measurement parameters in each frequency band. In noisy bands, spurious measurements are known to worsen the estimation accuracy of the scene model. The omission of such noisy frequency bands is an important problem since it can lead to improvements in the model estimation accuracy which subsequently leads to better analysis either for classification problems [31] or target/anomaly detection [32]. Unsupervised methods facilitate for on-board processing and faster throughput in terms of obtaining time-critical results. There are many prior works that justify the need for frequency band selection from the standpoint of data redundancy as well as model over-fitting. Nonetheless, in a pure regression sense, estimation error of scene parameters are reduced in a penalised sense when noisy measurements in the signal are removed. Typically, many existing methods [31], [32], [33], [34] assume prior knowledge and carry this out in a supervised sense but we assume no prior knowledge on the existing classes in the scene and aim to carry out unsupervised frequency band selection. In this study, we use the EM algorithm, which is an iterative procedure, to estimate band parameters that correspond to a Gaussian Mixture model. We use an objective function for scoring bands such that theoretical maximum likelihood of the model is guaranteed in a penalized sense for M number of bands. It is shown via simulation that the model estimation error converges in a reasonable number of iterations. In section 3.4.1, we motivate the use of maximum likelihood as the optimality criteria. In section 3.2.1 we provide justification for why we consider Gaussian mixture models. We briefly consider two different scoring approaches in section 3.2.2 and argue why the non-linear approach is more suited to the problem. In sections 3.4.4 and 3.4.2 we develop the methodology behind our proposed approach which is tested with some simulations on synthetic Gaussian Mixture data in section 3.5. Importantly, in section 3.6 we describe the limitations of such an approach

in-terms of band selection criteria which form the catalyst for the work conducted in chapter 4.

3.2 Background

3.2.1 Gaussian Mixtures for Hyperspectral Data

In this chapter and in chapter 4 frequency band measurements gathered by an overhead platform of a scene contain more than one material class, which is typically the case. A Gaussian mixture probability distribution has been used in previous studies, [11], [10] to model hyperspectral data with D bands. The non-Gaussian nature of hyperspectral data is evident in numerous studies which use elliptic-T models [12], [13]. The probability measure is a many-to-one mapping representing the likely spectral response values of *any* spectral band. A one-to-one mapping would be a probability assigned to the possible values across each spectral band. Hence, the many-to-one probability can be considered a joint probability. Sawo *et. al.* in [35] show that if the joint pdf is a Gaussian mixture, the marginal is also a Gaussian mixture but with a correlation factor influencing the number of Gaussian components. In terms of frequency bands, the marginal probability of a D dimensional Gaussian mixture could be for each dth band, where $d = 1 \dots D$. We can also extend this definition to say that the marginal pdf across a localised set of neighbouring bands is a multivariate Gaussian mixture since the joint probability of measurements across each D bands is also a Gaussian mixture.

3.2.2 Nonlinear vs Linear Band Scoring

Consider the following definition:

Definition 5: A localised *frequency band-subset* is a set of neighbouring bands that are correlated due to the overlap in spectral response. A localised band-subset can contain anywhere between 3-20 bands [3] according to the band configuration in the sensor

payload.

The scoring techniques proposed in section 3.4.4 depend on the practitioner's intent. There are two choices: (1) If the practitioner is happy in finding out P optimal localised frequency band-subsets out of a possible R total band-subsets, then linear likelihood scoring should suffice. In the linear case, the global likelihood score is additive of the likelihoods of individual band-subsets. Thus $P - R$ band-subsets would contain a lower likelihood score than the individual scores of the R band-subsets. (2) However, if exact knowledge of M out of D exact bands is needed, under a locally correlated model a non-linear scoring technique is required to find the optimal bands. This is because a band's contribution to the the global likelihood score is dependant on the neighbouring bands. The context-specific nature of bands introduces non-linearity into the problem if we wish to find the M exact frequency bands. From an application perspective, the exact knowledge of M optimal bands is especially useful due to hyperspectral bands having a narrow spectral bandwidth. For example atmospheric interference maybe negligible in some bands in a localised band-subset but prominent in others within the same subset [3]. The fine spectral resolution provided by hyperspectral sensors provides a rich measurement set, which introduces non-linearities that make a non-linear scoring approach more useful.

3.3 Existing Work

Du *et. al.* in [36] adopt a similarity measure to find the subset of frequency bands that are most dissimilar to each other in terms of projecting the most orthogonal components after the application of an orthogonal subspace projection or least similar in terms of being linear combinations of each other. In this method there is a reliance on the initial chosen band subset where another pre-processing step is required to select the initial bands. In our proposed method we use a single criteria throughout the band-selection process and do not rely on any other heuristic. The machine learning community refers to bands as features and a cluster of studies deal with problem of

feature selection and unsupervised learning. Graham and Miller [37], use the EM algorithm to estimate the unknown Gaussian components of a D-dimensional mixture. The saliency of each of the D frequency bands/features in describing the component is determined by the change in the component distribution when considering remaining features. Whenever two component distributions share similar parameters across a particular feature, that particular feature is deemed irrelevant. Law *et. al.* [38], apply a similar saliency principle by applying a soft feature saliency for all components but measure the feature's usefulness by considering the relative increase in number of Gaussian components in the model as a result of its presence. Our work differs, in that we steer away from combinatorial approaches to evaluate the usefulness of each feature. We deem a feature as useful in the context of its neighbours and how well it describes the global model whilst assuming a fixed number of model components. Therefore, we address the estimation accuracy of the D dimensional model as opposed to using sparsity to improve identifiability of Gaussian components such as the latter study.

Roth and Lange [39] carry out feature selection in a regression setting, where the problem is formulated as an extension of Linear Discrimanant Analysis (LDA) problem [40]. A scoring procedure [41], referred to as LASSO [42], is used to quantify each feature's contribution in terms of improving the class membership of observations to the global Gaussian mixture. The philosophy of Roth and Lange is similar to the one adopted in this chapter, however, eigen-decompositions of the original feature set are used to overcome large combinatorial problems. In this study, we use a convex relaxation of the original problem to overcome this issue.

In this sense, our work is closest in flavour to Joshi *et. al.* in [43], who address the problem of selecting $M \subset D$ correlated measurements which improves the estimation accuracy of a Gaussian model fitting the measurements. The estimation accuracy for the linear Gaussian is characterised by an error covariance and performance is measured by calculating the volume of the confidence ellipsoid and subsequently relaxing

the scoring problem to choose M measurements. If D measurements were independent, estimation accuracy is simply an additive function of the number of measurements which requires a search of $DlogD$. To address measurement correlation and to avoid a computationally intensive search procedure (for large D), we also use a convex relaxation procedure but deal with non-Gaussian measurements unlike Joshi *et. al.* in [43]. We ensure the M frequency bands selected are optimal in terms of how well they describe the measurements.

3.4 Maximum Likelihood Criteria for Band Scoring

3.4.1 Motivation

The non-Gaussian nature of hyperspectral data can be attributed to multiple material classes in the scene as well as numerous noise sources listed in chapter 1. Noise and atmospheric interference introduced during the measurement process is sporadic in nature, specific only to certain frequency bands and hence difficult to model [3]. In anomaly detection applications, the scene model is a representative background model since true anomalies tend to be sparse in comparison relative to the land-cover and natural vegetation. Thus, elimination of noisy frequency bands can thus be a catalyst for improving the likelihood of the background model due to removal of sporadic measurements. This translates to an improvement in anomaly detection performance. In this chapter, we attempt to eliminate the noisy frequency bands and maximise likelihood of the model.

Maximum likelihood criteria for frequency band selection also has some nice theoretical properties. If a set of M frequency bands are selected out of a possible D bands such that the likelihood of the model is a maximum. We can say that the fewest number of additional assumptions are made about the data generation process. Berger *et. al.* [44] show that the maximum likelihood estimate of a model from the exponential family is the dual equivalent of a model where the entropy is maximum with respect to its natural parameter. The maximum entropy principle states that such a model makes

the fewest number of assumptions of the data generation process. Selecting M optimal bands out of D could provide the most certainty in terms of model estimation accuracy which could eliminate additional sources such as atmospheric interference that are difficult to characterise. For the band selection application this could mean in a theoretical sense that the chosen bands are the ones that contain the least amount of band noise, atmospheric interference and the least number of redundant information sources. Thus, the measurements are more likely to be explained by their true chemical properties when maximum likelihood criteria is used due to fewer assumptions about the measurements' sources.

3.4.2 Proposed Model

Definition 6: We consider a spatial subset of a snapshot captured by the hyperspectral sensor, where we consider a total of N *spatial samples* captured by each of the $d = 1 \dots D$ spectral frequency bands. We assume the samples are independent and identically distributed (i.i.d) in a spatial sense, which means objects in neighbouring locations do not affect the radiation gathered at the location of interest. This is not always true in reality where pixel reflectance is influenced by radiation reflected by neighbouring pixels around it, referred to as the adjacency effect. Nonetheless, we argue that any evidence of spatial correlation is due to the homogeneity of natural scenes and signal measured is pre-dominantly an accurate representation of the material(s) residing within the pixel. Healey *et. al.* in [45] claim that the adjacency effect is small enough to be neglected in most cases other than hazy atmospheric conditions or if small objects lie within the pixel. We assume adjacency effects are minimal in the context of this study and note that any small man-made objects occupying the pixels are sparse in number in relation to the scene observed. In a broad sense, this validates the use of probabilistic models that assume the i.i.d condition.

Definition 7: Each *co-ordinate* of the $Q(r)$ dimensional vector represents a frequency band. A $Q(r) \subset D$ of co-ordinates is referred to in plural form as the *rth* subset of *co-ordinates*. The co-ordinates of the $Q(r)$ dimensional vector are correlated due to

the sensor's frequency response, where neighbourhood correlation is due to the proximity of the sensor bands as mentioned earlier. Since this is localised to neighbouring bands and is known prior to measurement [11], [3], this makes the measurements in *rth* band-subset independent to the remaining $R-1$ band-subsets given $Q(r) \forall r = 1 \dots R$. Thus, each *nth pixel* is a multi-dimensional vector whose co-ordinates are denoted by $y(1,r,n), \dots y(Q(R),R,n)$.

Definition 8: Each *spectral sample*, denotes the measurements of the *nth* pixel, $y(1,r,n)$, $\dots y(Q(R),R,n)$, across the *rth* frequency band subset. Each spectral sample is i.i.d (since the R band-subsets are independent).

Thus, the entire set of pixels in the scene consisting of spatial samples across all D frequency bands and spectral samples across R frequency band subsets is summarised by a matrix $\bar{Y}$.

$$\bar{Y} = \begin{bmatrix} y(1,1,1) & y(2,1,1) & \cdots & y(Q(R),R,1) \\ y(1,1,2) & y(2,1,2) & \cdots & y(Q(R),R,2) \\ \vdots & \vdots & \vdots & \vdots \\ y(1,1,N) & y(2,1,N) & \cdots & y(Q(R),R,N) \end{bmatrix}$$

where, $\bar{Y}$ is an $N \times D$ measurement matrix, where each of the N rows represents N pixels and each column represents the $Q(r)$ th frequency band from the *rth* frequency band subset.

If a random variable $Y(r) \in \mathbb{R}^{Q(r)}$ represents all $Q(r)$ dimensional measurements in the *rth* frequency band-subset. We let random variable $Y \in \mathbb{R}^D$ where $Y = \prod_{r=1}^{R} Y(r)$ be the possible measurement values across D bands for the N pixels. In this chapter we represent $Y(r)$ as a Gaussian Mixture model given the spectral measurements in the *rth* band-subset.

$$P_{Y(r)} = \sum_{j=1}^{K} \mathcal{N}(;\mu_j(r),\Sigma_j(r))\pi_j(r), \tag{3.1}$$
$$\forall r = 1 \dots R \text{ frequency frequency band-subsets}$$

where $\mu_j(r) \in \mathbb{R}^{Q(r)}$ is the *jth* Gaussian component mean of the *rth* frequency band-subset, $\Sigma_j(r) \in \Re^{Q(r)\times Q(r)}$ is the covariance of the Gaussian and component proportions are denoted by $\pi_j(r)$. The probability across R independent frequency band-subsets,

$$P_Y(;\theta) = \prod_{r=1}^{R} P_{Y(r)}(;\theta(r)) \tag{3.2}$$

where, $\theta \in \Theta \subset \mathbb{R}^q$. In this chapter, we assume prior knowledge of the number of Gaussian components denoted by K since estimating model order is not the primary focus of the study. The problem now is two-fold. The first is to estimate the maximum likelihood of θ in (3.2) and the second is to select M frequency bands which are optimal in terms of maximising the likelihood.

3.4.3 EM Algorithm

Parameter estimation in Gaussian Mixture models cannot be performed analytically but only iteratively, if membership of data to Gaussian components in the model is unknown. We refer to such datasets as un-categorised. The model parameters often fail to converge to an asymptotic value [24] if parameter complexity is high as is the case with high-dimensional hyperspectral data. This phenomena is addressed by Fessler and Hero in [24], who reduce the data into smaller partitions when the measurement space is i.i.d. The chosen subset is smaller, less-informative and is modified in-between EM iterations. The authors demonstrate that analytic convergence of likelihood is achieved in spite of a modification to measurement the space as long as the chosen partitioned subset adheres to the SAGE condition. The SAGE condition states that if measurements of $\bar{Y}$ and a partitioned subset of measurements described by $Z^S \subset \bar{Y}$ are independent, the probability of the entire measurement space denoted by random variable Y is given by the product of conditional of Y given Z^S and the probability of Z,

$$P_Y = P_{Y|Z^S}(;\theta^{\tilde{S}})P_Z(;\theta^S) \tag{3.3}$$

where Z denotes the random variable describing Z_S measurements. In this chapter, $Z \equiv Y(r)$ for any *rth* and refers to N measurements contained in the *rth* frequency

band subset, $\theta^{\tilde{S}}$ represents the Gaussian Mixture parameters which belonging to the remaining $R-1$ band-subsets. Its worth re-emphasising that the condition only holds true as long as measurements in Z^S and those described by Y are independent. The likelihood of the θ given the measurement space Y is given by the product of likelihoods,

$$L(y_1 \dots y_N; \phi) = L(Z_S; \theta^S) \times L(y_1 \dots y_N | Z_S; \theta^{\tilde{S}}), \tag{3.4}$$

where, the parameter ϕ characterises all measurement values in Y. We maximise the likelihood in equation (3.4), w.r.t to θ^S which means the latter term after the product can be ignored. We have,

$$\arg\max_{\theta^S \in \Theta} \log(L(Z_S; \theta^S) L(y_1 \dots y_N | Z_S; \theta^{\tilde{S}})) \equiv \arg\max_{\theta^S \in \Theta} \log(L(Z_S; \theta^S)), \tag{3.5}$$

which, is the maximisation performed in the standard EM algorithm defined in section 3.4.3, where we alternate between the $R-1$ frequency band-subsets for Z_S till maximum likelihood convergence is reached. Let $Z_S = Y(r)$ and
$\theta_S = \{\mu_1(r), \dots \mu_K(r), \Sigma_1(r) \dots \Sigma_K(r), \pi_1(r) \dots \pi_K(r)\}$ for some value of r,
the un-categorised log-likelihood is given by,

$$L(y_1(r) \dots y_N(r); \theta^S) = \prod_{n=1}^{N} \sum_{j=1}^{K} \pi_j(r) \mathcal{N}(y_n(r); \mu_j(r), \Sigma_j(r)) \tag{3.6}$$

where $v_{j,n}(r) \in \{0, 1\}$ and represents the membership of the nth spectral measurement to the jth component in the rth frequency band-subset. The *complete* data log-likelihood (neglecting constant terms not dependent of the parameters and assuming that $\pi_j(r) > 0$) is thus,

$$\log(L_c(y_1(r) \dots y_N(r); \theta^S) = \ell_c(y_1(r) \dots y_N(r); \theta^S) = \sum_{n=1}^{N} \sum_{j=1}^{K} v_{j,n}(r) \dots \left(\log \pi_j(r) - 1/2 \left((y_{r,n} - \mu_j(r))^T \Sigma_j^{-1}(r)(y_n(r) - \mu_j(r))\right)\right) \tag{3.7}$$

Since the dataset is un-categorised, the membership of V is unknown and we consider the expected value of V. The EM algorithm thus involves computing the conditional expectation of $\ell_c(y_1(r) \dots y_N(r); \theta^S)$. This requires computation of $\mathrm{E}\left\{V_{j,n}(r) | y_1(r), \dots, y_N(r)\right\}$. It is straightforward to show [20] that

$$p_{r,j,n} = \mathrm{E}\left\{V_{j,n}(r) | y_1(r), \dots, y_N(r)\right\} = \frac{\pi_j(r) \ \prod_{n=1}^{N} \mathcal{N}(y_n(r); \mu_j(r), \Sigma_j(r))}{\sum_{j=1}^{K} \pi_j(r) \ \prod_{n=1}^{N} \mathcal{N}(y_j(r); \mu_j(r), \Sigma_j(r))}. \tag{3.8}$$

We denote this quantity by $p_{j,n}(r)$. The M step computes maximising argument for each *mth* estimate of the expected complete log-likelihood given by,

$$\mathrm{E}\left\{\ell_c(y_1(r),\ldots,y_N(r);\theta^S)\right\} = \sum_{j=1}^{K}\sum_{n=1}^{N} p_{j,n}(r)\log\pi_j(r) - \frac{p_{j,n}(r)}{2}\left((y_n(r)-\mu_j(r))^T\Sigma_j^{-1}(r)(y_n(r)-\mu_j(r))\right). \tag{3.9}$$

The m+1th parameter update is thus,

$$\hat{\pi}_j^{(m+1)}(r) = \frac{1}{N}\sum_{n=1}^{N} p_{j,n}^{(m)}(r),\ \hat{\mu}_j^{(m+1)}(r) = \frac{\sum_{n=1}^{N} p_{j,n}^{(m)}(r)y_n(r)}{\sum_{n=1}^{N} p_{j,n}^{(m)}(r)}$$
$$\hat{\Sigma}_j^{(m+1)}(r) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K} p_{j,n}^{(m)}(r)AA^T \tag{3.10}$$

where, $A = (y_n(r) - \mu_j^{(m+1)}(r))$. Once we derive a sufficiently accurate set of parameter estimates across the D frequency bands, $D-M$ sub-optimal bands are removed and we can continue running the EM steps until the total likelihood for the M models are reached. The SAGE condition shows that the parameter estimates provide the maximum likelihood of the model.

When the E-M steps are repeated L times across each *rth* frequency band-subset $\forall R$ band-subsets. We have parameter estimates $\hat{\mu}_j(r), \hat{\Sigma}_j r), \hat{\pi}_j(r) \forall r = 1\ldots R, \forall j = 1\ldots K$. If K is assumed to be the same across R band-subsets, the expected complete log-likelihood of the R independent independent band-subsets, is given by,

$$\mathrm{E}\left\{\ell_c(y_1,\ldots,y_N;\hat{\phi})\right\} = \sum_{r=1}^{R}\sum_{j=1}^{K}\sum_{n=1}^{N} p_{j,n}(r)\log\hat{\pi}_j(r) - \frac{p_{j,n}(r)}{2}\left((y_n(r)-\hat{\mu}_j(r))^T\hat{\Sigma}_j^{-1}(r)(y_n(r)-\hat{\mu}_j(r))\right). \tag{3.11}$$

where, $\hat{\phi} = \{\hat{\theta}_1^S \ldots \hat{\theta}_R^S\}$ denotes mixture parameter estimates across all R band subsets, which are estimated by carrying out the E and M steps indicated above.

3.4.4 Non-linear Band Scoring using Convex Optimisation

Consider random variable X which denotes a linear transformation of Y and is given by,

$$X = T.Y \tag{3.12}$$

where, $X \in \Re^M$ and represents the possible measurement values made across M optimal frequency bands, T is $M \times D$ binary matrix whose elements take the values $\{0, 1\}$, $T_{m,d}$ represents whether the dth frequency band is represented by the mth element of X. In this section we consider the non-linear band scoring problem. We wish to determine a subset, M of the D available bands of given size $M < D$ which best represents the scene under analysis. By this we mean selecting those M bands which maximise the likelihood compared to any other subset of M bands. Taking into account the neighbourhood correlation, this is a combinatorial optimisation problem which is NP-Hard [43]. So rather than performing a "hard" assignment of bands, we follow the approach of [43], and perform a "soft" assignment. The soft assignment problem can be viewed as an extension to the treatment of the "hard" assignment problem described as follows. A combinatorial problem of selecting from $\binom{n}{k}$ combinations is simply avoided by applying a relaxation to the band selection parameter and rounding it suitably to represent the binary matrix T.

Parameter estimates for the linearly transformed random variable Y across all R frequency band-subsets are obtained by concatenating the parameters into a single vector or matrix all R frequency band-subsets. This includes $\bar{\mu}_j \in \Re^D$ which is the concatenated component mean, $\bar{\Sigma}_j \in \Re^{D \times D}$ which is a block diagonal-covariance and $\bar{\pi}_j \in \Re^D$ is concatenated component membership to the jth component. It is assumed that each rth band-subset has the same number of components, K which is reasonable given that the all bands are observing the same scene. We consider the problem of finding M optimal bands and subsequently estimating X. The optimisation objective is given by,

$$\begin{aligned} \hat{T} = \arg\max_{\tilde{T}} \ell_x(\hat{\phi}, \tilde{T}) \quad &\equiv \quad \arg\max_{\tilde{T}} \sum_{n=1}^{N} \sum_{j=1}^{K} log\mathcal{N}(y_n; \tilde{T}\bar{\mu}_j, \tilde{T}\bar{\Sigma}_j\tilde{T}^T) \\ &\text{s.t } \tilde{T} \in [0, 1],\ 1^T\tilde{T} = M \end{aligned} \tag{3.13}$$

where, $\ell_x(\phi, \tilde{T})$ is the likelihood of the D frequency band model, $\tilde{T} \in [0,1]^{M \times D}$ is the parameter with respect to which the optimisation is performed. The first constraint is a convex inequality and the second equality constraint ensures that rows and columns of $\tilde{T}$ sum to M. The log-likelihood of D frequency band model can be written as,

$$\begin{aligned} \ell_x(\hat{\phi}, \tilde{T}) &= -1/2 \sum_{j=1}^{K} (N(\log |2\bar{\pi}_j \tilde{T} \bar{\Sigma}_j \tilde{T}^T| - \\ & \quad 1/2 \sum_{n=1}^{N} \sum_{j=1}^{K} (y_n - \tilde{T}\bar{\mu}_j)^T \tilde{T} \bar{\Sigma}_j^{-1} \tilde{T}^T (y_n - T\bar{\mu}_j) \end{aligned} \tag{3.14}$$

This objective can be reduced to the following in order to simplify the optimisation,

$$\begin{aligned} \ell_x(\hat{\phi}, \tilde{T}) &= -N/2 \sum_{j=1}^{K} (\log |2\bar{\pi} \tilde{T} \bar{\Sigma}_j T^T| + tr(\tilde{T} \bar{\Sigma}_j^{-1} \tilde{T}^T S) \\ &+ (\tilde{y}_n - \tilde{T}\bar{\mu}_j)^T \tilde{T} (\bar{\Sigma}_j^{-1}) \tilde{T}^T (\tilde{y}_n - \tilde{T}\bar{\mu}_j) \end{aligned} \tag{3.15}$$

where S denotes the sample covariance of D-dimensional random variable Y, $\tilde{y} \in \Re^D$ is the sample mean across all N pixels and $tr()$ denotes the trace operator. The objective can be shown to be log-concave w.r.t to $\tilde{T}$. Please refer to section 3.4.6 for the complete proof. Since the objective is long-concave and we have convex in-equality and linear equality constraints the problem can be solved as a convex optimisation problem. The use of the standard EM algorithm, where parameter estimates converge to their respective maximum likelihood estimates as well as the equality constraint ensures that there is sufficient sparsity. This was found to be the case in the simulations conducted even though there is no specific condition stating that values of $\hat{T} \approx \{0,1\}$. The new M dimensional subset X is estimated using equation (3.12) where $X = \hat{T}.Y$.

3.4.5 EM CVX Algorithm

1. Repeat E and M steps a maximum of L times across each rth frequency band subset $\forall r = 1 \ldots R$. In the experiments conducted, convergence across each frequency band subset was set at to a maximum value of $L = 20$ iterations. If the likelihood converged at an earlier stage the subsequent parameter estimates were accepted to be optimal in a maximum likelihood sense.

2. Solve for $\hat{T}$ in equation (3.13) using the objective in (3.15)

3. Normalise $\hat{T}$ to values between 0 and 1 and compute the M dimensional linear transformation $X = \hat{T}.Y$.

3.4.6 Proof of Concavity for the Band Selection Objective

Consider the log-likelihood function given in equation (3.14). We will show that this expression is a sum of two concave expressions and hence in turn also concave, which means that we can find a global maxima and solve equation (3.13) as a convex optimisation problem. Properties of convex functions were mainly obtained from Boyd *et. al.* in [15].

Consider the first term in equation (3.14),

$$N/2 \sum_{j=1}^{K} \log |2\bar{\pi} T \bar{\Sigma}_j T^T| \tag{3.16}$$

where $|.|$ denotes determinant. Let function $f(Z) = \log |Z|$, where, $Z = \tilde{T}\Sigma_j \tilde{T}^T$, is a symmetric positive definite matrix. Consider the function, $g(t) = \log |Z + tV|$, restricts $g(t)$ to a positive number, where $Z, V \in |\Re^{D \times D}$. We can say that, $f(Z)$ is a concave function, if $g(t)$ is a concave function as shown.

$$\begin{aligned} g(t) &= \log |Z + tV| \\ &= \log |Z| + \log |I + tZ^{-1/2} V Z^{-1/2}| \\ &= \log |Z| + \sum_{d=1}^{D} \log(1 + t\lambda_d) \end{aligned} \tag{3.17}$$

where, $\lambda_1 \ldots \lambda_D$ are eigenvalues of $Z^{-1/2} V Z^{-1/2}$. When we take the derivative of $g(t)$ with respect to t, we have,

$$\begin{aligned} g'(t) &= \sum_{d=1}^{D} \frac{\lambda_d}{1 + t\lambda_d} \text{ and,} \\ g''(t) &= -\sum_{d=1}^{D} \frac{\lambda_d^2}{(1 + t\lambda_d)^2} \end{aligned} \tag{3.18}$$

where $g'(t), g''(t)$ denote the derivative and second derivative respectfully. Since, $g''(t) < 0$ f is concave. From the term in equation (3.16), we have an affine combination of a concave expression, which makes entire term concave if we ignore the

negative multiplier before $N/2$. Consider the second term in equation (3.14). Let,

$$\ell_B(\hat{\phi}, T) = 1/2 \sum_{n=1}^{N} \sum_{j=1}^{K} (y_n - T\bar{\mu}_j)^T T \bar{\Sigma}_j^{-1} T^T (y_n - T\bar{\mu}_j) \tag{3.19}$$

We shall prove that this quadratic expression is concave by proving that $\frac{\partial^2 \ell_B}{\partial \tilde{T}^2} < 0$. First, consider the derivative of this expression:

$$\begin{aligned} \frac{\partial \ell_B}{\partial \tilde{T}} &= (y_n - \tilde{T}\bar{\mu}_j)^2 (\tilde{T}^2 \bar{\Sigma}_j^{-1}) \\ &= (y_n - \tilde{T}\bar{\mu}_j)^2 2\tilde{T}\bar{\Sigma}_j^{-1} - 2\tilde{T}^2 \bar{\Sigma}_j^{-1} (y_n - \tilde{T}\bar{\mu}_j) \\ &= 2\tilde{T}\bar{\Sigma}^{-1} (y_n - \tilde{T}\tilde{\mu}_j)(y_n - \tilde{T}\tilde{\mu}_j - \tilde{T}) \\ &= 0 \end{aligned} \tag{3.20}$$

Therefore the solution for $\tilde{T}$ include, $0, y_n - \tilde{T}\tilde{\mu}_j, \frac{y_n}{\mu_j}$. This leaves the only possible solution for which the $\frac{\partial^2 \ell_B}{\partial \tilde{T}^2}$ exists.

$$\frac{\partial \ell_B}{\partial \tilde{T}} (y_n - \tilde{T}\bar{\mu}_j) = -\bar{\mu}_j < 0 \tag{3.21}$$

Therefore, the expression is concave. Since both expressions have a negative multiple, concavity is maintained by dividing through by -1 and maximising the negative of the objective in (3.14).

3.5 Experiment A

The dataset in each experiment consists of $N = 1000$ samples which are generated by $K = 2$ component Gaussian mixtures where the mean is randomly selected across all R frequency band-subsets for each trial in each experiment as listed in Table 3.5. The number of bands in each band-subset is set to $Q_r = 3$ which is fixed for all experiments but the total number of band-subsets R vary for each experiment. Covariance is uniform across K components and R across K components and R frequency band-subsets but varied for each trial, where correlation between $Q(r)$ bands in a band-subset is varied arbitrarily in a constrained sense for each rth band-subset. White-Gaussian noise is added in some experiments at a specified SNR, where the SNR calculation is computed in decibels (dB) using $20 \log \frac{Y}{R}$. The estimation error for dth band is given by

Exp. No.	SNR	No. of Bands (D)	Desired Bands (M)	No. of Trials
A1	0	10	5	7000
A2	0	100	20	1000
A3	20	200	25	1000
A4	5,10,20	100	10	1000

Table 3.1. Band Selection Experiment Summary

$E_d = \{\|\theta - \hat{\theta}\| + \|\pi - \hat{\pi}\|\}$ and for M sensors $\left\|\sum_d^P E_d\right\|$. Initialisation conditions were kept constant throughout the experiment. Further details of the experiments are provided in table 3.5.

The primary goals whilst conducting these experiments is to demonstrate that a) the algorithm is robust under different noise conditions and b) the frequency bands selected are the ones with the lowest-SNR. Hence, all testing is conducted with synthetic Gaussian mixture data where noise conditions are adequately controlled. The former aim is reflected in the convergence of parameter estimation error towards an arbitrarily small value as indicated in figures 3.1,3.2, 3.3, 3.5. The latter is tested by adding white Gaussian noise at varying noise ratios and examining which frequency bands are selected. The bands chosen are the ones with a SNR of 20 as opposed to 10 and 5 at two different initialisation points. Figure 3.4, shows the slow convergence of the standard EM, for when no frequency bands are removed. This highlights the benefit of the algorithm performing the band selection within-the-EM loop. In both the noise-less and noisy case, where $\tilde{\phi} \approx \phi_{\tilde{ML}}$, the estimation error converges towards an arbitrarily small value.

3.6 Conclusions and Limitations

The proposed approach finds frequency bands which provide the best model estimate, from which we can infer that the measurements are *likely* to be less corrupted by spurious noise sources and atmospheric effects and hence provide a more useful model for

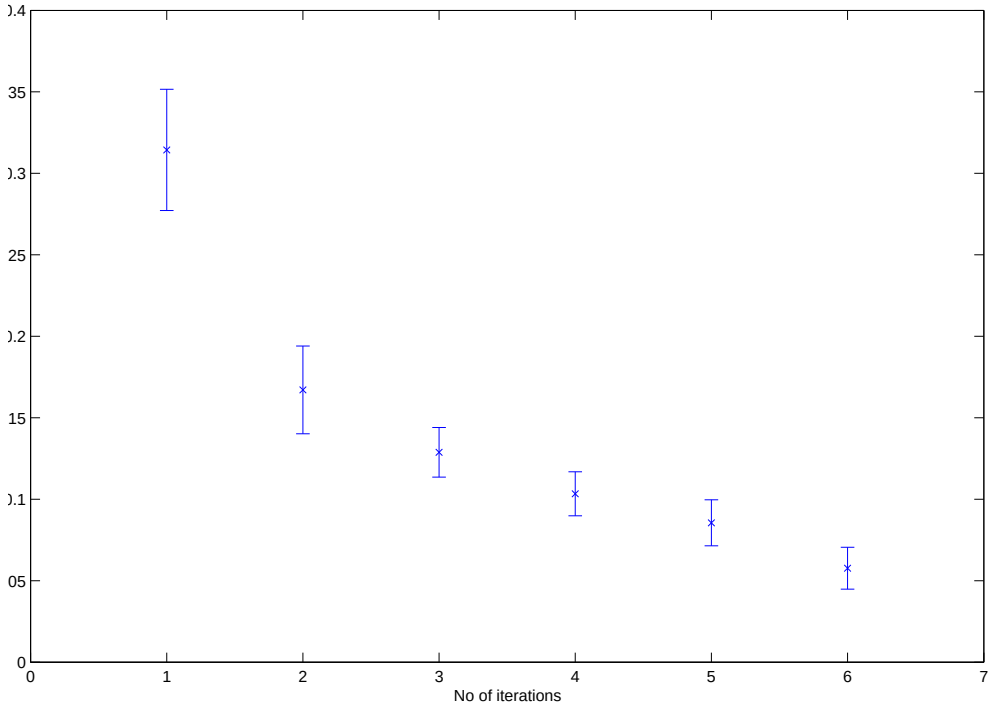


Figure 3.1. Exp. A1: Combined Band MSE; EM-CVX Algorithm; Selected 5 out of 10 frequency bands, 0 dB SNR across all frequency bands

anomaly or target detection. Our results indicated in fig. 3.3, 3.4 demonstrate this to be the case. Furthermore, EM also fails when no bands are removed which highlights the utility of removing bands when performing clustering on a HSI scene whilst retaining the true measurements. The proposed formulation ensures that the bands selected are optimal in-terms of model estimation accuracy relative to those bands that were left out. However, we cannot necessarily guarantee that the collection of bands identified are *optimal* for finding anomalies in the given scene, since its not part of the performance criteria. In the subsequent chapter, the band selection criteria is one which is useful for determining the presence of anomalies in the scene if there are any. Moreover, we also have to specify M, the desired number of bands, which is overcome in chapter 4.

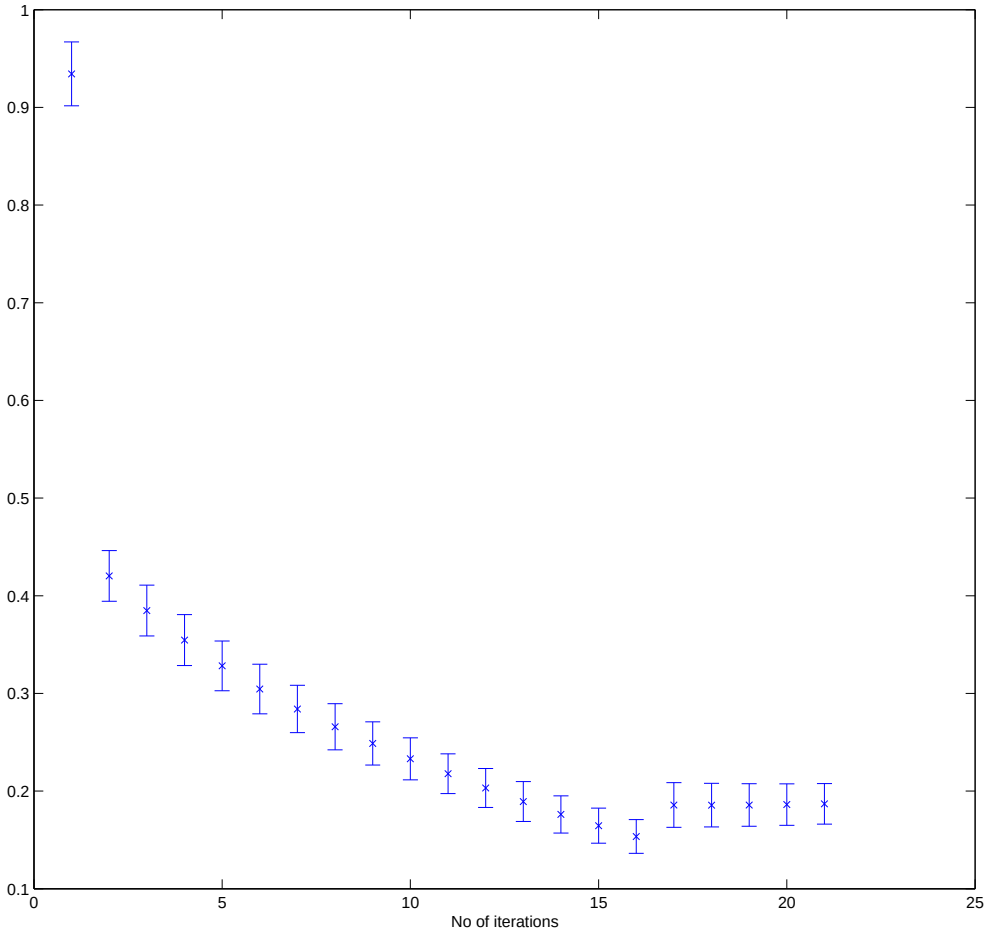


Figure 3.2. Exp. A2: Combined Band MSE; EM-CVX; Select 20 out of 100 frequency bands, 0 dB SNR across all channels

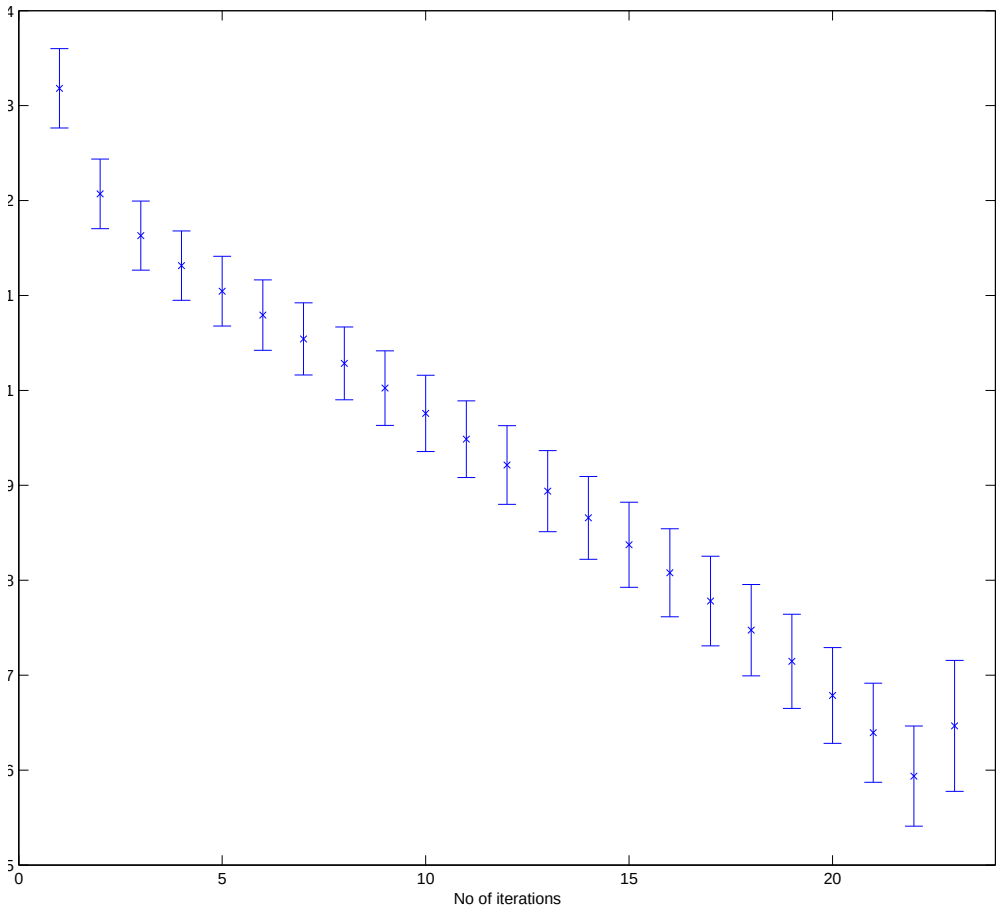


Figure 3.3. Exp. A3: Combined Band MSE; EM-CVX; Select 25 out of 200 frequency bands, 20 dB SNR across all frequency bands

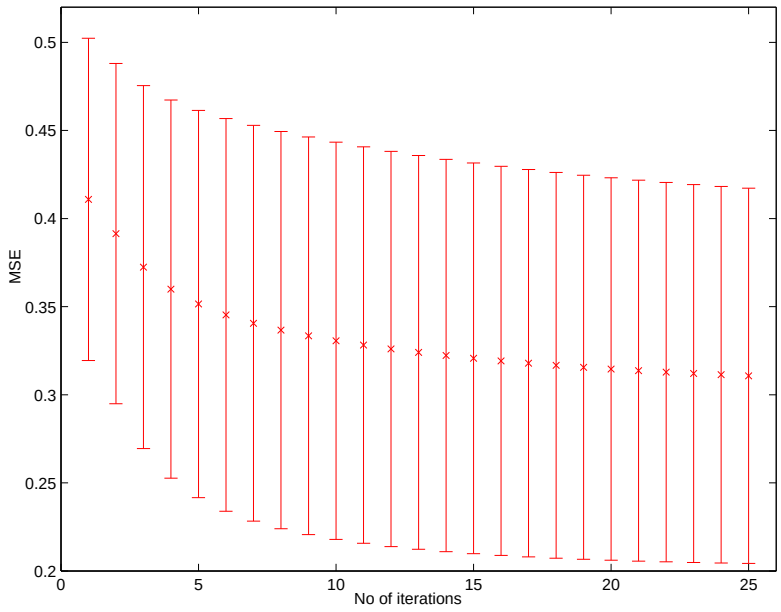


Figure 3.4. Combined Band MSE; EM-CVX; Select 10 frequency bands, no removal of frequency bands

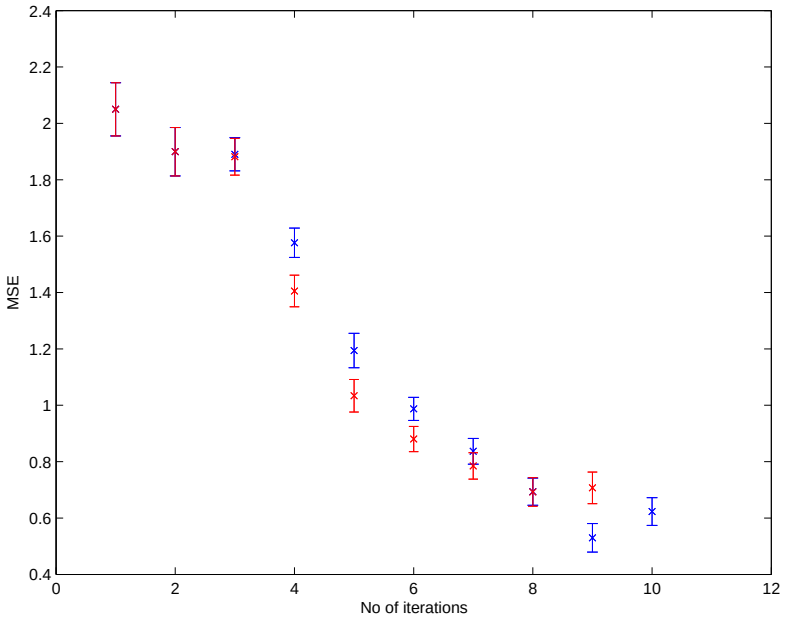


Figure 3.5. Exp. A5: Combined Band MSE; EM-CVX; Select 10 out of 50 frequency bands, 20, 10 ,5 dB SNR evenly distributed across all frequency bands for two arbitrary initialisation points

Chapter 4

Inferring Appropriate Bands To Find True Anomalies

IN Hyperspectral Imaging, a sensor with hundreds of sensor-elements gathers measurements of a spatial region of interest. The measurements collected by each sensor-element are considered a snapshot of the same spatial area but at a unique frequency referred to as a spectral band. In a scene, where there are sparse man-made materials (anomalies) relative to natural background, we show that it is possible to infer a ranking of the spectral bands that help identify anomalies. We consider the problem of simultaneously identifying anomalies and inferring spectral bands where anomalies vary significantly from background. Simultaneous identification of critical band frequencies that help identify anomalies, can lead to: improvements in sensor design, confirm the presence of true man-made materials and improve data throughput. In existing literature, eigen-decomposition methods alter the physical band-structure whilst unsupervised band-selection schemes do not guarantee the presence of anomalies. We develop a partition-based anomaly-clustering scheme that groups non-Gaussian measurements according to the extent to which the groups are divergent from one another. The SAGE condition guarantees groups are described by a locally optimal model whilst a convex relaxation scheme is used to evaluate suitability of group members. Simulations conducted demonstrate acceptable probability of detection, false-alarm-rates and band-ranking accuracy with synthetic non-Gaussian data with a sparse number of anomalies as well as real hyperspectral measurements gathered by the HyMap airborne sensor.

4.1 Introduction

A problem addressed in hyperspectral surveillance is the detection of sparse man-made materials relative to natural background such as soil and vegetation. This study is motivated by two specific surveillance problems:

1. The detection of visually similar man-made and natural materials such as green grass, green paint or camouflage.
2. Man-made materials which are small in size in relation to the pixel it occupies.

In both these cases, the man-made materials are likely to exhibit distinct differences in only a subset of bands. In the former case, this is due to the similarity of materials across certain wavelengths whilst in the latter the spectral measurement of the pixel is dominated by the surrounding background. In both cases, the exact bands across which the material varies the most is unknown apriori to measurement.

The problem of identifying critical bands is important for various reasons: (1) By virtue of the wavelengths identified, critical bands provide confirmation that at least some of the anomalies detected are anomalies and not noisy background, (2) a detection and false alarm metric associated with critical bands provides a quantitative way to assess the usefulness of the sensor-band design in terms of the ability to detect anomalies as well as improving the signal-to-noise ratio of critical bands, (3) the identification of critical bands allows the future designer to cue different bands for different tasks, which would also reduce the deluge of measurements collected in an intelligent manner, (4) schemes such as Principal Components Analysis (PCA) provide a linear-transformation such that maximum information content lies in an ordered set of principal components. This enables data compression since the latter set of components can be discarded. However, PCA does not guarantee explicitly or implicitly that the data transformation results in anomalies being found. Furthermore, the physical band-structure of the data is altered which means we do not have any information regarding the utility of each band. Metrics such as probability of detection (P_D) and false alarm rate (FAR) used in this study provide the hyperspectral data analyst a more tangible indication of the efficacy of the band-ranking step.

4.2 Existing Work

The following studies are studies carried out in Spectral Band Selection as well as Sensor or Feature Selection from signal processing and machine learning communities. They are representative of the work carried out in these fields, hence the list is not exhaustive:

4.2.1 Band Selection Criteria

Paper: Du *et. al.* in [33]
Contribution: Unsupervised Band Selection Using Compositional Models and Matrix Factorisation

Du *et. al.* assume a sub-pixel compositional model of hyperspectral data, where each pixel is a convex combination of a set of pure materials at varying fractions, where the pure materials are unknown but the maximum possible number of materials is fixed. They measure the contribution of each band towards re-constructing pure materials for each pixel measured. We use Gaussian mixture models to describe HSI data and attempt to identify bands that may reveal the presence of anomalies, whereas Du et.al. use pixel re-construction accuracy as opposed to anomaly detection accuracy as the band selection criteria. Nonetheless, both methods are unsupervised.

Paper: Keshava *et. al.* in [34]
Contribution: Supervised Combinatorial Band Selection Using Spectral Angle

Keshava *et. al.* in [34] add a band to a critical band set if the spectral angle between a reference and a test signature is maximised due to the addition of the band to the critical band set. Various two band combinations are used as initial bands in the set. The method relies on prior knowledge of a target spectral library to be used as a reference which is not the case in our method. We use convex relaxation to reduce computational

burden in band selection which is not the case in [34], which relies on cycling through many combination of bands.

Paper: Guo *et. al.* in [31]
Contribution: Information Theoretic Criteria for Band Selection

Guo *et. al.* visually identify spectral bands that show the greatest separability between pixels across a region of interest. A discretised reference map across each band in the identified range is created for the entire scene by averaging the class membership revealed by each band measurement in the identified range across each pixel. The mutual information between the averaged reference and each band measurements is estimated for the entire scene, where the maximum mutual information is revealed by bands which show the greatest similarity to the estimated reference map. Our proposed technique explicitly handles scenarios where there are sparse number of man-made materials in the scene whereas such information maybe lost during the averaging process suggested by Guo *et. al.* in [31]. We seek bands where the mutual information between potential targets and backgrounds is a minimal, without the use of reference maps.

Paper: Stein *et. al.* [14]
Contribution: Unsupervised Band Selection using Likelihood Ratio

Stein *et. al.* carry out hypothesis tests to determine whether a pixel belongs to a multivariate Gaussian background estimated from the scene or a multi-variate target distribution. Bands are removed until the likelihood ratio for a set of pixels in the scene is maximised. Although similar in spirit, we do not assume any knowledge of the target library and apply a convex relaxation to overcome combinatorial complexity. Furthermore, we also assume a Gaussian mixture as opposed to multivariate Gaussian used by [14].

Paper: Law et. al [38] (Machine Learning)
Contribution: Unsupervised Feature Selection for Gaussian Mixtures using Model Identifiability Criteria

Law *et. al.* improve identifiability of Gaussian components by removing bands or measurement features. Our method varies due to the fact that we apply a divergence criteria to estimate unique groups of Gaussian mixtures and seek the bands that again maximise divergence between a cumulative larger group of mixtures, where we are also able to identify measurements that are anomalies as well as identifying critical bands. We do not seek to remove bands but merely rank them.

4.2.2 Band Selection or Reduction Process

Paper: Green *et. al.* in [7]
Contribution: Band Reduction via Eigenvector Analysis

Propose a method called MNF (Maximum Noise Fraction) that maximises measurement/scene SNR through eigenvector analysis and a noise covariance, from which measurement dimensionality and thus data size can be reduced. However, the method relies on the rotation of the measurement axes which does not say anything about a band's utility or contribution in finding anomalies which is possible through our work. Moreover, the technique requires an analyst to carefully select a region of interest for estimating noise covariance, this is not required for our method.

Paper: Joshi *et. al.* [43] (Signal Processing)
Contribution: Unsupervised Sensor Selection for Gaussian Measurements using Convex Relaxation

Joshi *et. al.* in [43] apply a relaxation to the combinatorial problem of selecting an optimal P out of N correlated measurements which improves the estimation accuracy

of the system parameter. We formulate a relaxed likelihood ratio test to group measurements as opposed to minimising the determinant of an error covariance ellipsoid which cannot be directly derived for non-Gaussian data with unknown parameters. Nonetheless, we also apply convex optimisation in our proposed method, to overcome combinatorial complexity.

Paper: Griffiths *et. al.* [46] (Machine Learning)
Contribution: Unsupervised Band Selection for Compositional Models using Bayesian Methods

The Indian Buffet Process proposed by [46] estimates the distribution of a sparse binary random matrix, where D rows correspond to bands and columns correspond to K latent classes. Each element (d,k) corresponds to the relevance of each kth latent probability model to the dth band for all spectral measurements collected in the scene. The optimal number of bands to represent the each pixel is modelled as a random variable where prior knowledge can influence the exact number of components, this is not the case in our work. The technique also requires each latent class be a D dimensional latent factor which may suffer from over-fitting for large values of D, which is not the case in our proposed method in this chapter. Nonetheless, the method offers a unique alternative to carry out online band selection.

4.2.3 Summary of Work and Contributions

We develop an algorithm to carry out detection of anomalies through the inference of a subset of critical bands. Please refer to Definitions $1-7$ for an elaboration of the italicised terms in this paragraph and to Fig. 4.1 for a visual representation of the proposed technique. The standard EM approach to finding maximum likelihood estimates of high-dimensional non-Gaussian data suffers from parameter complexity and requires dimensionality reduction methods such as PCA to find a low-dimensional mapping before carrying out the clustering. In this study, (1) the model proposed is framed such

that local convergence of high-dimensional parameter estimates to maximum likelihood values is still achieved without a rotation of the principal axes and a reduction in measurement-dimensionality. We partition the measurement space in each *band-subset* and estimate parameters for *outlier* and *partial background measurements* using a Gaussian mixture model. By applying the SAGE condition to the problem, we show that the algorithm is still a standard EM algorithm [24] with the same implicit theoretical guarantees as the standard EM for maximum likelihood convergence of Gaussian Mixture parameter estimates. In the hyperspectral domain, to the author's knowledge, the model is the first of its kind where standard EM convergence guarantees apply without any physical alteration to the measurements' axes. (2) We formulate a novel scoring scheme using convex optimisation to establish membership of each *spectral measurement* in a band-subset to an outlier or partial background. Convex relaxation methods have been previously applied for combinatorial measurement selection by Joshi *et. al.* in [43], however the models were fully Gaussian and the cost function to evaluate an optimal subset of measurements is an analytically derived error covariance ellipsoid. In this study, we apply convex relaxation to a measurement selection scenario where the measurements are non-Gaussian and an error covariance ellipsoid cannot be derived. (3) To evaluate class membership and estimate outlier and partial background classes, we maximise a value proportional to the KL divergence between outlier and partial background classes with respect to an indicator matrix with convex constraints and then estimate class parameters using SAGE-EM. To the author's knowledge, this is the first instance of a convex-relaxation and KL divergence formulation used in the context of anomaly detection in a multi-dimensional non-Gaussian setting. (4) Subsequently we evaluate whether *a pixel*, which contains a combination of outlier and partial background measurements belongs to a *anomaly* or a standard background class. The constraints used in the optimisation problem enables class membership evaluation to be carried out based on a *critical band* rank which also enables the system designer to build in prior band knowledge for detecting anomalies. The information may include sensor band correlation, atmospheric bands and knowledge about band-subsets that are more useful in identifying metals, paints, etc. Such prior knowledge is useful for

maximising P_D and FARs. This feature is the first of its kind in hyperspectral band ranking schemes.

4.3 Methodology

In this section, we first setup the problem by proposing a statistical model that describes Hyperspectral data in terms of *outliers* and *partial backgrounds*. We define both these terms as well as providing criteria and justification for evaluating both these labels. We subsequently use this model to infer *anomalies* and *critical frequency bands*, where the problem is described in-terms of a constrained optimisation formulation.

Our solution comprises of the following features : (1) We resolve for outlier and partial background labels using Mahalanobis distance and convex relaxation in section 4.3.2. (2) We estimate the intractable model parameters using EM and show that standard convergence guarantees still apply in section 4.3.3 (3) We show how the procedure for determining unknown labels and model parameters is a maximisation of the KL divergence between outlier and partial background models in section 4.3.4 (4) Using these parameter estimates we again apply a convex relaxation with respect to an indicator variable to find anomalies and critical bands in section 4.3.5. We consider these unknown man-made materials as *anomalies* and the bands that reveal their presence as *critical bands*.

4.3.1 Problem Formulation

We first propose a novel representation of hyperspectral data. A hyperspectral image is three-dimensional data cube which contains signals measured by D spectral frequency bands across $X - Y$ spatial co-ordinates. Consider a spatial window within the $X - Y$ co-ordinates containing a total of N pixels determined by an operator, the D bands in this region can be partitioned into $r = 1 \dots R$ band-subsets containing $Q(r)$ contiguous bands in each subset where each rth band-subset is independent of the remaining

$R-1$.

This is possible due to the spectral correlation properties of the sensor and material correlation assumptions discussed in the Introduction. The measurement values across each band-subset can be described by random variable $Y(r)$, that describes the N pixel measurement values $y_1(r)\dots y_N(r) \in \mathbb{R}^{Q_r}$ in the spatial window across the rth band-subset. If $Y(r)$ is described by a statistical model, each nth pixel measurement across the rth band-subset can either be labelled as either an *outlier* of this model or a *partial background*.

Definition 9: An *outlier* lies on the tails of the probability distribution describing $Y(r)$ and *partial backgrounds* lie closer to the mean of this distribution. The term partial is used since we are referring to pixel measurements across a band subset and not the full set of bands. Outliers in real world are likely to be partial measurements of rare background materials in the window, noisy backgrounds or man-made anomalies.

Definition 10: *Anomalies* are pixels that are sparse in number and contain a number of spectral measurements that are outliers as depicted in Fig.4.1b. The degree of sparsity and number of outliers to deem a pixel a anomaly is determined by a likelihood ratio test between outlier and partial backgrounds across all bands. The P highest ratio values refer to P anomalies. On the corollary, a pixel that contains predominantly partial backgrounds is most likely to be natural background such as vegetation.

Definition 11: *Critical bands* are bands that contain the most number of outliers.

If $\theta(r) \in \Theta$ is a parameter describing $Y(r)$, the likelihood of $\theta(r)$ is given by,

$$\begin{aligned} L(y_1\dots y_R;\theta(r)) \quad = \quad & \prod_{n=1}^{N}\{\sum_{j=1}^{K}\pi_j(r)\mathcal{N}(y_n;\mu_j(r),\Sigma_j(r))^{\tau_n(r)} + \\ & \sum_{w=1}^{W}\pi_w(r)\mathcal{N}(y_n;\mu_w(r),\Sigma_w(r))^{\tau_n^c(r)}\}, \forall r = 1\dots R \end{aligned} \tag{4.1}$$

where, the term after the summation describes the mixture model for outliers and preceding the summation describing partial backgrounds, $\theta(r) = \{\mu_{j/w}(r) \in \mathbb{R}^{Q(r)}, \Sigma_{j/w}(r) \in \mathbb{R}^{Q(r)\times Q(r)}, \pi_{j/w}(r) \in \mathbb{R} \forall j = 1 \dots K, w = 1 \dots W\}$ are Gaussian Mixture model parameters, respectively, describing outlier and partial backgrounds across the rth band-subset, $\tau_n(r) \in \{0, 1\}$ and its complement denoted $\tau_n^c(r)$ is an indicator vector that says the nth pixel measurement across the rth band subset is described by either the outlier or partial background models. Please refer to the introduction for a justification of why we consider Gaussian mixtures as appropriate models for hyperspectral data. The Gaussian mixture parameter estimates as well as the indicator variables are unknown. The indicator variable is determined by,

$$\tau_n(r) = \begin{cases} 0 & \text{if } \sum_{d=1}^{Q(r)} t_{d,n}(r) \leq \rho * Q(r); \\ 1 & \text{if } \sum_{d=1}^{Q(r)} t_{d,n}(r) \geq \rho * Q(r), \end{cases} \tag{4.2}$$

where, ρ is a tolerance threshold empirically determined by the operator prior to the experiment, $t_n(r) \in \{0, 1\}^{Q(r)}$ is a vector weight determined by the following objective,

$$\begin{aligned} \hat{t}_n(r) &= \arg\max_{t_n(r)} \sum_w (t_n(r) \circ (y_n - \hat{\mu}_w(r)))^T \hat{\Sigma}_w^{-1}(r)(t_n(r) \circ (y_n - \hat{\mu}_w(r))) \\ &- \sum_j (t_n^c(r)(\circ y_n - \hat{\mu}_j(r)))^T \hat{\Sigma}_j^{-1}(r)(t_n^c(r) \circ (y_n - \hat{\mu}_j(r))), \\ &s.t. \quad t_n \in \{0, 1\} \forall r = 1 \dots R \end{aligned} \tag{4.3}$$

where $\hat{}$ denotes known estimates of mixture parameters, $\circ$ denotes the Hadamard or element-wise product. The value of $t_n(r) = 1$ when a measurement across a band is close to the measurement mean and 0 when they are further away, thus providing a vector indicator of the degree to which a pixel measurement across a band subset is an outlier or partial background. Thus, the expression in (4.3) measures how well pixel $y(n)$ is described by the model with respect to each frequency band.

The optimisation problem is equivalent to maximising the Mahalanobis distance between the outlier and partial background models which is desirable since it reduces the possibility of false alarms in anomaly detection. The justification for $t_n(r)$ being a vector as opposed to a scalar is that we get a richer metric to work with in-terms of measuring the sensitivity of each frequency band before labelling it as an outlier

or partial background. This is especially relevant for atmospheric bands that affect all pixels. Contribution of known atmospheric bands in determining whether a pixel measurement across a band-subset is an outlier can be reduced without having to remove the bands altogether in a supervised manner. Thus, a set of S outliers in the rth band-subset is given by, $z_1^{\tilde{A}}(r)\dots z_S^{\tilde{A}}(r) \in Z(r) \subset \mathbb{R}^{Q(r)}$, where $Z(r) = \tau_n(r) \circ Y(r) \in \mathbb{R}^{Q(r)}$ is random variable representing the outliers and $\tau_n(r) = 1 \forall s = 1 \dots S(r)$ unknown pixel measurements from a total N. Alternatively the $Z^c(r)$ represents all $N - S(r)$ measurements, $z_1^{\tilde{B}}(r)\dots z_{N-S(r)}^{\tilde{B}}(r) \in Z^c(r) \subset \mathbb{R}^{Q(r)}$, that are partial backgrounds. The parameter estimates for these random variables are unknown. We denote the respective parameter estimates as $\hat{\theta}^{\tilde{A}}(r)$ and $\hat{\theta}^{\tilde{B}}(r)$, where $\hat{\theta}(r) = \{\hat{\theta}^{\tilde{A}}(r), \hat{\theta}^{\tilde{B}}(r)\} \forall r = 1 \dots R$.

4.3.2 Labelling Outliers and Partial Backgrounds using Convex Relaxation

In this section we solve for $\hat{t}_n(r) \forall r = 1 \dots R$ for the problem stated in (4.3) and propose an iterative algorithm to label pixel measurements across each rth band subset. We assume that $\hat{\theta}(r) = \{\mu_{j/w}(r) \in \mathbb{R}^{Q(r)}, \Sigma_{j/w}(r) \in \mathbb{R}^{Q(r)\times Q(r)}, \pi_{j/w} \in \mathbb{R} \forall j = 1 \dots K, w = 1 \dots W\}$ have been found in a separate step.

Solving for vector $t_n(r)$ from the objective stated in (4.3) is an intractable problem since the co-ordinates of $t_n(r)$ are correlated spectrally which means the value of a single co-ordinate in t_n is dependant on the values of $Q(r) - 1$ remaining co-ordinates which means there are a total of $\binom{Q(r)}{Q(r)-1} \times \dots \binom{Q(r)}{1} \times R \times N$ calculations for each window of pixels. If total number of pixels, N is in the order of thousands and if R and $Q(r)$ are large, combinatorial strategies of evaluating membership become computationally intractable for an entire scene which can have up to hundred thousand pixels. Therefore, we relax the constraints and solve the problem using convex optimisation, where $t_n(r) \in [0,1]^{Q(r)}$ is a convex in equality and the objective is a quadratic function of $t_n(r)$ which is concave. Both these factors mean, a solution to the problem via convex optimisation can be found [15] through semi-definite programming or interior point

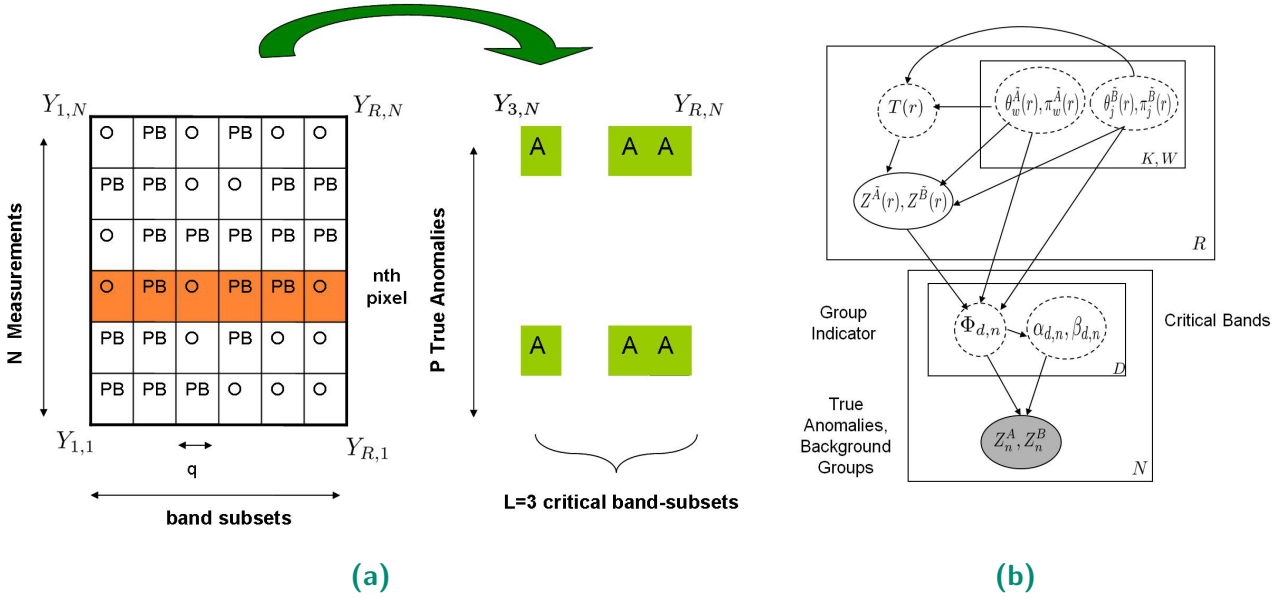


Figure 4.1. a) Proposed algorithm simultaneously identifies critical band-subsets which reveal the presence of anomalies. Outlier (O) and partial backgrounds (PB) measurements are first identified across R band-subsets. The diagram on the right shows $P < N$ anomaly pixels (A) that produce the greatest KL divergence between PB and O distributions using a subset of bands from a critical band rank. b) The graphical model describes the generative process for each r-th band-subset which contains $Q(r)$ bands. Indicator variable $T(r)$ indicates the membership of a spectral sample from the r-th band-subset to outlier, $Z(r)$ and partial background $Z(r)^c$ subsets. The full circles are random variables whilst the square plates around the circles indicate number of measurements, bands or components of the variable in the circle whilst dotted circles represent parameters that are non-random variables. A measurement window contain a maximum of P anomalies where anomalous pixels exceed $P+1$ thresholds. Both anomalies and critical bands are derived from convex matrix Φ which is restricted by inequality constraints λ_1, λ_2. Z_n^A, Z_n^B are binary matrices that indicate membership of the nth pixel to anomaly and background groups.

methods. The revised objective is given by,

$$\begin{aligned} \hat{t}_n(r) &= \arg\max_{t_n(r)} \sum_w (t_n(r) \circ (y_n - \hat{\mu}_w(r)))^T \hat{\Sigma}_w^{-1}(r)(t_n(r) \circ (y_n - \hat{\mu}_w(r))) \\ &- \sum_j (t_n^c(r) \circ (y_n - \hat{\mu}_j(r)))^T \hat{\Sigma}_j^{-1}(r)(t_n^c(r) \circ (y_n - \hat{\mu}_j(r))), \\ s.t. \quad & t_n \in [0,1] \forall r = 1 \ldots R \end{aligned} \tag{4.4}$$

The idea is to alternate between finding the parameters $\hat{\theta}(r)$ and $t_n(r) \forall n = 1 \dots N$ for each rth band-subset.

1. Assume all measurements, $y_1(r) \dots u_N(r)$ across each rth band subset are partial backgrounds. Estimate $\hat{\theta}^{\tilde{B}(l)}(r) \forall k = 1 \dots K$ fixed number of components, where l denotes the lth iteration of the algorithm. However, assume a fixed proportion of these measurements are outliers according to the likelihood score and estimate $\hat{\theta}^{\tilde{A}(l)}(r)$.

2. Using parameter estimates $\hat{\theta}^{l}(r)$ estimate, $t_n(r)$ from (4.4) $\forall n = 1 \dots N$ pixel measurements.

3. Solve for $\tau_n(r) \forall n = 1 \dots N$ from (4.2) and compute membership of pixel measurement to the outlier $Z(r)$ or partial background class $Z^c(r)$

4. Solve for $\hat{\theta}^{l+1}(r)$ using newly labelled measurements from the previous step.

5. Repeat steps $2 - 4$, $l = 1 \dots \tilde{L}$ times or until $\sum_s t_s(r) \forall s = 1 \dots S(r)$ pixel measurements is maximum.

6. Repeat steps $1 - 5 \forall r = 1 \dots R$ band-subsets

Note that we do not repeat these steps until the likelihood is maximum since the desirable scenario is for tight clusters, that is when the outliers and partial backgrounds are as close as possible to the component means which results in a small variance. This does not correspond to a high likelihood value, since the likelihood is greater when the model tries to explain all measurements. Nonetheless, since we repeatedly switch between estimating the partial background and the outlier models, there is convergence in the combined likelihood of both models which further justifies the use of a relaxation approach on likelihood.

In section 4.3.3, we obtain maximum likelihood estimates for $\hat{\phi}$, where, $\{\hat{\phi} = \hat{\theta}(1) \dots \hat{\theta}(R)\}$ are the parameter estimates for R band-subsets, where $\hat{\phi}$ parameterises the selected window of pixels. This is denoted by random variable $Y \in \mathbb{R}^D$ with D bands.

4.3.3 Maximum Likelihood Estimation of Gaussian Mixture Band-Subsets

We seek the maximum likelihood estimates of $\hat{\phi}$ to establish a sense of optimality for parameters describing the scene. In this section we show how maximum likelihood guarantees are attained as well as stating the equations to derive the parameter estimates. We assume that the sets $Z(r), Z^c(r)$ are derived from a previous step and fixed throughout the parameter estimation process.

Parameter estimation in Gaussian mixture models cannot be performed analytically but only iteratively, if membership of data to Gaussian components in the model is unknown. We refer to such datasets as un-categorised. The model parameters often fail to converge to an asymptotic value [24] if parameter complexity is high as is the case with high-dimensional hyperspectral data. This phenomena is addressed by Fessler and Hero in [24], who reduce the data into smaller partitions when the measurement space is i.i.d. The chosen subset is smaller, less-informative and is modified in-between EM iterations. The authors demonstrate that analytic convergence of likelihood is achieved in-spite of a modification to measurement the space as long as the chosen partitioned subset adheres to certain conditions.

From chapter 3, if $\bar{Y}$ represents all pixel measurements described by Y and a partitioned subset of measurements described by $Z^S \subset \bar{Y}$ are independent, the **SAGE condition** states that the probability of the entire measurement space Y is given by the product of conditional of Y given Z^S and the probability of Z_S,

$$P_Y = P_{Y|Z^S}(;\theta^{\tilde{S}})P_Z(;\theta^S) \tag{4.5}$$

where, θ^S is a parameter estimate characterising Z^S measurements, $\theta^{\tilde{S}}$ represents the parameterisations of the remaining measurements in $\bar{Y}$. Its worth re-emphasising that the condition only holds true as long as measurements in Z^S and Y are independent. In this chapter, we restrict the choice of Z^S to represent either the r-th partial background

or outlier measurements, assuming that the labels are known, where Z^S is a permissible subset since $y_n \forall n = 1 \dots N$ pixels are are i.i.d, and enable partitioning within the band-subset and the partition between the r-th band-subset and the remaining $R-1$ band-subsets are independent. Therefore, $\theta^S = \hat{\theta} \in \{\hat{\theta}^{\tilde{A}}, \hat{\theta}^{\tilde{B}}\}$. Partitioning within a band-subset was not carried out in Chapter 2.

We now state the likelihood equation describing all measurements in Y and an EM algorithm to estimate $\hat{\theta}(r)$ for any r-th band-subset. The likelihood of ϕ given the measurement space Y is given by the product of likelihoods,

$$L(y_1 \dots y_N; \phi) = L(y_1(r) \dots y_{S(r)}(r); \theta^S) \times L(y_1(r' \neq r) \dots y_N(r' \neq r); \theta^{\tilde{S}}), \tag{4.6}$$

where, $r' \neq r$ refers to the remaining $R-1$ band subsets, the parameter ϕ characterises all measurement values described by Y. We maximise the likelihood in equation (4.6), w.r.t to θ^S which means the latter term after the product can be ignored. We have,

$$\begin{aligned} &\arg\max_{\theta^S \in \Theta} \log(L(y_1(r) \dots y_N(R); \theta^S) L(y_1(r' \neq r) \dots y_N(r' \neq r) \theta^{\tilde{S}})) \equiv \\ &\arg\max_{\theta^S \in \Theta} \log(L(y_1(r) \dots y_N(R); \theta^S)), \end{aligned} \tag{4.7}$$

which, is the maximisation performed in the standard EM. We alternate the measurement subsets Z^S between outlier and partial background in-between EM iterations in each r-th band-subset for all band-subsets. If $Z^S = z_1^{\tilde{A}}(r) \dots z_S^{\tilde{A}}(r)$, the likelihood of $\hat{\theta}^{\tilde{A}}(r)$ is given by,

$$L(y_1(r) \dots y_N(R); \theta^S) = L(Z(r); \hat{\theta}^{\tilde{A}}(r)) = \prod_{s=1}^{S(r)} \sum_{w=1}^{W} \pi_w^{\tilde{A}}(r) \mathcal{N}(z_s^{\tilde{A}}(r) | \mu_w^{\tilde{A}}(r), \Sigma_w^{\tilde{A}}(r)) \tag{4.8}$$

If we knew which component in the mixture each measurement came from, then analytic maximisation of (4.8) is straightforward because the log likelihood function is Gaussian. This defines the so-called *complete data* - the measurements augmented by the knowledge of which component of the mixture each measurement comes from. , the complete data likelihood of $\hat{\theta}^{\tilde{A}}(r)$ is given by,

$$L_c(Z(r); \hat{\theta}^{\tilde{A}}(r)) = \prod_{s=1}^{S(r)} \prod_{w=1}^{W} \pi_w^{\tilde{A}}(r)^{v_{w,s}(r)} \mathcal{N}(z_s^{\tilde{A}}(r) | \mu_w^{\tilde{A}}(r), \Sigma_w^{\tilde{A}}(r))^{v_{w,s}(r)} \tag{4.9}$$

where $v_{w,s}(r) \in \{0,1\}$ and represents the membership of the s-th sample to the wth component in the r-th band-subset. The *complete* data log-likelihood (neglecting constant terms not dependent of the parameters and assuming that $\pi_w(r) > 0$) is thus,

$$\log(L_c(Z(r);\hat{\theta}^{\tilde{A}}(r))) = \ell_c(\theta(r)^{\tilde{A}}) = \sum_{s=1}^{S(r)} \sum_{w=1}^{W} v_{w,s}(r) \ldots$$
$$\left(\log \pi_w(r) - 1/2 \left((z_s^{\tilde{A}}(r) - \mu_w(r))^T \Sigma_w^{-1}(r)(z_s^{\tilde{A}}(r) - \mu_w(r))\right)\right) \tag{4.10}$$

Since the membership of V is unknown the expected value of V is computed before estimating $\theta(r)^{\tilde{A}}$. This forms the iterative EM algorithm which consists of two steps. For the mth iteration, the E step involves computing the conditional expectation of $\ell_c(\theta^{\tilde{A}}(r))$ given measurements $z_s^{\tilde{A}}(r)$ and parameter estimates from the previous iteration $\theta^{\tilde{A}(m-1)}(r)$. The expected value of V is denoted by
$\mathrm{E}\left\{V_{w,s}^{(m)}(r)|z_1^{\tilde{A}}(r),\ldots,z_{S(r)}^{\tilde{A}}(r),\hat{\theta}^{\tilde{A}(m-1)}(r)\right\}$ and it is straightforward to show [20] that,

$$\mathrm{E}\left\{V_{w,s}^{(m)}(r)|z_1^{\tilde{A}}(r),\ldots,z_{S(r)}^{\tilde{A}}(r),\hat{\theta}^{\tilde{A}(m-1)}(r)\right\} = p_{w,s}^{(m)}(r) =$$
$$\frac{\hat{\pi}_w^{(m-1)}(r)\ \prod_{s=1}^{S(r)} \mathcal{N}(z_s^{\tilde{A}}(r);\hat{\mu}_w^{(m-1)}(r),\hat{\Sigma}_w^{(m-1)}(r))}{\sum_{w=1}^{W} \hat{\pi}_w^{(m-1)}(r)\ \prod_{s=1}^{S(r)} \mathcal{N}(z_s^{\tilde{A}}(r);\hat{\mu}_w^{(m-1)}(r),\hat{\Sigma}_w^{(m-1)}(r))}. \tag{4.11}$$

The M step computes maximising argument for each m-th estimate of the expected complete log-likelihood,

$$\mathrm{E}\left\{\ell_c(\hat{\theta}^{\tilde{A}(m)}(r))|z_1^{\tilde{A}}(r),\ldots,z_{S(r)}^{\tilde{A}}(r),\theta^{\tilde{A}(m-1)}(r)\right\} =$$
$$\sum_{w=1}^{W}\sum_{s=1}^{S(r)} p_{w,s}^{(m)}(r)\log\hat{\pi}_w^{(m-1)}(r) - \frac{p_{w,s}^{(m)}(r)}{2} \times$$
$$(z_s^{\tilde{A}}(r) - \hat{\mu}_w^{(m-1)}(r))^T \hat{\Sigma}_w^{-1(m-1)}(r)(z_s^{\tilde{A}}(r) - \hat{\mu}_w^{(m-1)}(r)). \tag{4.12}$$

The *m+1*-th parameter update is given by

$$\hat{\pi}_w^{(m+1)}(r) = \frac{1}{S(r)}\sum_{s=1}^{S(r)} p_{w,s}^{(m)}(r),\ \hat{\mu}_w^{(m+1)}(r) = \frac{\sum_{s=1}^{S(r)} p_{w,s}^{(m)}(r) z_s^{\tilde{A}}(r)}{\sum_{s=1}^{S(r)} p_{w,s}^{(m)}(r)}$$
$$\hat{\Sigma}_w^{(m+1)}(r) = \frac{1}{S(r)}\sum_{s=1}^{S(r)}\sum_{w=1}^{W} p_{w,s}^{(m)}(r) AA^T \tag{4.13}$$

where, $A = (z_s^{\tilde{A}}(r) - \hat{\mu}_w^{(m+1)}(r))$. These two steps are repeated using the estimates (4.13) as the values of the parameters in (4.11) for the next iteration. It is well-known that the expected-complete log-likelihood in (4.12) increases monotonically until a local

maximum is found. So, now we have a complete description of set $\{Z(r), V\}$. Similar update equations are applied to partial backgrounds, where samples from $Z^c(r)$ rather than $Z(r)$ and $\theta^S = \hat{\theta}^{\tilde{B}}(r)$, with K Gaussian components instead of W.

4.3.4 A Kullback-Leibler Divergence for Maximising Partially Labelled Gaussian Mixtures

In this subsection we show that the algorithm for determining outlier and partial background distributions listed in section 4.3.2 is equivalent to maximising the Kullback-Leibler divergence between these distributions. This provides a sense of optimality to these distributions.

Consider the expected log likelihood in equation (4.12). This expression can be parameterised as

$$\begin{aligned}
&\mathrm{E}\left\{\ell_c(\hat{\theta}^{\tilde{A}(m)}(r))|\tau_1(r)\circ y_1(r),\ldots,\tau_N\circ y_N(r),\theta^{\tilde{A}(m-1)}(r)\right\} = \\
&\sum_{w=1}^{W}\sum_{n=1}^{N} p_{w,n}^{(m)}(r)\log\hat{\pi}_w^{(m-1)}(r) - \frac{p_{w,n}^{(m)}(r)}{2}\times \\
&(\tau_n(r)\circ y_n(r) - \hat{\mu}_w^{(m-1)}(r))^T\hat{\Sigma}_w^{-1(m-1)}(r)(\tau_n(r)\circ y_n(r) - \hat{\mu}_w^{(m-1)}(r)). \qquad (4.14)
\end{aligned}$$

where the N values of $\tau_n(r)$ are known. Since $\tau_n(r) = \sum_d^{Q(r)} t_{d,n}(r)$, we can say that maximising the difference between the expected complete outlier and partial background likelihoods with respect to τ_n is proportional to maximising the objective in (4.4). By the monoticity property of expectations, the objective is equivalent to

$$\begin{aligned}
&\mathrm{E}\left\{\ell_c(y_n(r);\hat{\theta}^{\tilde{A}}(r),t_n(r))\right\} - \mathrm{E}\left\{\ell_c(y_n(r);\hat{\theta}^{\tilde{B}}(r),t_n^c(r))\right\} \\
&= \mathrm{E}\left\{\ell_c(y_n(r);\hat{\theta}^{\tilde{A}}(r),t_n(r)) - \ell_c(y_n(r);\hat{\theta}^{\tilde{B}}(r),t_n^c(r))\right\} \\
&= \mathrm{E}\left\{\log L(y_n(r);\hat{\theta}^{\tilde{A}}(r),t_n(r)) - \log L(y_n(r);\hat{\theta}^{\tilde{B}}(r),t_n^c(r))\right\} \\
&= \mathrm{E}\left\{\log\frac{L(y_n(r);\hat{\theta}^{\tilde{A}}(r),t_n(r))}{L(y_n(r);\hat{\theta}^{\tilde{B}}(r),t_n^c(r))}\right\} \\
&= D(P(y_n;\hat{\theta}^{\tilde{A}}(r),t_n(r))||P(y_n;\hat{\theta}^{\tilde{B}}(r),t_n^c(r))) \qquad (4.15)
\end{aligned}$$

where, $P(y_n;\hat{\theta}^{\tilde{A}}(r),t_n(r))$ and $P(y_n;\hat{\theta}^{\tilde{B}},t_n^c(r)))$ are respectively the outlier and partial background probability distributions of the nth pixel measurement from the rth band

subset. The distributions are parameterised by the indicator vector t_n instead of τ_n due to the proportionality established above. $D()$ denotes the Kull-back Leibler divergence due to the expectation and log-likelihoods. Maximising the Mahalanobis distance between measurements described by fully-categorised Gaussian Mixtures is proportional to maximising the Kullback Leibler divergence with respect to outlier and partial background models for un-categorised Gaussian Mixtures.

For scalable solutions we consider optimising for all N samples without any iterative means. Maximising the expression in (4.15) is equivalent to the following,

$$\begin{aligned}\hat{T}(r)^{(m+1)} &= \arg\min_{T(r)} \mathbf{tr} \sum_{w=1}^{W} (\hat{\Sigma}_w^{-\frac{1}{2}(m)} T(r) \circ (Y(r) - \vec{\hat{\mu}}_w^{\tilde{A}(m)}))^T \hat{\Sigma}_w^{-\frac{1}{2}(m)} T(r) \circ (Y(r) - \vec{\hat{\mu}}_w^{\tilde{A}(m)}) \\ &- \frac{1}{2} \sum_{j=1}^{K} (\hat{\Sigma}_j^{-\frac{1}{2}(m)} T(r)^c \circ (Y(r) - \vec{\hat{\mu}}_j^{\tilde{B}(m)}))^T \hat{\Sigma}_j^{-\frac{1}{2}(m)} T(r)^c \circ (Y(r) - \vec{\hat{\mu}}_j^{\tilde{B}(m)}) \\ &\text{s.t } T(r) \in [0,1]^{Q(r) \times N} \end{aligned} \tag{4.16}$$

where, $T(r) \in [0,1]^{Q(r) \times N}$ indicator matrix, $Y(r)$ represents all N measurements from the r-th band-subset and $\vec{\hat{\mu}}$ represents N copies of the mean. A larger value of $T(r)$ results in the minimisation of the expression, where the complement, $T(r)^c$ minimises the objective further and thus maximises the divergence. This improves the certainty in membership, where values of $T(r)$ are either maximum or minimum subject to a good fit of the model being maintained.

The equation in (4.16) contains the trace of two terms, let $\sum_w A_w^T A_w$ represent the first expression before the addition operator and $\sum_j B_j^T B_j$ represents the second term. Since the trace of a square is equivalent to the square of the Frobenius norm and the triangle in-equality, we have,

$$\begin{aligned}\mathrm{tr}(\sum_w A_w^T A_w + \sum_j B_j^T B_j) &= \left\| \sum_w A_w \right\|_F^2 + \left\| \sum_j B_j \right\|_F^2 \\ &= \left\| \sum_w A_w + \sum_j B_j \right\|_F^2 \leq \left\| \sum_w A_w \right\|_F^2 + \left\| \sum_j B_j \right\|_F^2 \end{aligned} \tag{4.17}$$

We use this lower-bound to simplify the optimisation and thus maximise the following expression,

$$\begin{aligned}\hat{T}^{(m+1)} &= \arg\max_{T(r)} -\| \sum_{w=1}^{W} \hat{\Sigma}_w^{-\frac{1}{2}(m)}(r)T(r) \circ (Y(r) - \vec{\hat{\mu}}_w^{\tilde{A}(m)}(r))^T \hat{\Sigma}_w^{-\frac{1}{2}(m)}(r)T(r) \circ \\ &\quad (Y(r) - \vec{\hat{\mu}}_w^{\tilde{A}(m)}(r)) + \sum_{j=1}^{K} \hat{\Sigma}_j^{-\frac{1}{2}(m)}(r)T^c(r) \circ (Y(r) - \vec{\hat{\mu}}_j^{\tilde{B}(m)}(r))^T \circ \\ &\quad \hat{\Sigma}_j^{-\frac{1}{2}(m)}(r)T^c(r)(Y(r) - \vec{\hat{\mu}}_j^{\tilde{B}(m)}(r))\|_F^2 \end{aligned} \tag{4.18}$$

Finding a closed from estimate of $T(r)$ from the above expression is difficult [15] and hence, we consider convex optimisation methods to solve the problem. The stated objective is concave since the negative of a Frobenius norm is convex with respect to the measurements. Since $T(r)$ is also bounded by a convex inequality the solution can be found through iterative convex optimisation methods. We use the `CVX` toolbox by Grant *et. al.* [47] to solve the problem.

4.3.5 Anomaly Detection and Band Ranking Using Convex Relaxation

In this subsection, we shift our attention to finding anomalies and critical bands once we have the maximum likelihood estimates, $\hat{\theta}(r) = \hat{\theta}^{\tilde{A}}(r), \hat{\theta}^{\tilde{B}}(r) \forall r = 1 \dots R$ after M_r iterations $\forall r = 1 \dots R$ band-subsets. If random variable $Y \in \mathbb{R}^D$ represents the pixel measurements $y_1 \dots y_N$ across all R independent band-subsets, the statistical model describing Y is defined by the following hypotheses,

$$\begin{aligned} H_0 &: y_1 \dots y_N \sim \sum_{k=1}^{K} \widetilde{\pi}_k \mathcal{N}(y_n; \hat{\theta}_k^0, \Psi^c) \\ H_1 &: y_1 \dots y_N \sim \sum_{w=1}^{W} \widetilde{\pi}_w \mathcal{N}(y_n; \hat{\theta}_w^1, \Psi), \end{aligned} \tag{4.19}$$

where, $\hat{\theta}_{k/w}^{0,1} = \{\widetilde{\mu}_{k/w} \in \mathbb{R}^D, \widetilde{\Sigma}_{k/w} \in \mathbb{R}^{D \times D}, \widetilde{\pi}_{k/w} \in [0,1] \forall k = 1 \dots K \forall w = 1 \dots W\}$ and $\Psi \in [0,1]^{D \times N}$ matrix with vectors $[\psi_1 \dots \psi_N]$ plays a similar role as t_n did previously where each dth element of the nth pixel indicates whether the measurement is a member of outlier or partial background. Since the R band subsets are independent, each

element of the Gaussian mixture component mean is the estimated mean across each rth band-subset, thus $\widetilde{\mu}_{k,/w} = [\hat{\mu}_{k/w}(1) \dots \hat{\mu}_{k/w}(R)]^T$. Similarly each Gaussian component covariance is block-diagonal and consists of the estimated covariances from R band-subsets $\widetilde{\Sigma}_{k,/w} = \text{diag}(\hat{\Sigma}_{k/w}(1) \dots \hat{\Sigma}_{k/w}(R))$. We use approximated values for Gaussian component proportions, $\widetilde{\pi}_{k/w}$, where we compute the average value of the kth and wth Gaussian component across all R band-subsets.

If $\eta(Y; \hat{\phi}, \Psi) = \prod_{n=1}^{N} (\frac{H_1}{H_0})$ represents the likelihood ratio, an optimisation problem is setup to determine whether each pixel is an anomaly or a background,

$$\begin{aligned} \Psi^* &= \arg\max_{\psi} \eta(Y, \hat{\theta}, \Psi) \\ \text{s.t} \qquad & \lambda_1 <= \sum_d \sum_n \Psi_{d,n} <= \lambda_2, \Psi_{d,n} \in [0,1] \end{aligned} \tag{4.20}$$

where, the indicator matrices are represented by $\Psi, \Psi^* \in [0,1]^{D \times N}$. Let,

$$\alpha_n = \sum_d \Psi^*_{d,n}, \beta_d = \sum_n \Psi^*_n \tag{4.21}$$

Summing the rows of α_n indicates whether the nth pixel is an anomaly, whilst summing the columns of β_d indicates whether the dth band is a critical band. The lower bound λ_1 ensures a pixel with a few outliers is not considered an anomaly and similarly, the upper-bound λ_2 ensures that a pixel with many outliers is simply a noisy or dead pixel. The exact values for these are empirically determined and can evaluated using training data prior to on-board processing where they remain fixed. Thus, the likelihood of any pixel y_n being a anomaly is dependant on the size of α, which reflects the number of outlier measurements in each pixel,

$$P(y_n \in Z(r)) = \begin{cases} 1 & \text{if } \alpha > \alpha_0 \\ 0 & \text{otherwise} \end{cases}$$

$\forall n = 1 \dots N$ pixels, where $\alpha \in \mathbb{R}^N$ contains N values and α_0 refers to the $P+1$th largest value of α, P denotes the maximum number of anomalies we expect to find within a window. It can be set based on how conservative the practitioner wishes to be with the FAR or P_D. In experiments conducted, $P \in [1, 10]$. This means we expect no more than $P = 10$ anomalies per window, where there are $N \in [200, 289]$ pixels in each window.

Under this formulation, the values of α_n need not necessarily be evaluated over all D bands but across a subset M of which which we consider the L highest values of β_n across the M bands. Thus, $L \subset M \subset D$. In the experiments conducted M, L were restricted to a maximum of $M = 60$ and $L = 20$ bands respectively.

Thus, it can be seen that the technique can be used to fix a false alarm rate for each window, through the setting of P, but at the cost of missing anomalies. The setting of M is optional and can be carried out if the practitioners wishes to use some prior knowledge about the anomalies, sensor or scene itself. The L most critical bands provides a quantification of band utility with respect to P_D and FAR.

A summary of the entire algorithm is provided below,
Steps:

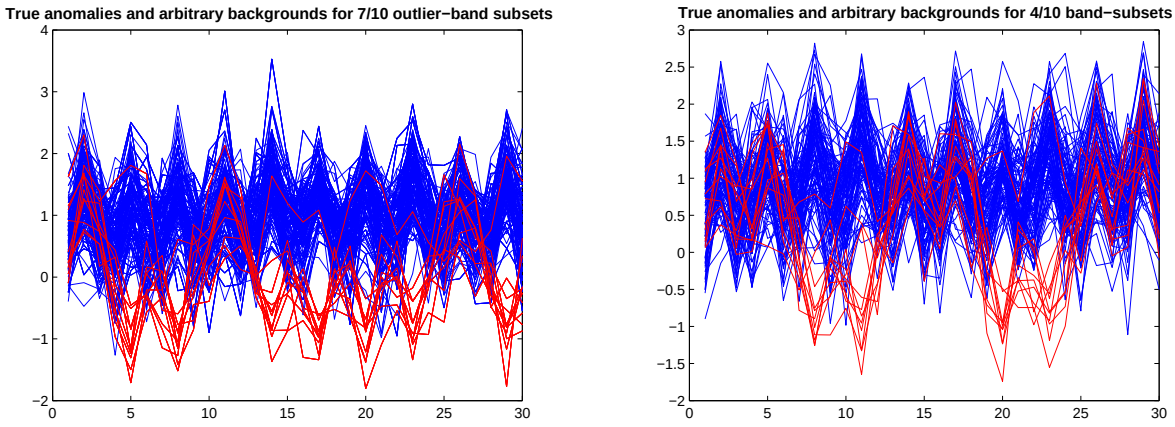
1. Let $Z_S = \bar{Y}$, where $Q(r)$ is set by the user $\forall r = 1 \dots R$.
2. Estimate $\hat{\theta}^{\tilde{B}(m+1)}(r)$ for l iterations using (4.13).
3. Calculate (4.12) for each n-th sample.
4. Let $z_1^{\tilde{A}}(r) \dots z_S^{\tilde{A}}(r)$ consists of samples with the S_r smallest scores.
5. Estimate $\hat{\theta}^{\tilde{A}(m+1)}(r)|Z_S = z_1^{\tilde{A}}(r) \dots z_S^{\tilde{A}}(r)$ for l iterations from (4.13).
6. Estimate $T^{(m+1)}$ using equations, (4.18), (4.2).
7. Repeat steps 2-6 until likelihood value in equation $\sum T^{(m+1)}$ converges or after $m = 1 \dots \tilde{M}$ iterations.
8. Estimate ψ^*, α, β to find anomalies and critical bands, using (4.20), (4.21). If any prior band knowledge is available, such as regions of potential contrast, such as those in experiment B, apply constraints across only those bands.
9. Set λ_1, λ_2 according to empirical experimentation based on prior knowledge of possible materials in the scene. Refer to section for further details on how these values were obtained for the experiments conducted.

10. Set P, M, L depending on tolerance to false alarms.

4.4 Experiments

4.4.1 Experiment A: Simulated Gaussian Mixture data

In this experiment we create a synthetic dataset with outliers spanning $M \subset D$ critical bands. We attempt to test the capacity of the proposed method in identifying anomalies as well as critical bands. Prior knowledge of critical bands and anomalies are subsequently used to illustrate performance. Experiments are repeated $e = 1 \dots E$ times, where each *eth* dataset consists of $N \times R$ independent samples drawn from R independent Gaussian Mixtures. The $Q(r)$ dimensional r-th band-subset consist either entirely of partial backgrounds or outliers that are drawn independently from $K = 2$ and $W = 2$ component Gaussian Mixtures, respectively. For every *eth* simulation and r-th band-subset, partial background and outlier means and co-variances are varied randomly, but the component weights are fixed for both experiments. Outliers are randomly generated M_e times for every *eth* simulation. Since we are interested in cases where only a subset of bands vary significantly, an upper bound is placed on M for all E simulations and possible number of outlier in each critical band is also fixed to a maximum of 10. Values for thresholds τ in equation (4.2) are set to a value of 0.95. The upper and lower bounds of the inequality constraints in equation (4.21) are fixed to $\lambda_1 = 100, \lambda_2 = 300$ for all E simulations for a total of $N = 200$ pixels, where $M \in [9, 21]$ are the range of bands that contain outliers out of a total of $D = 30$ bands, with $Q(r) = 3 \; \forall r = 1 \dots R$. The total possible number of anomalies for each *eth* simulation are fixed to $P = 10$, on the assumption that there are no more than 5 percent of pixels in the window that are anomalies. We evaluate P_D and FAR for $P = 10$ anomalies and evaluate $L = 10$ critical bands to check whether specific outlier measurements were artificially placed in these bands. Please refer to table 4.2 for a summary of the experiment parameters and Fig. 4.2 for visualisation of input data. The number of critical bands identified are compared with the actual band-subsets that contain outliers. The experiment is repeated for $E = 1000$ simulations. Initial values for each model component is random



(a) (b)

Figure 4.2. Experiment A: a) Subset of backgrounds (blue) and all true anomalies (red) detected for the least difficult scenario considered where there are 7/10 band-subsets that contain outliers. b)Arbitrary backgrounds (blue) and all true anomalies (red) detected for the tougher case where there are 4/10 band-subsets that contain outliers

for the mean and co-variance but fixed for each Gaussian component prior throughout all simulations. The best probability of detection and lowest FAR achieved out of ten different initialisation conditions were chosen for each *eth* simulation. Results are also expressed as a function of the cumulative sum of Cauchy-Schwarz distance [48] between outlier and partial background distributions across $E = 1000$ simulations. We show the P_D vs. FAR for all M cases. Error bars indicate the maximum and minimum probability of detection at these thresholds across E simulations. Refer to Fig. 4.3 for results of Experiment A.

4.4.2 Experiment B: Real Hyperspectral Data

In this section we test the efficacy of the proposed algorithm with real Hyperspectral data. The dataset was collected by the Rochester Institute of Technology (RIT) using a HyMAP sensor with 126 bands collected over Cooke City, Montana, USA in 2006 [49], [16]. The airborne sensor was flown at height of 1.4 km from the ground and yields a ground spatial resolution of 3m per pixel. The dataset is publicly available including ground-truth information and can be accessed at (http://dirsapps.cis.rit.edu/blindtest/).

Table 4.1. Experiment B: Anomaly Details

Anomaly Code	Anomaly Type	No. of Anomalies	Pixel Resolution
F1	Red Cotton Cloth	9	3m x 3m
F2	Yellow Nylon	9	3m x 3m
F3a, F3b	Blue Cotton	9,9	1m x 1m, 2m x 2m
F4a, F4b	Red Nylon	9,9	1m x 1m, 2m x 2m

True anomaly details can be found in table 4.1. We examine the critical bands inferred whilst detecting anomalies, F1, F2, F3, F4, from a total of 10000 pixels. This subset of pixels in the RIT dataset was chosen such that it is inclusive of varying backgrounds, forest, grass and soil as well as known anomalies. The problem is to examine the bands inferred in the detection of these anomalies at varying P_D and FARs. We fix the window-size to $N = 289$ pixels and apply the window sequentially until we reach 10000 pixels, Fig. 4.4 illustrates the difficulty of the problem. We set the number of bands in each *rth* band-subset, to $Q(r) = 10$ for all R band-subsets, taking into account the maximum spectral overlap of the HyMAP sensor [50] i.e. 10 bands is also sufficient for the convergence of the parameters given the number of samples considered in the window. For the experiment conducted, $\tau = 0.5$ in (4.2) across known atmospheric bands enabled convergence of likelihood across these bands, where setting it to a higher value meant that global convergence of outlier and background parameters was never reached and resulted in an infeasible solution to the optimisation problem in (4.20). The atmospheric band-subsets correspond to $R = 6$ which contains bands $61 - 70$, $\tau = 0.95$ across the remaining band subsets accounting for some band noise which may affect estimation performance. Prior knowledge regarding the spectral regions of contrast provides a means of reducing FAR and is incorporated into the optimisation problem in (4.20). The first summation in the in-equality constraint is restricted to a particular band range. We add two unique constraints, for band ranges $10 - 30$, $70 - 90$, which are maintained throughout the entire experiment. These regions were empirically determined from visual inspection as regions of *potential contrast*, refer to Fig. 4.4. Two penalties are effectively applied due to the constraints, hence if the indicator $T_{d,r,n}$ is equal to 1 contiguously across some or

Table 4.2. Experiments A,B: Parameter Summary (* Parameter setting for Atmospheric Bands)

Exp.	$Q(r)$	τ	R	α_0; P=10,10	β'; L=10,20	λ_1, λ_2	Prior Bands
A	3	0.95	10	α_{11}	$\beta'_1 \dots \beta'_{10}$	100, 300	-
B	10	0.95 / *0.5	12	α_{11}	$\beta'_1 \dots \beta'_{20}$	400, 1500	10-30, 70-90

Table 4.3. Critical Bands, True Anomaly Detection Summary

Anom. Code	Inferred Critical Bands	P_D (FAR < 0.012)
F1	22-44	1.0
F2	20-30, 61-62	0.778
F3	70-80, 64, 66-69	1.0
F4	6-33, 110-115	0.278

all of these regions, the possibility of the pixel containing a anomaly is more likely. The inference of a band rank provides the exact wavelengths that are critical to detection of these anomalies and are quantified according to certain FARs and P_D. Hence, the upper and lower bounds in (4.20) are adjusted from the previous experiment to $\lambda_1 = 400, \lambda_2 = 1500$ due to the increase in window size and size of $Q(r)$, through a manual process across a window consisting of a known set of anomalies. If the resultant inference of bands across a certain window corresponds to a low rank of either bands $10-30$ or $70-90$, the presence of a anomaly in the window is unlikely. On the corollary, results show that if the bands inferred do not fall within the region of potential contrast, the FAR is likely to be high. For experiment B we set $P = 10, L = 20$, where the aim was to maximise P_D assuming that there are no more than 10 anomalies per window. For evaluating anomalies the threshold values correspond to summing the rows, $\alpha_n = \sum_m \phi^*_{m,n} \in \mathbb{R} \forall n = 1 \dots N$ evaluated across the pre-specified band range, M. Subsequently, the threshold values used to evaluate the presence of critical bands correspond to summing the columns across N pixels, $\beta_m = \sum_n \phi^*_{m,n} \in \mathbb{R} \forall m = 1 \dots M$. The ROC curves are shown in Fig. 4.5. Inferred critical bands are provided in table 4.3. Bands are shown in the order in which they were ranked.

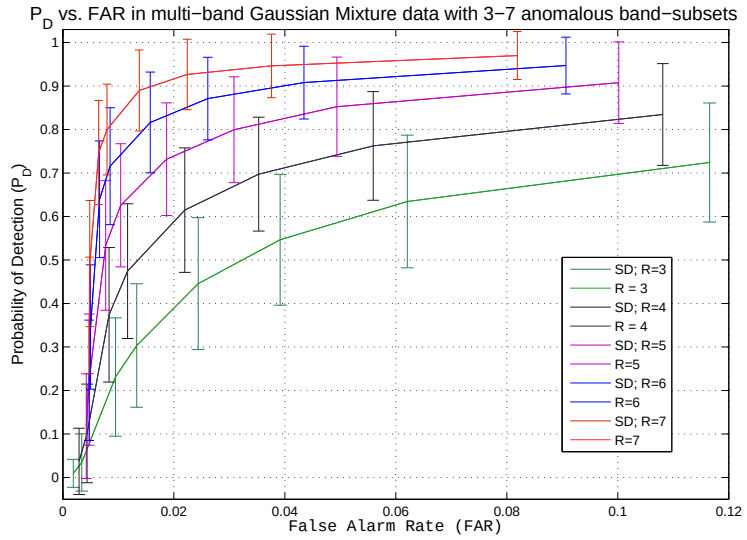
4.5 Discussion

It is important to note that all false alarms occur locally since we have used a window-based technique which maybe more acceptable from a surveillance perspective. Furthermore, a rank and window-based technique to determine anomalies also implicitly provides a method to fix the false-alarm rate which is advantageous. Receiver Operator Characteristic (ROC) curves from the RIT test image indicate the difficulty (as can be expected) in finding sub-pixel anomalies at low false-alarm rates. The critical band ranks obtained especially for F1, F2, F3 in experiment B confirm the utility of contiguously placed bands in hyperspectral sensor design since they reveal these anomalies. Results obtained from the detection of F4 suggest that bands that fall outside the pre-selected list of bands are likely to contribute to a greater number of false alarms. This not only motivates the notion of scene-dependant hyperspectral measurement across critical bands but also highlights the value of measuring prior band utility (specified by the variable M) for anomaly detection. The process of constraint addition accommodates for multiple types of anomalies as indicated by the different bands used to detect F4, F3. Prior knowledge of specifying useful bands is achieved through visual inspection. Although this process does not have any notion of optimality, results indicate that it was reasonably accurate and perhaps a suitable visual metric can be developed in subsequent work. Furthermore, the process of adding or selecting the appropriate bands and constraints does not need to be repeated once it has been conducted for the combination of scene background and anomaly. Similarly, the values for λ_1 and λ_2 for the experiments considered were determined for the simulated data through a process of trial and error but was fixed throughout the experiment. For the real dataset, the thresholds were increased appropriately to a value based on window-size and size of $Q(r) \forall r = 1 \ldots R$. Finding optimal values for λ_1, λ_2 is not necessarily a search problem since we would like to use the least number of bands to identify the maximum number of anomalies. Nonetheless, we note that future work shall include empirical testing to determine the effects of varying λ_1, λ_2, as well as exploring an adaptive procedure to determine their exact values that incorporates belief in the number of anomalies in the scene as well as minimum number of critical bands required for doing so. Results

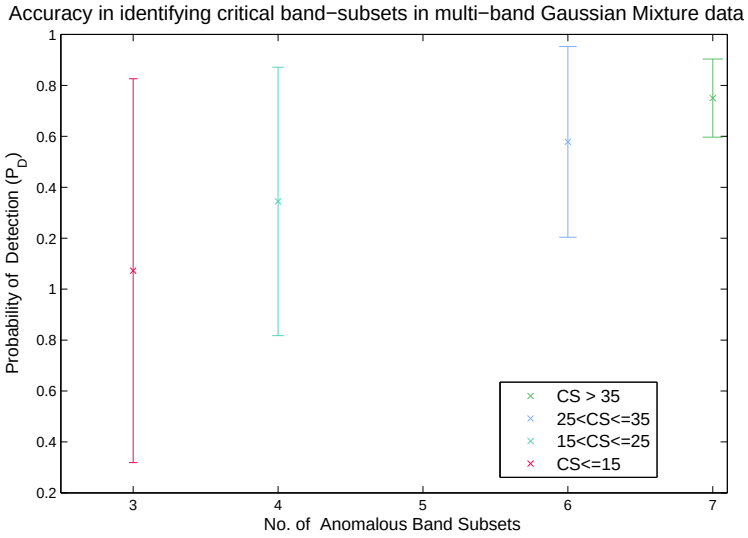
from both experiments show that anomaly detection and band subset selection performance are not a direct function of D but rather a subset M of appropriate bands. If the inequality constraints in equation (4.20) are fixed, irrespective of D, performance is dependant only on the accuracy of the parameter estimates and the extent to which anomalies vary in terms of distance from the background. Results in experiment A, show that performance deteriorates as distances and number of bands with artificially placed outliers reduce. The use of fully Bayesian techniques to infer model parameters may result in greater accuracy but at the cost of analytic local convergence in likelihood offered by the standard EM approach. We also note that critical bands quoted in table 4.3 are band numbers and not wavelengths. Further details regarding exact wavelengths chosen are a subject for future study since the study is a proof-of-concept.

4.6 Conclusion

The technique proposed in the study provides an avenue for simultaneous anomaly detection and critical band-subset selection. The novelty lies in (1) the proposed generative model and anomaly inference process, (2) the iterated inference of an indicator matrix and system parameters for a Gaussian mixture model according to a KL divergence measure whilst maintaining local convergence guarantees using SAGE-EM. (3) a technique to identify critical bands in surveillance problems using convex optimisation as well as detecting low SNR anomalies in hyperspectral data. Anomaly detection performance is demonstrated through acceptable P_D and FARs for multi-dimensional measurements that vary significantly across a subset of co-ordinates or critical bands. Band-subset selection accuracy is also measured as a function of distance between anomaly and background and shows that the accuracy deteriorates as distance reduces (as is the case with sub-pixel anomalies) but produces acceptable results when the anomalies vary across a sufficient number of bands. The results motivate the idea of picking and choosing which bands to measure from, based on a combination of prior information and a performance metric using the test data available, which is pursued in the next chapter.

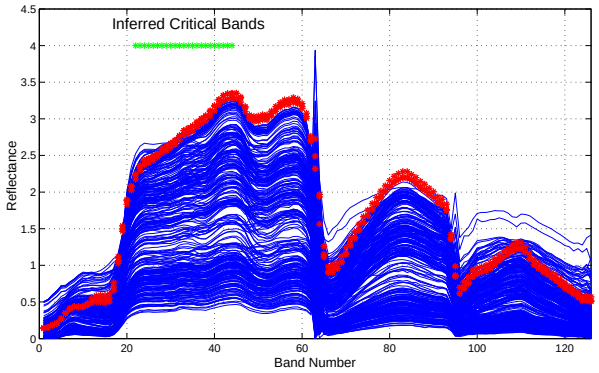


(a)

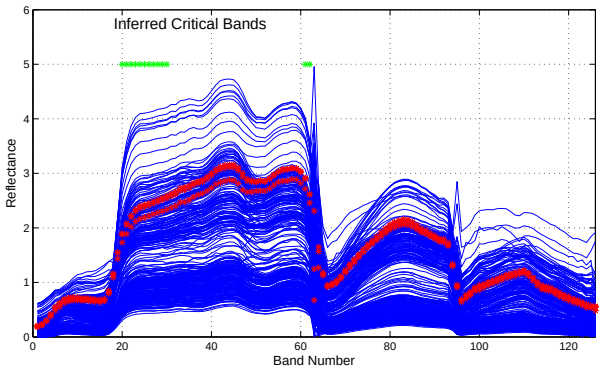


(b)

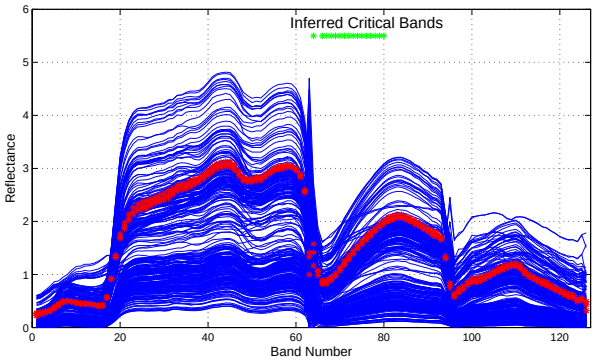
Figure 4.3. Experiment A: a) P_D vs. FAR for $M = 7,6,5,4,3$ anomalous bands-subsets out of $R = 10$ total band-subsets. They are represented by red, blue, purple, black and green curves respectively. Each band-subset consists of $Q(r) = 3$ bands hence making the total number of bands equal to 30. Error bars (referred to as SD in the legend) indicate the accuracy range for $E = 1000$ simulations. b) Band Ranking performance is measured according to how many times critical band-subsets that contain unique outliers were actually chosen. It is represented as a function of the cumulative Cauchy-Schwarz distances between partial background and outlier distributions and as a function of the anomalous band subsets is indicated in the legend.



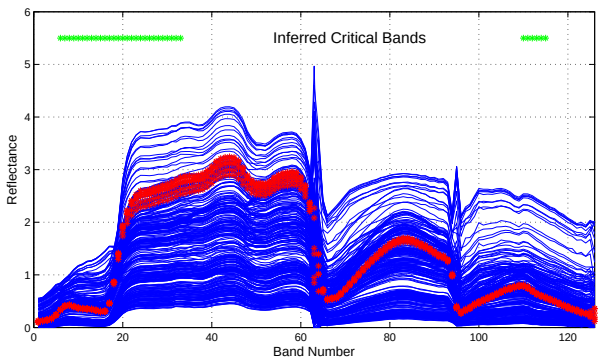
(a)F1 (red) vs. Background



(b)F2 (red) vs. Background



(c)F3 (red) vs. Background



(d)F4 (red) vs. Background

Figure 4.4. Spectral measurements of anomaly vs background in local window consisting grass, tree and soil. Green asterix indicates critical bands inferred for each material. Note how the locations vary. For anomaly, $F4$, which produced the worst result in terms of P_D, 34 bands are required to obtain the result. Critical bands identified for $F1$ can be validated visually, whereas the inferred critical bands are not so obvious for the others.

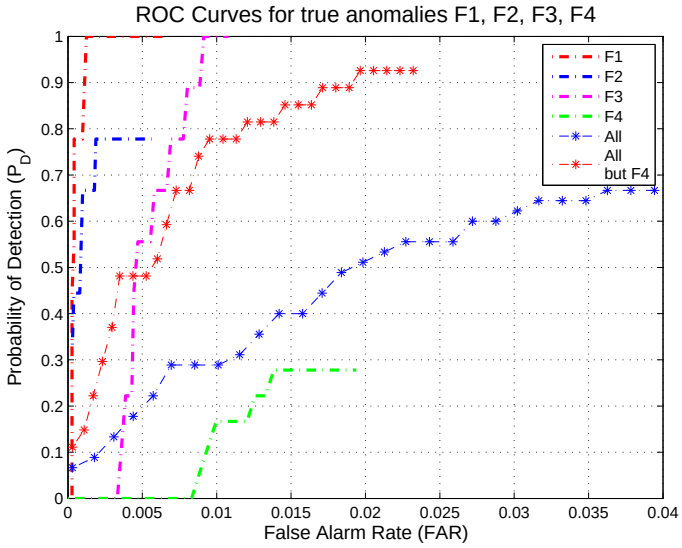


Figure 4.5. Experiment B: P_d vs FAR for finding 54 anomalies out of 10000 pixels: F1,F2,F3,F4.

Chapter 5

Band Sparsity for Compositional Models in Hyperspectral Imaging

THIS chapter deals with the problem of online band selection in Hyperspectral Imaging. The advent of Adaptive Focal Plane Arrays (AFPAs) [9] enable a sensor array to be electronically tuned to measure signals across a varying subset of sensor bands for each pixel gathered. AFPA payloads are also light-weight which make them suitable for spaceborne imaging where weight restrictions become a consideration. Thus AFPAs, promise a reduction in power, processing and bandwidth requirements making it suitable for applications such as persistent surveillance. To the authors' knowledge however, a suitable technique to determine the subset of bands to cue at each pixel is not yet available. The proposed algorithm alleviates the need for eigen-decomposition based dimensionality reduction schemes that operate in batch mode and rely on the entire scene being collected before dimensionality reduction takes place. Given a library of possible scene signatures, we first design a sub-pixel model and then subsequently measure the influence of each band on abundance estimation performance. Subsequently, we design a recursive algorithm that combines prior band utility and bands used in previous pixels to estimate the most likely band subset to be cued for each subsequent pixel. All experiments are conducted on synthetic and real AVIRIS-Cuprite data used by Mittelman *et. al.* in [2]. Both abundance and endmember estimation accuracy are better in-term of (Sum- Squared Error) SSE than state-of-the-art techniques operating in batch-mode.

5.1 Introduction

In hyperspectral imaging, the signal used to represent each pixel in the scene is a convex combination of a finite set of signals rather than a single unique signal. This is mainly due to the ground sampling distance (GSD) that can vary anywhere between one to twenty meters [3], depending on the distance of the sensor to the ground below and is referred to as spatial resolution. The linear sub-pixel model is represented by,

$$\begin{aligned} y_n &= \sum_{k=1}^{K} g_{k,n} x_{k,n} + w_n \\ \text{s.t. } x_{k,n} &\in [0,1],\ c = \sum_{k} x_{k,n} = 1 \end{aligned} \tag{5.1}$$

where, $y_n \in \mathbb{R}^D$ represents the signal measured across D spectral bands which is the radiation reflected by the nth spatial location or pixel on ground for $n = 1 \dots N$ total number of pixels, $g_{k,n} \in \mathbb{R}^D \forall k = 1 \dots K$ represents K possible constituents or *endmembers*, $x_{k,n} \in [0,1]^K$ is the abundance or fractional contribution of each constituent signal and w_n represents the additive sensor noise. This model is an instance of a *compositional model* where each observation under the model does not fall under one unique class but is a member of K classes. If $x_{k,n}, g_{k,n} \forall k = 1 \dots K, n = 1 \dots N$ are treated as samples of random variables, X_k, G_k described in Chapter 2. This means each pixel falls under a joint distribution of endmember and abundance, where N samples or pixels are independent and identically distributed. A common problem in hyperspectral imaging is to solve for unknowns $g_{k,n}, x_{k,n} \forall k, n$ and is referred to as spectral unmixing and is not restricted to probabilistic treatment of endmember and abundance. In fact, the spectral unmixing problem can be thought of as resolving a linear system of equations, where we alternate between estimating $g_{k,n}$ and $x_{k,n} \forall k, n$. If we fix the values of $g_{k,n} \forall K$ and resolve for $x_{k,n}$, since $D > 100$ and $K << D$, the system is overdetermined. The total number of linearly independent equations is restricted to less than $K + 1$. The number of rows corresponding to D equations is greater than K unknown variables that correspond to the abundances in each row. If the sum-to-one constraint is ignored briefly, this implies that there could either be single, infinite or no solution depending on the exact number of linearly independent equations which is unknown. Nonetheless, the natural sum-to-one abundance constraint prevalent in

sub-pixel HSI data implies the solution for $x_{k,n}$ lies in a $K-1$ dimensional convex subspace, since $D >> K$. This has subsequently motivated a plethora of techniques [51] that seek a mapping of $G_k : \mathbb{R}^D \rightarrow \mathbb{R}^{K-1} \forall k = 1 \dots K$ endmembers, where the transformed $K-1$ dimensional basis vectors are independent to one-another. Typically, the methods are eigen-decomposition methods where the basis vectors correspond to K principal eigenvectors. Although the mapping enables the practitioner to deal with a convex subspace and reduces the solution space which improves computational complexity, it removes natural chemistry variations in the data.

Hyperspectral signals are spectrally correlated due to overlap in neighbouring spectral band responses and also due to chemistry variations across across the infra-red spectrum, which also differ for each material. Given these conditions, the signal measured at each pixel exhibit non-Markovian behaviour, where a band measurement is influenced not only by the value across the previous band but also by whether its a member of a larger chemical absorption or reflectance feature that encompasses many bands. The modification of the measurement axes, as is the case eigen-decomposition methods is intended to remove this correlation structure, which can lead to poor unmixing performance especially for low Signal-to-noise-Ratio (SNR) materials in surveillance applications [52]. Furthermore, eigen-decomposition methods also imply batch mode processing which imply that all pixels have already been gathered by the sensor.

We do not attempt such a mapping in this study, we wish to preserve the natural physical structure and cue $Q_n \subset D$ bands to collect measurements across each pixel, where Q_n varies $\forall n = 1 \dots N$ pixels. We hypothesise that not all D spectral measurements are required for an accurate estimate of the abundance in each pixel. We also assume that the remaining $N-n$ pixels are yet to be collected. We seek a reasonable method to select Q_n conditional on some prior knowledge of the utility of each band as well using recursive estimates based on bands used in previous pixels. In this study, we seek a technique to find $Q_n \forall n = 1 \dots N$ pixels and carry out online band selection under a novel sub-pixel model where no eigen-decomposition is applied. We assume that

$g_{k,n}, x_{k,n}$ are samples of random variables, G_k, X_k, that represent measurements which lie in $G_k \in \mathbb{R}^D, X \in [0,1]^{K-1}$.

5.1.1 Motivation and Significance

Our primary motivation for online band selection is for Hyperspectral sensors with Adaptive Focal Plane Arrays (AFPA). AFPA's [9] are electronically tunable focal plane that offer an alternative mode of collecting signals than a conventional pushbroom scanning in a hyperspectral sensor. Rather than collect signals in a mechanical manner across all bands, an AFPA can be electronically tuned to gather signals across only certain bands and can be adaptively tuned for each pixel gathered. An AFPA facilitates for cueing of hyperspectral bands which reduces the amount of information gathered, reducing computational storage, network throughput and processing requirements, thus minimising system costs. Moreover, AFPA systems are also compact and lightweight compared their mechanical counterparts which make them more relevant for spaceborne systems. However, a systematic automated procedure to select appropriate bands for each pixel observed is not widely documented. It is unknown on what basis the bands are selected and whether prior scene knowledge is required to adaptively tune bands. In surveillance applications, it is realistic to assume that some prior knowledge of possible materials in the scene is available via spectral libraries, satellite data and/or previous collections over the region. This can have serious implications for sensor design, where measurement collection can be cued across a small subset of bands out of a large number of available bands in the array based on prior knowledge and those used across previous pixels. To our knowledge the combination of prior knowledge and unsupervised recursive band selection are not used in AFPAs. Moreover, evaluating the utility of a band prior to collection also enables potential improvement in SNRs for that bands if it is deemed useful for a particular scene type.

5.1.2 Summary of Work and Contributions

Our overall contribution is the novel sub-pixel model, Gibbs Sampling algorithm for inferring the abundance and procedure for carrying out recursive band selection in sub-pixel hyperspectral data. To our knowledge, this is the first instance of such an algorithm in hyperspectral literature. We briefly provide an overview of the proposed algorithm and identify novel contributions:

1. A Gaussian Process is used to spectrally approximate prior end-member means and covariance from a spectral library, where each endmember is drawn from a Gaussian probability distribution. Posterior update of endmember means and covariance follow the Kalman Filter update equations after each Gibbs Sampling iteration. This approach to evaluating endmembers is novel, in terms of sequential update of means and covariances whilst preserving the physical structure of the endmembers. Moreover, even when original band information is preserved as is the case for sparse regression techniques [51], within class-variance information of each endmember class is not used unlike our proposed method.

2. Unlike conventional approaches that use truncated Gaussian distributions to represent the abundance vector, we model the abundance parameter using a Gamma distribution. The posterior abundance estimates are also Gamma distributed conditioned from a joint Gamma Gaussian model. We apply two unique techniues to enforce sum-to-one constraints since the Gamma distribution does not provide sum-to-one constraints on the abundance. In the first method, we use the Gamma-Dirichelet relation to enable a simple update of the abundance hyperparameters which improves the sparsity of future abundance candidates. We select the abundance candidate that provides the maximum likelihood estimate of pixel parameters from an arbitrary number of Gibbs Sampling iterations. The use of the Dirichelet-Gamma equivalency for hyperparameter update and modelling of the abundance estimate using a Gamma distribution is novel. In the second sampling technique, a nonlinear transformation of the abundance random variable is applied resulting in an implicit enforcement of the sum-to-one constraint. Both techniques are unique.

3. The current study does not facilitate the use of large spectral libraries but we apply the sparsity principle in the number of bands used to estimate the abundance. A Beta process recursively estimates the number of bands required to estimate the abundance in each pixel for all pixels in the test scene using some prior knowledge. Whilst we adopt the Gibbs sampler used by Thibaux *et. al.* in [29] to construct and sample from the posterior Beta process, this study is the first instance of such a process being applied to hyperspectral imaging. The base distribution of the Beta process encodes prior knowledge by measuring the number of times each band is used to estimate abundance from a training dataset which contains similar materials. The formulation and application of the stochastic process to band selection using sub-pixel mixing criteria is novel.

4. A set of sub-optimal band weights are derived from a convex relaxation procedure that is formulated to estimate band utility for the base distribution. The formulation takes into account band correlation which is not pre-specified or constrained in any way. The formulation and application of the convex optimisation problem is also novel.

In section 5.2 we provide an overview of existing unmixing methods that carry out Bayesian unmixing and highlight where our technique differs. For purposes of clarity for the reader we present the problem in two parts a) sub-pixel mixing b) band selection. The sub-pixel mixing component is presented in section 5.3, where we introduce our proposed sub-pixel Bayesian model consisting of Gaussian Processes to capture endmember means and covariance and Gamma distributions to capture abundances. Subsequently, we describe the Gibbs Sampling procedure required to infer unknown parameters and hyperparameters and derive the posterior probabilities for endmember and abundance estimates. In section 5.6.1, we introduce a convex relaxation approach with training data to estimate the base measure of a stochastic Beta Process to recursively cue bands for each pixel. In section 5.6 we update our existing Bayesian model for the test dataset we introduce the Beta process that determines the number and the exact bands to be used to describe each pixel. In section 5.6.3, we provide a summary of the proposed Recursive Sparse Band Selection (RSBS) algorithm.

We conduct simulations with both synthetic and real-data in section 5.7, where the former is used to benchmark abundance estimation accuracy with state-of-the-art unmixing algorithms for comparative purposes and the latter tests show the endmember reconstruction accuracy with real data. In section 5.8 we discuss design aspects that are significant to performance. We conclude in section 5.9 with some ideas for future work.

5.2 Existing Work

In terms of band selection we believe our work is the first of its kind that carries out an online recursive band selection approach using a sub-pixel model, therefore scope for comparison does not exist. However, the sub-pixel mixing problem has received considerable treatment. We provide an overview of existing sub-pixel mixing methods and then illustrate how the design of our model differs to accommodate for the online recursive band selection aspect.

Existing approaches can be divided into two categories; (1) those methods that rely explicitly on training data - to model endmembers $G_k \forall k = 1 \dots K$ this includes pure signals collected and stored in spectral libraries. In this category the nth abundance, $P(X_k = x_{k,n})$ is inferred given observation y_n and estimates $\hat{G}_k = g_{k,n} \forall K$; and (2) those that rely on the scene to estimate both $g_{k,n}, x_{k,n}$. The latter category includes geometric, and statistical techniques that rely on the entire scene being gathered before represent endmebers. Regression and sparse coding techniques fall under the former category, where the reliance is on spectral libraries, which facilitate for online unmixing. Bioucas-Dias *et. al.* in [51] provide an elaborate account of methods which fall under both categories. However, both categories have their disadvantages. Methods that rely on training data suffer from being specific in terms of being localised to certain sensor instrumentation and also suffer from an inability to handle spectral variability such as atmospheric noise which is often scene specific [51]. Geometric and statistical techniques such as Zare et. al in [53] and Nascimento *et. al.* in [54] pose a heavy computational storage burden and does not facilitate for online processing. The proposed

technique in this study falls under the former category.

Techniques that fall under the former category are referred to as sparse regression or sparse coding methods. These techniques are applied to cases where spectral libraries are quite large i.e. > 100. The term sparse refers to the fact that, there is typically no more than 10 materials that are present in a single pixel thus zeroing abundance estimates of the remaining endmembers. Sparse regression methods use raw spectral training data, where differences in sensor instrumentation, atmospheric conditions and natural spectral variability of test conditions are not accounted for. These shortcomings are addressed by sparse coding methods, where posterior estimates of possible spectra in each pixel is learnt from a spectral library or referred to as a dictionary. We also apply a sparsity principle in our study but assume that there is a small number of bands are sufficient to describe the abundance of each pixel. We estimate the posterior of the dictionary given the pixels but assume the number of endmembers in the dictionary is small and hence does not require sparsity to be imposed either on the abundance [55] or the dictionary [56], [57] itself.

Given some prior knowledge of endmembers in the scene, a fully Bayesian model aims to capture the spectral variation between training data and the scene by computing the posterior of the endmembers given the pixel measurement and abundance. Eches *et. al.* in [58] represent the endmember as a draw from a Gaussian distribution, with known means derived from the spectral library with diagonal covariance. The variance in each band is also a random variable and modelled as an inverse Gamma distribution which provides convenient conjugacy properties during sampling. However, in the study they do not compute the posterior probability of the endmember given the measurement and therefore it remains fixed throughout the Gibbs Sampling procedure. We treat the endmember as a draw from a Gaussian but where each endmember mean is drawn from a Gaussian Process and also re-sampled during each Gibbs Sampling iteration using a Kalman Filter update unlike [58]. Dobigeon *et. al.* in [59] and Mittelman *et. al.* in [2] apply similar priors as Eches *et. al.* in [58] but use the scene to estimate endmember means using endmember extraction procedures like Vertex Component

Analysis [54] which is not relevant for online band selection. Both studies also use PCA to reduce endmember dimensionality, which does not provide an indication of useful bands that are required for AFPAs and is hence not pursued in this study.

The abundance posterior in [58] [2] and [59] is a draw from a truncated multivariate Gaussian where the sum-to-one constraints are enforced through a partial estimate of the abundance. In this study, we propose two separate techniques to enforce sum-to-one constraints, where in the first we use the Dirichelet-Gamma equivalency to obtain posterior abundance estimates that maintain the sum-to-one constraints whilst in the second we rely on a nonlinear transformation of the abundance random variable to implicity enforce a sum-to-one constraint. Other approaches to estimate the abundance parameter are optimisation approaches, where the parameter is not a random variable such as the adaptive Lasso in [55]. Mittelman *et. al.* in [2] propose a hierarchical approach and spatial smoothing to estimate sparse abundances. In the proposed method, abundances are sparsified by propagating group membership probabilities from the top of a quadrature tree, where the leaves of the tree represent each pixel. Posterior estimates of the abundances include the multi-layer group membership probabilities which zero small abundance values. The degree of sparsity achieved is reflected in abundance estimation accuracy evident in table 5.2 in the Results section.

5.3 Problem Formulation

We analyse the sub-pixel mixing aspect of the proposed model before analysing the band selection component. The proposed generative model representing the sub-pixel mixing problem is listed in figure 5.1. In the proposed model, the prior distributions

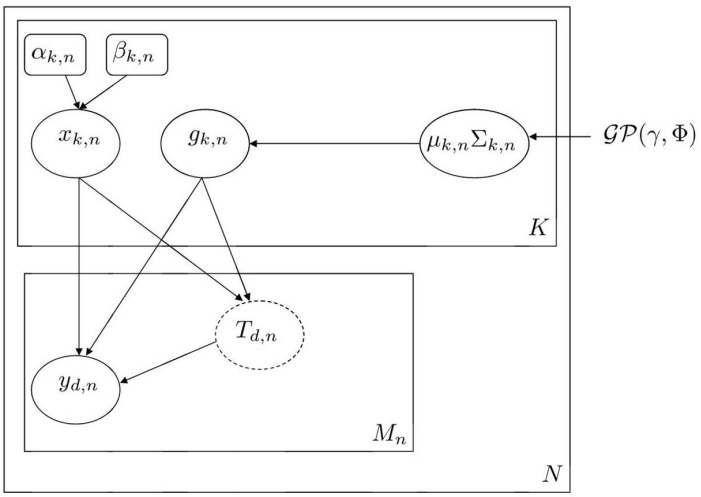


Figure 5.1. This directed graphical model represents the generative model used to capture linear sub-pixel mixing phenomena described in equation (5.1). Random variables (circles) and hyperparameters (smooth boxes) are unknown and inferred using a Gibbs Sampler. The lower-case symbols used inside the circles represents samples of those random variables. Arrows indicate the dependencies between random variables. The exception to this rule is the pixel y_n which is an observation. In this model the kth abundance of the nth pixel is represented by $x_{k,n}$ and $g_{k,n}$ is the kth endmember that is present in the nth pixel and sampled from posterior probabilities of random variables X_k, G_k. The endmember and abundance are conditionally dependant given the measurement at the nth pixel y_n. Hyperparameters $\alpha_{k,n}$ (shape) and $\beta_{k,n}$ (scale) vary for each nth pixel. In this model, the measurement at each nth pixel is assumed to be independent of remaining $N-1$ pixels. Indicator matrix $T_{d,n}$ is not a random variable and is iteratively inferred to determine whether M_n bands are sufficient to describe the pixel. y_n. The band selection aspect of the model is specific to training data and is used to estimate the base distribution of the Beta process B_0.

are described by the following,

$$
\begin{aligned}
P(G_k = g_{k,n} | \mu_{k,n}, \Sigma_{k,n}) &\sim \mathcal{N}(\mu_{k,n}, \Sigma_{k,n}), \\
P(X_k = x_{k,n} | \alpha_{k,n}, \beta_{k,n}) &\sim Ga(\alpha_{k,n}, \beta_{k,n}), \\
w_n &\sim \mathcal{N}(0, R) \\
\tilde{\sigma}^2 &\sim \mathcal{IG}(v_0, \rho) \\
\alpha_0, \beta_0, \beta_{k,n}, \alpha_{k,n}, &\sim U(.)\ \forall k = 1 \ldots K
\end{aligned}
\tag{5.2}
$$

where, random variable $P_k = cX_k \in [0,1]$ is a scalar transformation whose samples are denoted by $p_{k,1} \dots p_{k,n}$ and $c = \frac{1}{\sum_k x_{k,n}}$, $w_n \in \mathbb{R}^{Q_n}$ is additive zero-mean independent Gaussian noise with a covariance $R \in \mathbb{R}^{Q_n \times Q_n}$ has an inverse-Gamma prior, $\mu_k \in \mathbb{R}^{Q_n}, \Sigma_k \in \mathbb{R}^{Q_n \times Q_n} \forall K$, are hyperparameters that represent endmember means and covariances obtained from a Gaussian Process that takes into account the correlation across bands, (this is further elaborated in subsection 5.4.1), the abundance estimate X_k is described by a Gamma distribution represented by $Ga(.)$ and parameterised by the shape $\alpha_{k,n}$, and scale $\beta_{k,n} \in \mathbb{R}$ parameters. Refer to subsection 5.4.2 for further explanation and justification for using the Gamma distribution. Since $P(cX_k \leq p_{k,n}) \equiv P(X_k \leq \frac{p_{k,n}}{c})$, any sample $x_{k,n} = \frac{p_{k,n}}{c}$. The likelihood or joint probability of the proposed generative model is given by,

$$
\begin{aligned}
&P(y_n, X_k = x_{k,n}, G_k = g_{k,n}, W = w_n, \beta_{k,n}, \alpha_{k,n}, \mu_{k,n}, \Sigma_{k,n} \forall k) = \\
&\prod_{k=1}^{K} P(y_n | \hat{X}_k = x_{k,n} \hat{G}_k = g_{k,n} \forall k) P(G_k = g_{k,n} | y_n, \hat{G}_{k'} = g_{k',n} \forall k', \hat{X}_k = x_{k,n}, \hat{\beta}_{k,n}, \hat{\alpha}_{k,n} \forall k) \dots \\
&P(X_k = x_{k,n} | y_n, \hat{G}_k = g_{k,n} \forall k, \hat{X}_{k'} = x_{k',n} \forall k', \hat{\alpha}_{k,n}, \hat{\beta}_{k,n} \forall k) \dots \\
&P(\alpha_{k,n}) P(\beta_{k,n}) P(G_k = g_{k,n} | \hat{\mu}_{k,n}, \hat{\Sigma}_{k,n}) P(X_k = x_{k,n} | \hat{\alpha}_{k,n}, \hat{\beta}_{k,n})
\end{aligned} \tag{5.3}
$$

where, $P(G_k = g_{k,n} | y_n, \dots)$ and $P(X_k = x_{k,n} | x_{k,n} \dots)$ refer to the posterior probability distributions of the endmember and abundance, where, parameters denoted with a $\hat{}$ refer to parameter estimates stored and inferred $\forall k = 1 \dots K, n = 1 \dots N$ samples, $\hat{G}_{k'} = g_{k',n}, \hat{X}_{k'} = x_{k',n}, \beta_{k',n}$ refer to inferred values of all $K-1$ instances apart from the kth. The terms $P(G_k = g_{k,n} | \hat{\mu}_{k,n}, \hat{\Sigma}_{k,n})$ is initialised using training data and a Gaussian process, $P(X_k = x_{k,n} | \hat{\alpha}_{k,n}, \hat{\beta}_{k,n})$ refers to prior of X_k drawn from a Dirichelet with hyperparameters initialised to 1 across all K samples. We derive both posterior probabilities before stating the Gibbs Sampling steps, including two separate sampling techniques for the posterior abundance.

5.4 Background

5.4.1 Representing End-members using Gaussian Processes

Given that the endmember is described by, $P(G_k = g_{k,n}) \sim \mathcal{N}(\mu_{k,n}, \Sigma_{k,n})$, Gaussian processes (GPs) provide a means to estimate the mean and covariance capturing neighbourhood band correlation and also non-Markovian correlation which are both reflected in the smoothness of hyperspectral signatures. Furthermore, GPs allow for differences in sensor types between the spectral library, Φ as well as from the test sensor in the aircraft. If $\tilde{D}$ band wavelengths in a spectral library are denoted by vector, $\mathbf{z} = z_1 \dots z_{\tilde{D}}$ and Q_n test wavelengths by $\mathbf{z}^* = z_1^* \dots z_{Q_n}^*$, the Gaussian mean and covariance of the *kth* endmember are given by,

$$\mu_{k,n} = \gamma_k(z^*, z)[\gamma_k(z, z) + \sigma_n^2 I]^{-1} \tilde{g}_k \tag{5.4}$$

$$\Sigma_{k,n} = \gamma_k(z^*, z^*) - \gamma_k^T(z^*, z)[\gamma_k(z, z) + \sigma_n^2 I]^{-1} \gamma_k^T(z, z^*),\ \forall z, z^* \forall k = 1 \dots K \tag{5.5}$$

where, γ is a covariance function and represents the inner product between a pair of input values, z, z^*, $\sigma_n^2 I$ is the diagonal noise for each endmember in the spectral library denoted $\tilde{g}_k \in \mathbb{R}^D \forall k = 1 \dots K$. The derivations for the endmember mean and covariance function γ is provided by Rasmussen *et. al.* in pp. 19 of [60]. The choice of the covariance function type or kernel, γ, plays a strong role in abundance estimation accuracy. We consider the use of the Matern covariance kernel that has traditionally been used to capture behaviour of geophysical processes [60]. The hyperparameters of this kernel are used to determine the sensitivity of the function to a spectrally varying process as well as accounting for spectral variability within the material class between training and test data. We discuss the choice of the covariance kernel in sec.5.7.

5.4.2 Using Gamma and Dirichelet Distribution to represent Abundance

Given that the endmember is a draw from the Normal distribution and the abundance has a Gamma prior, the joint probability is a Normal-Gamma distribution. The justification for using a Gamma prior is the analaytic tractability it offers with the Gaussian, it

also allows for an easier hyperparameter update given its relationship to the Dirichelet distribution. Subsequently the hyperparameter update is used to narrow the search space of the abundance parameter estimates. Since the Dirichelet distribution implicitly enforces sum-to-one constraints. Its usage is advantageous. We briefly describe how this can be couched into the Gaussian Gamma framework and highlight the effects of updating the hyperparameters, $\alpha_{k,n}, \beta_{k,n}$.

Since $p_{k,n} = \frac{x_{k,n}}{\sum_k x_{k,n}}$, where $P(X_k = x_{k,n}) \sim Ga(\alpha_{k,n}, \beta_{k,n})$ then for K samples, $p_{1,n} \dots p_{K,n} \sim Dir(\frac{\alpha_{1,n}}{\beta_{1,n}}, \dots, \frac{\alpha_{K,n}}{\beta_{K,n}})$ is represented by a Dirichelet distribution, where $\sum_k p_{k,n} = 1$. The parameters $\alpha_{k,n}, \beta_{k,n} \forall k = 1 \dots K$ is the normalised set of hyperparameters, that correspond to the shape and scale parameters respectively. The mean of X_k at the nth pixel is given by $\mathrm{E}\{X_k\} = \frac{\alpha_{k,n}}{\beta_{k,n}}$ and variance by, $\mathrm{E}\{X_k^2\} = \frac{\alpha_{k,n}}{\beta_{k,n}^2}$. Therefore the smaller the value of $\alpha_{k,n}$ the smaller the scalar mean and variance and more certain the abundance estimate. If $\alpha_0 = \sum_k \frac{\alpha_{k,n}}{\beta_{k,n}} \forall k = 1 \dots K$, the Dirichelet abundance mean corresponds to, $\mathrm{E}\{P_k\} = \frac{\frac{\alpha_{k,n}}{\beta_{k,n}}}{\alpha_0}$ and variance by $\mathrm{E}\{P_k^2\} = \frac{\alpha_{k,n}(\alpha_0 - \frac{\alpha_{k,n}}{\beta_{k,n}})}{\alpha_0^2(\alpha_0+1)}$, where a reduction in $\alpha_{k,n}$ also translates to a smaller mean but larger variance increasing the possible number of sample values for that random variable. Although samples drawn from a Dirichelet can be described as those from a normalised Gamma, a reduction in hyperparameter $\alpha_{k,n}$ results in different outputs. We adjust the hyperparameters to reduce the search space for abundance parameter. In this study we also adjust the hyperparameters and re-sample from the Dirichelet to penalise and reward reasonable Gamma generated abundances that fall within a certain threshold. A reduction in hyperparameters or penalty results in increased uncertainty over an abundance estimate, whereas an increase in hyperparameters or reward narrows the solution space of that random variable. Sudderth *et. al.* in [27] provide a succinct introduction to the Dirichelet distribution. The Dirichelet-Gamma relation is explored in [61] whilst Fox *et. al.* in [62] apply this condition to estimate transition probabilities whilst maintaining conjugacy relations.

5.5 Posterior Probability Estimates for a Naive Gibbs Sampler

We model unknown parameters as random variables and perform Bayesian inference using a Gibbs Sampler. In a Gibbs Sampler, posterior inference of the kth state out of K possible states of a parameter is computed given the remaining $K-1$ states of the parameter, the pixel measurement and the values of other dependant random variables in the Bayesian model. When the procedure is iterated across all possible states of a random variable and across all random variables, for a certain number of Monte-Carlo iterations the parameter estimates are known to converge to a global maximum [27], [26]. We derive the posterior probabilities of the kth endmember or abundance given prior knowledge $K-1$ estimates of that parameter as well as all K estimates of the remaining corresponding parameters.

5.5.1 Estimating the Endmember Posterior

We consider the posterior probability $P(G_k = g_{k,n}|y_n, \ldots)$ given measurement y_n, $K-1$ other endmembers of $g_{k',n} \forall k' \ldots K-1$ and K estimates of $x_{k,n}$. Let,

$$\tilde{y}_{k,n} = y_n - \sum_{k' \neq k} g_{k',n} x_{k',n} = g_{k,n} x_{k,n} + w_n \tag{5.6}$$

where $w_n \sim \mathcal{N}(0, R)$. Given that we are using a Gibbs Sampler to perform inference and we have linear Gaussian system with additive noise we can improve the estimation accuracy of the endmember mean and variance by using previous estimates of the same variable. A Kalman Filter update of the estimate of the endmember means and covariance during the lth Gibbs Sampling iteration is given by,

$$\begin{aligned}
\hat{\mu}_{k,n}^{G(l)} &= \hat{\mu}_{k,n}^{G(l-1)} + K(\tilde{y}_{k,n} - x_{k,n}\mu_{k,n}) \\
\hat{\Sigma}_{k,n}^{G(l)} &= \hat{\Sigma}_{k,n}^{G(l-1)} - K(x_{k,n}\Sigma_{k,n}) \\
K &= \hat{\Sigma}_{k,n}^{G(l-1)} x_{k,n} (x_{k,n}^2 \hat{\Sigma}_{k,n}^{G(l-1)} + R)^{-1} \\
& P(G_k = g_{k,n}|y_n, \ldots) \sim \mathcal{N}(\hat{\mu}_{k,n}^{G(l)}, \hat{\Sigma}_{k,n}^{G(l)})
\end{aligned} \tag{5.7}$$

where, $\mu_{k,n}, \Sigma_{k,n} \forall k = 1 \ldots K$ are the original endmember mean and variance obtained from the Gaussian Process. In Kalman Filter terms $x_{k,n}$ maps the true state estimate

$g_{k,n}$ into the observed space $\tilde{y}_n$. There are many studies that describe the Kalman Filter update equations, Welch *et. al.* in [63] provide a concise summary.

5.5.2 Estimating the Abundance Posterior

The true conditional posterior distribution we seek requires joint sampling of parameters from the following distribution,

$$P(P_1 = p_{1,n} \dots P_K = p_{K,n} | y_n, \hat{G}_k = g_{k,n}, \hat{X}_k = x_{k,n} \forall k = 1 \dots K) \sim Dir(\frac{\tilde{\alpha}_{1,n}}{\tilde{\beta}_{1,n}} \dots \frac{\tilde{\alpha}_{K,n}}{\tilde{\beta}_{K,n}}) \mathcal{N}(y_n - \sum_k g_{k,n} x_{k,n} | 0, R) \tag{5.8}$$

where, $p_{k,n} \forall k = 1 \dots K$ is a sparse posterior abundance drawn from the joint Dirichelet-Gaussian distribution, $\tilde{\alpha}_{k,n} \forall k = \dots K$ and $\tilde{\beta}_{k,n} \forall k = \dots K$ are hyperparameters that support sparse abundance estimates. Finding a conditional distribution of each *kth* abundance $x_{k,n} \forall k = 1 \dots K$ is intractable for a Gibbs sampler or any of its block variants since the joint distribution is in-tractable. The tractable Gamma Gaussian form alleviates this issue but at the cost of ignoring the sum-to-one constraints. In this study, we provide two techniques to overcome this issue. The Gibbs Sampler for both the proposed Unmixing techniques are provided in the subsequent two subsections 5.5.5,

5.5.3 Abundance Sampling - Technique A: Gamma Dirichelet Relation

When X_k is a Gamma random variable, the conditional posterior probability distribution, $P(X_k = x_{k,n} | y_n, \hat{G}_k = g_{k,n} \forall k, \hat{X}_{k'} = x_{k',n} \forall k', \alpha_{k',n}, \beta_{k',n} \forall k')$ is also Gamma distributed [64], where,

$$P(X_k = x_{k,n} | \hat{X}_{k'} = x_{k',n} \forall k', \hat{G}_k = g_{k,n} \forall k = 1 \dots K, y_n) \sim Ga(\tilde{\alpha}_{k,n}/2, \tilde{\beta}_{k,n}/2), \forall k, n \tag{5.9}$$

where, $\tilde{\alpha}_{k,n} = \hat{\alpha}_{k,n} + Q_n$ and $\tilde{\beta}_{k,n} = \hat{\beta}_{k,n} + (g_{k,n} - \mu_{k,n})^T R^{-1} (g_{k,n} - \mu_{k,n})$, where $\mu_{k,n} \in \mathbb{R}^{Q_n}, R \in \mathbb{R}^{Q_n \times Q_n} \forall k = 1 \dots K$ is the prior means from the Gaussian Process and the measurement noise covariance, respectively. Q_n denotes the degrees of freedom. Refer to the appendix in sec. 5.10 for an immediate proof. At this stage, $x_{k,n} \mathbb{R} \forall k = 1 \dots K$

and sum-to-one constraints do not apply.

We artificially impose a constraint on the posterior abundance estimates via an update of the posterior Gamma hyperparameters. We penalise posterior abundances that exceed a certain threshold and reward those abundance values that fill within a certain range. Experimentally it was found that, when the procedure was carried out across many iterations, the number of the abundances estimates that sum to an approximate value of 1 increased in comparison to without the update. The hyperparameter update is carried out is given by,

$$\hat{\alpha}_{k,n} = \left\{ \begin{array}{ll} \frac{\tilde{\alpha}_{k,n}}{2} - c_1 & \text{iff } x_{k',n} \geq \tau_1 \\ \frac{\tilde{\alpha}_{k,n}}{2} + c_2 & \text{iff } x_{k',n} \leq \tau_2 \\ \frac{\tilde{\alpha}_{k,n}}{2} - c_3 & \text{iff } x_{k',n} < \tau_3 \end{array} \right\} \tag{5.10}$$

The exact values c_1, c_2, c_3 used for the hyperparameter update are discussed in Step 5 of the Gibbs Sampler. In our final step, we re-sample the abundance posterior from the Dirichelet using the updated hyperparameters, which provides a sparse set of abundance estimates as well as ensuring the sum-to-one constraints are preserved.

$$P(p_{1,n}, \ldots, p_{K,n}) \sim Dir(\frac{\hat{\alpha}_{1,n}}{\beta_{1,n}} \ldots \frac{\hat{\alpha}_{K,n}}{\beta_{K,n}}), \forall k, n \tag{5.11}$$

Each element of this vector drawn from a Dirichelet distribution is Gamma distributed, where

$$\begin{aligned} P(X_k = p_{k,n} | \hat{\alpha}_{k,n}, \hat{\beta}_{k,n}) &\equiv P(\frac{x_{k,n}}{\sum_k x_{k,n}} | \hat{X}_{k'} = x_{k',n} \forall k', \hat{G}_k = g_{k,n} \forall k = 1 \ldots K, y_n) \\ &\sim Ga(\hat{\alpha}_{k,n}, \hat{\beta}_{k,n}) \end{aligned} \tag{5.12}$$

We propagate the samples values of, the linearly transformed random variable P_k, $p_{k,n} \forall k = 1 \ldots K, \forall n = 1 \ldots N$ throughout the rest of the chain to ensure that sum-to-one constraints are preserved. Recursive updates of $\alpha_{k,n}$ and $\beta_{k,n}$ results in sparse abundance estimates. The maximum likelihood posterior abundance estimate is selected from a list of estimates as the final solution, where the likelihood value refers to pixel likelihood given all parameters.

This procedure can be likened to a Metropolis-Hastings Rejection Sampling scheme, where the desired distribution is stated in equation (5.8) and the actual distribution is given by (5.11). However, further work is required to demonstrate that w are indeed sampling from (5.8). Nonetheless, Samples estimates are drawn from (5.11) and rejected until the maximum likelihood solution is achieved and subsequently preserved and propagated throughout the rest of the chain.

5.5.4 Abundance Sampling - Technique B: Non-Linear Transformation

Consider the following form,

$$\begin{aligned} y_n &\sim \sum_k G_k X_k + W \\ &\sim \frac{\sum_k G_k X_k}{\sum_k X_k} + W \\ &\sim \frac{\sum_{k' \neq k} G_{k'} X_{k'} + G_k X_k}{\sum_{k' \neq k} X_{k'} + X_k} + W \end{aligned} \tag{5.13}$$

If $S_{k'} = \sum_{k' \neq k} X_{k'}$ and $P_k = \frac{X_k}{S_{k'} + X_k}$. The joint posterior probability can be expressed as,

$$\begin{aligned} P(X_k = p_{k,n}, X_{k' \neq k} = x_{k',n} \forall k', y_n, G_k = g_{k,n} \forall k = 1 \dots K) \quad = \\ P(P_k = p_{k,n}) \mathcal{N}(y_n - p_{k,n} g_{k,n} - (1 - p_{k,n})(\frac{\sum_{k' \neq k} x_{k',n} g_{k',n}}{S_{k'}}), R). \end{aligned} \tag{5.14}$$

By the law of non-linear transformation of random variables, where samples of $p_{k,n}$ are drawn according to the following probability distribution, the prior probability of P_k may be computed using the following equation rather than the Dirichelet used equation (5.11) in Technique A,

$$P(P_k = p_{k,n}) \sim Ga(\hat{\alpha}_{k,n}, \hat{\beta}_{k,n}) \frac{\partial P_k}{\partial X_k} \tag{5.15}$$

where the mean $\mu_k^P = \frac{\hat{\alpha}_{k,n} \hat{\beta}_{k,n}}{\hat{\alpha}_{k,n} \hat{\beta}_{k,n} + S_k}$ and its easy to show the variance is given by $\sigma_k^{2P} = \hat{\alpha}_{k,n} \hat{\beta}_{k,n}^2 (\frac{S_k}{(S_k + \mu_k^P)^2})^2$. In a Monte-Carlo sense $\hat{\alpha}_{k,n}, \hat{\beta}_{k,n}$ correspond to the hyperparameter values used in the previous Monte-Carlo simulation. Subsequently, the conditional posterior probability of X_k is computed from a Gamma distribution according to conditions stated in the appendix in sec. 5.10, and is given by,

$$P(X_k = p_{k,n} | X_{k' \neq k} = x_{k',n} \forall k', P_k = p_{k,n}, y_n, G_k = g_{k,n} \forall k = 1 \dots K) \sim Ga(\tilde{\alpha}_{k,n}, \tilde{\beta}_{k,n}) \tag{5.16}$$

where, $\tilde{\alpha}_{k,n} = \alpha_0 + Q_n$, $\tilde{\beta}_{k,n} = \hat{\beta}_{k,n} + (y_n - p_{k,n}g_{k,n} - (1-p_{k,n})(\frac{\sum_{k' \neq k} x_{k',n} g_{k',n}}{S}))^T R^{-1}(y_n - p_{k,n}g_{k,n} - (1-p_{k,n})(\frac{\sum_{k' \neq k} x_{k',n} g_{k',n}}{S_{k'}}))$.

5.5.5 Gibbs Sampler using Abundance Sampling Technique A

The following steps outline the Gibbs sampling algorithm:

Step 1: Estimate $\mu_{k,n}, \Sigma_{k,n}$ from a Gaussian process using (5.4) and spectral library Φ for K possible endmembers in the scene.

Step 2: Sample $g_{k,n}$ using Gaussian parameters from (5.4), and initialise $p_{1,n}^{(l)} \dots p_{K,n}^{(l)}$ using a uniform Dirichelet distribution, where $\alpha_{k,n} = 10$, $\beta_{k,n} = 0.5 \forall k = 1 \dots K$. Thus,

$$\begin{aligned} P(G_k = g_{k,n}^{(l)}) &\sim \mathcal{N}(\mu_{k,n}, \Sigma_{k,n}), \forall k = 1 \dots K \\ P(P_1 = p_{1,n}^{(l)}, \dots P_K = p_{K,n}^{(l)}) &\sim Dir(\frac{10}{0.5} \dots \frac{10}{0.5}) \end{aligned} \tag{5.17}$$

where, $D \times D$ noise covariance R, is initialised to a reasonable value and the value of $\beta_{k,n}$ remains fixed throughout the entire experiment.

Step 3: Using equation (5.7), estimate the lth endmember posterior estimate given previous values of $\hat{X}_k = x_{k,n}^{(l-1)}, \hat{\alpha}_{k,n}^{(l)}, \hat{\beta}_{k,n}^{(l)}$, re-sample $g_{k,n}^{(l)}$ using the updated means and covariances.

Step 4: Estimate the posterior of the abundance stated in (5.9) and derived in section 5.10, using re-sampled values of $g_{k,n}^{(l)} \forall k = 1 \dots K$ from step 3.

Step 5: Update hyperparameter $\hat{\alpha}_{k,n}$, using posterior estimates of $x_{k,n}^{(l)} \forall k = 1 \dots K$ from step 4,

$$\hat{\alpha}_{k,n} = \left\{ \begin{array}{ll} \frac{\tilde{\alpha}_{k,n}}{2} - c_1 & \text{iff } x_{k',n}^{(l)} \geq \tau_1 \\ \frac{\tilde{\alpha}_{k,n}}{2} + c_2 & \text{iff } x_{k',n}^{(l)} \leq \tau_2 \\ \frac{\tilde{\alpha}_{k,n}}{2} - c_3 & \text{iff } x_{k',n}^{(l)} < \tau_3 \end{array} \right\} \tag{5.18}$$

where, k', represents a sum of all values till k', constants $c_1 = c_2 = c_3 = 10$, $\tau_1 = 0.8, \tau_2 = 0.6, \tau_3 = 0.1$. The hyperparameter update reflects prior belief in the content of each pixel. The probability of higher abundances of a material is considered to be low, hence we penalise , which reduces the possible abundance values of that material as evident in the first constraint. Similarly, we believe for mapping problems, a material if present will exist at proportions greater than $c = 0.05$. Finally, we believe that the most likely scenario is for abundances to exist between $0.05 - 0.8$, hence we increase the variance for abundances that fall within this range.

Step 6: Re-sample $x_{k,n}^{(l)}$ according to (5.11) using posterior updates values of α from Step 4 and estimates of β from Step 2, where,

$$P(P_1 = p_{k,n}^{(l)}, \dots P_K = p_{K,n}^{(l)}) \sim Dir(\frac{\hat{\alpha}_{1,n}^{(l)}}{\hat{\beta}_{1,n}} \dots \frac{\hat{\alpha}_{K,n}^{(l)}}{\hat{\beta}_{K,n}^{(l)}}) \tag{5.19}$$

Step 7: Using posterior estimates, $g_{k,n}^{(l)}$ in Step 3 and $p_{k,n}^{(l)} \forall k = 1 \dots K$ from Step 6, the posterior noise variance is given by,

$$\sigma_n^{2(l)} \sim \mathcal{IG}(v_0 + 1, \rho + 1 + ||y_n - \sum_k g_{k,n} p_{k,n}||) \tag{5.20}$$

where, parameter values $v_0 = 20, \rho = 0.1 \forall n = 1 \dots N$. We choose the Inverse Gamma distribution since it has been successfully applied previously by Eches et. al in [65]. The parameter values were empirically determined to ensure that the noise variance is not too large. Bernardo et. al in [61] provide further details on the Inverse-Gamma distribution.

Step 8: Re-estimate $\mu_{k,n}, \Sigma_{k,n} \forall k = 1 \dots K$

$$\hat{\mu}_{k,n}^{(l)}, \hat{\Sigma}_{k,n}^{(l)} | g_{k,n}, p_{k,n} \forall k = \mathcal{GP}(g_{k,n} \forall k = 1 \dots K, \gamma), \tag{5.21}$$

where, $g_{k,n}^{(l)}, x_{k,n}^{(l)}$ are posterior draws, from steps 3 and 6.

Step 9: Repeat Steps $2-8$, for $l = 1 \dots L$ Monte-Carlo iterations. Calculate the likelihood across all $l = 1 \dots L$ iterations,

$$P(y_n, p_{k,n}^{(l)}, g_{k,n}^{(l)}, \mu_{k,n}^{G(l)}, \Sigma_{k,n}^{G(l)}, \hat{\alpha}_{k,n}^{(l)}, \hat{\beta}_{k,n}^{(l)} \forall k = 1 \dots K) =$$
$$Dir(p_{k,n}; \hat{\alpha}_{k,n}^{(l)}, \hat{\beta}_{k,n}^{(l)} \forall k = 1 \dots K) \prod_{k=1}^{K} \mathcal{N}(y_n; \sum_{k=1}^{K} g_{k,n}^{(l)} p_{k,n}^{(l)}, R) \mathcal{N}(g_{k,n}^{(l)}; \mu_{k,n}^{G(l)}, \hat{\Sigma}_{k,n}^{G(l)}) \quad (5.22)$$

Alternatively the first term before the product, can be substituted with a Gamma distribution with the same hyperparameters and multiplied K times.

Step 10: Identify the abundance estimate, $\hat{x}_n^{(l)} \in [0,1]^{K-1}$, and endmember estimates $g_{1,n}^{(l)} \dots g_{K,n}^{(l)} \in \mathbb{R}^{Q_n \times K}$ that produces the maximum likelihood value in (5.22).

5.5.6 Gibbs Sampler using Abundance Sampling Technique B

1. Apply Steps 1-3 according to Technique A.
2. Sample $p_{k,n}$ according to (5.16) $\forall k = 1 \dots K$.
3. Omit Step 5 and 6 from Technique A and continue with Steps $7-10$.

5.6 Recursive Band Selection using Beta Processes

Imposing sparsity to have a reduced band representation of each pixel facilitates for improved throughput, but to have a large proportion of bands consistently used throughout the scene to collect data is desirable since we can save cost and improve band Signal-to-Noise-Ratio (SNR) across a smaller subset of bands. Determining the exact bands as well as the exact number required to describe the contents of each pixel is still an open problem to be solved in the hyperspectral community. We setup a convex optimisation problem using training data to estimate a set of band weights to estimate the base measure of a Beta process. Whilst the base measure provides prior knowledge regarding the utility of each band for unmixing a particular set of materials, the posterior Beta process encodes the recursive aspect based on bands previously used without

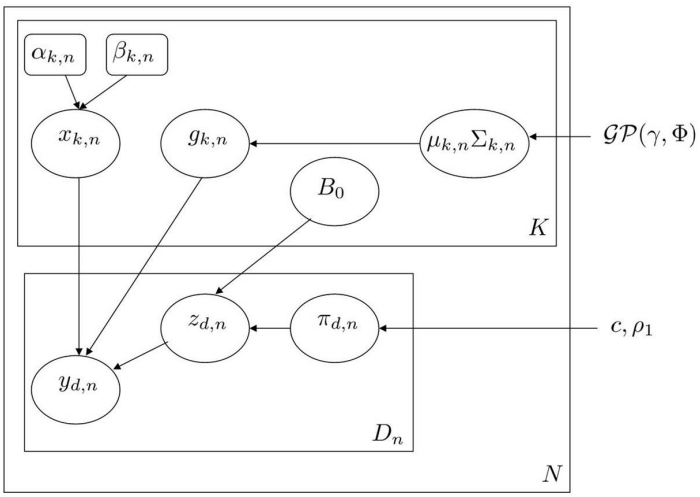


Figure 5.2. The following graphical model applies to test data where prior band utility is described by a base measure B_0 obtained from training data. Each nth pixel is represented by no more than $Q_n \subset D$ total number of bands in the sensor array, where value Q_n is drawn from a Poisson random variable, $\pi_n(d)$ is the posterior band utility which forms the posterior Beta process that depends on base distribution B_0 derived from training data and prior band labels. Hyperparameters for the Beta process are indicated by ρ_1, c. Both sets of weights are combined in estimating binary random matrix $Z = z_1(d), \ldots z_N(d)$ which is a result of successive draws from a Bernoulli random variable and forms the posterior Bernoulli process.

specifying which bands and how many to use. We update the existing sub-pixel model by introducing a binary matrix Z whose elements $z_n(d)$ describe whether the dth band is useful in describing the nth pixel. The updated model is shown in Fig. 5.2.

5.6.1 Estimating Base Measure using Convex Relaxation

In this section, we estimate the discrete base probability measure B_0, which is a band utility measure that provides prior knowledge of a band's potential usefulness using $\forall n = 1 \ldots N_t$ training pixels. The base measure is subsequently used by a stochastic Beta Process to carry out recursive online band selection on a test dataset. Band utility is measured across the training set through a convex optimisation procedure in Steps 11 and 12. Whereas, these two steps are omitted when using the test set and Steps 13 and 14 listed below carry out the recursive procedure. We first describe the convex

optimisation step to estimate the base measure B_0 for a training set.

Step 11: We seek $M_{s+1} \subset D$ bands that minimise abundance estimation error, where M_s denotes the M number of bands used to estimate the abundance at the sth iteration of the algorithm. M_s bands are removed for $s = 1 \dots S$ iterations of Steps 1-10, where S corresponds to the number of iterations until the inequality stated in (5.24) fails. We assume that best-case estimate even for pixels which are difficult to describe given the library, are resolved in a maximum L number of iterations in Steps $1 - 10$. We treat this Step as an outer loop to remove as many bands as possible before the abundance estimation error exceeds a certain threshold. The reason for removing just M_s bands at a time has to do with the fact that we do not know the exact number of bands before we exceed the threshold. The justification behind iterating through Steps 1-10 is that abundance estimates are likely to change when measurement set describing the endmembers changes in structure.

For the experiments conducted, $\xi = 0.1$, bands are removed as long as estimation performance does not suffer beyond this threshold in-comparison to the full-set. The combinatorial optimisation problem is given by,

$$\begin{aligned} \hat{T}_n^{M_{s+1}} &= \arg\min_{T_n^{M_s}} \left\| T_n^{TM_s}(\hat{g}_n^{M_s}(\hat{x}^{Q_n} - \hat{x}_n^{M_s})) \right\| < \xi \\ \text{s.t. } T_n^{M_s} &\in \{0,1\}, \sum^{M_s}(d)\hat{T}_n^{M_s}(d) == M_{s+1}, \forall n = 1 \dots N_t \text{ pixels}, \end{aligned} \tag{5.23}$$

where, $T_n^{TM_s} \in \{0,1\}^{1 \times M_s}$ is an indicator vector that denotes the critical bands for unmixing performance for each nth pixel, $\hat{g}_n \in \mathbb{R}^{M_s \times K}$ endmember matrix containing K M_s band endmembers, $\hat{x}_{k,n}^{M_s} \forall k = 1 \dots K$ is obtained after L iterations of steps $1 - 10$ and the value of M_{s+1} was set to maximum of $M_s - 50$ during the initial iterations and adjusted to $M_s - 10$ for the latter iterations when $\xi > 0.08$. The optimisation criteria is M_{s+1}^{th} band subset that minimise the difference in norm between the original abundance and the M_s band case. The parameter estimates $\hat{x}_n^D, \hat{x}_n^{M_s} \in [0,1]^{K \times 1}, \hat{g}_n$, are maximum likelihood abundance and endmember estimates obtained from measurements with $Q_n = D$ original bands and $Q_n = M_s < D$ reduced number of bands, after

L runs of Steps $1-10$. Online band selection for each of the $n = 1 \dots N_t$ training pixels is carried out as a function of the accuracy of the abundance estimate.

The problem stated in (5.23) is a combinatorial problem that is difficult to solve since measurements across certain bands are correlated which means dependency may exists across an unknown number of contiguous bands. The problem however can be addressed if we relax the constraints on the indicator variable, $T_n^{TM_s}$. The relaxed optimisation problem is given by,

$$\begin{aligned} \hat{T}_n^{M_{s+1}} &= \arg\min_{T_n^{M_s}} \left\| T_n^{TM_s}(\hat{g}_n^{M_s}(\hat{x}_n^P - \hat{x}_n^{M_s})) \right\| < \xi \\ \text{s.t. } 0 \leq T_n^{M_s} &\leq 1, \sum_{d=1}^{M_s} \hat{T}_n^{M_s}(d) == M_{s+1}, \forall n = 1 \dots N_t \text{ pixels,}. \end{aligned} \tag{5.24}$$

This problem is a convex optimisation problem since the objective, which is a L_2 norm, is convex with respect to the optimisation variable, the in-equality constraints are convex and equality constraints are affine. The problem is addressed in terms of its equivalent dual space via Lagrangian variables and solved using semi-definite programming. Refer to Chapter 2 and Boyd *et. al.* in [15] for further details. The final step in the band selection procedure is given by,

Step 12: Update bands for $y_n, g_{k,n} \forall K, N_t$ for the $s+1th$ iteration with,

$$\begin{aligned} y_n^{s+1} &= \hat{T}_n^{M_{s+1}} \circ y_n^{M_s T}, \\ \hat{g}_{k,n}^{M_{s+1}} &= \hat{T}_n^{M_{s+1}} \circ \hat{g}_n^{M_s T} \\ \forall n &= 1 \dots N_t \text{ pixels.} \end{aligned} \tag{5.25}$$

We compute the probability of the dth band being useful by counting the number of times it has been used over all N_t training image pixels. Thus,

$$B_0(d) = \frac{N(d)}{N_t} \forall d = 1 \dots Q_n \text{ bands} \tag{5.26}$$

where $N(d) = \sum_n \hat{T}_n(d) \forall n = 1 \dots N_t$ pixels and $B_0 \in [0, 1]$.

5.6.2 Beta and Bernoulli Processes

Consider a random measure B on a space $\Theta \in \mathbb{R}$, where Θ is a random variable and $\theta_n(1), \ldots \theta_n(Q) \sim \Theta$ are Q_n samples of Θ. The samples correspond to Q_n independent partitions of the space relevant to the nth pixel. $B(\theta_n(1)), \ldots B(\theta_n(Q))$ are masses assigned to independent partitions, if the masses assigned are also independent and do not sum to 1, the stochastic process is referred to as a Levy process characterised by a Levy measure, which is a measure on $\Theta \times [0,1]$. A Beta process is a positive Levy process, where $B \sim BP(c, B_0)$ whose Levy measure depends on a concentration parameter c and a fixed discrete measure $B_0(\Theta)$. B has the form, $B = \sum_d \pi_n(d)\delta_{\theta_n}(d)$, where, $\pi_n(d) \sim Beta(1,c)$, $\theta(d) \sim B_0(\Theta)$ is a draw from the discrete base measure $B_0(\Theta)$ and B represents the probability mass associated at each location $\theta_n(d)$.

Consider a Poisson process whose base measure is given by $\nu(\theta, \pi) = c\pi^{-1}(1-\pi)^{(c-1)}B_0(\Theta)$, where Θ, π are finite and c is uniform across all Θ. Such a process is also used to generate a probability mass for $\theta(d) \times \pi_n(d)$. It is shown by Thibaux *et. al.* in [29] that expectation of this process is finite if B_0 is finite for both discrete and continuous cases. If Θ is finite and fixed, this implies that there is a finite possible number of probabilities that will result from repeated draws of the Beta process and represents the Bernoulli distribution drawn from independent subsets. The Bernoulli process is a conjugate of the Beta process and is denoted by $Z|B \sim BeP(B)$ for $n = 1 \ldots N$ draws of B, where values of Z are represented by $z_n \in \{0,1\}^D$. Each row of Z constitutes a binary vector, z_n, with Q_n number of ones.

For this problem, $\theta_n(d)$ represents the maximum likelihood estimates of the nth pixel across each dth band representing posterior parameter estimates, where, $\theta_n(d) = \{\hat{g}_{k,n}(d), x_{k,n}(d), \pi_{k,n}(d) \forall k = 1 \ldots K\}$ represents the endmember measurement values across the dth band, $\hat{x}_{k,n}(d) \forall k = 1 \ldots K\}$ is uniform across Q_n bands and $\pi_n(d)$ is the probability that the dth band is useful. Due to neighbouring band correlation, typically the number of independent partitions is less than the total number of bands, $Q_n < D \forall n = 1 \ldots N$ pixels. Given that we want to estimate the most useful bands for

each pixel, we seek the posterior Beta process that provides the most likely bands for each subsequent pixel according to bands chosen to represent previous pixels as well as those bands with a large probability mass as per the base measure. We follow Thibaux et. al's [29] algorithm who construct the posterior Beta process as sum of independent Beta Processes, where,

$$\hat{B}_n = \hat{B}_{n-1} + \sum_{d=1}^{Q_n} \pi_n(d)\delta_{\theta_n}(d) \tag{5.27}$$

where, $B = \sum_n \hat{B}_n$, $\pi_n(d) \sim \text{Beta}(1, c+n-1)$ and Q_n is the number of bands chosen to represent the nth pixel. The algorithm continues to propose a series of steps to generate the posterior Beta process given by,

$$B|Z_{1\ldots N} \sim BP(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{n=1}^{N} Z_n) \tag{5.28}$$

where by using the independence property of Beta process and induction, this is equivalent to the following steps,

Step 13: Estimating the posterior Beta process, B

1. Sample $Q_n \sim \text{Poi}(\frac{c\rho_1}{c+n-1})$, where $\rho_1 = \sum_d B_0(d)$

2. Sample Q_n new band locations $\theta(d)$ according to $\frac{1}{\rho_1}B_0$

3. Sample the weight, $\pi_n(d) \sim \text{Beta}(1, c+n-1) \forall Q_n$ bands.

Step 14: Estimating the Bernoulli process, $Z_n|Z_1 \ldots Z_{n-1}, B_0$

Subsequent to computing the Beta processes, bands are included or omitted from the representation of each pixel, through a Hadamard product between the Bernoulli variable Z and the pixel y_n,

$$\tilde{y}_n = Z_n \circ y_n. \tag{5.29}$$

where, we sample Z_{n+1} according to a Bernoulli process given by,

$$Z_{n+1} \sim \text{BeP}(\frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{n=1}^{N} Z_n)\forall n = 1 \dots N. \tag{5.30}$$

where, the Bernoulli process is represented as two independent Bernoulli processes, $Z_{n+1} = U + V, U \sim BeP(\frac{c}{c+n}B_0), V \sim BeP(\sum_d \frac{m_n(d)}{c+n}\delta_{\theta_n(d)})$. The latter Bernoulli process V is equivalent to estimating Q_n and identifying new band locations (carried out in Step 13) hence it is sufficient that we estimate U, where $m_n(d)$ represents the number of times band d is used to represent abundance in $1 \dots n-1$ previous pixels. The estimation of $\pi_n(d)$ from Step 12 maybe omitted since its not required to estimate Z in spite of being part of the generative model.

5.6.3 RSBS Algorithm Summary

1. Implement unmixing algorithm using either Technique A or B and steps $11-12$ using a training dataset with $n = 1 \dots N_t$ pixels, to estimate B_0.

2. Apply unmixing using either Technique A or B followed by steps 13, 14 on a test dataset for $n = 1 \dots N$ pixels.

5.7 Experiments

We conduct two experiments with synthetic and real data. The purpose of these experiments is to show that the RSBS algorithm proposed in this study can do a comparable job in terms of unmixing performance relative to state-of-the-art algorithms using a much smaller subset of bands. The performance measures used include sum-squared error (SSE) for abundance estimates in synthetic data where the original abundance in each pixel is known and a mean SSE and standard deviation for the end-member reconstruction accuracy for a test dataset whose endmember ground-truth is well known. Both unmixing techniques are tested with the synthetic dataset but Technique B is omitted for the real dataset since it did not perform as favourably as Technique A.

Abundance estimates for 3 out of 5 materials were more accurate using Technique A rather than B. Moreover, Technique B required a greater number of Monte-Carlo simulations to reach a stable solution, hence it was also not used to estimate band weights using training data.

5.7.1 Experiment A

For the synthetic case, we use the data provided by Mittleman *et. al.* in [2] to benchmark the performance of our proposed method with current state-of-the-art techniques. We partition the synthetic data into training and test sets applying the convex-relaxation step to 100 out of 10000 pixels to measure band utility and derive the discrete base distribution B_0 shown in figure 5.3. Subsequently we apply the Beta process to the test set to carry out recursive band selection and measure the SSE across each pixel to characterise the abundance estimation performance across all endmembers. The synthetic data consists of upto 5 end-members arbitrarily selected from the USGS spectral database [1] where abundances of each pixel are randomly drawn [2] from a Dirichelet distribution where some pixels are represented by just three of these end-members. Additive Gaussian noise at $10dB$ Signal-to-Noise Ratio (SNR) was added to each pixel. We assume knowledge of the five possible end-members from the spectral library that can be present in any pixel and do not rely on end-member extraction techniques such as VCA [66] used by Mittelman *et. al.* in [2] to represent our initial end-members. Table 5.1 shows the minimum SSE of three existing algorithms, VCA (Vertex Component Analysis) , BLU (Bayesian Linear Unmixing) [59] and SCU (Spectral Constrained Unmixing) [2] as well as our proposed approach. The abundance SSE results show that overall SSE is lower than VCA , BLU and SCU algorithm. Performance across end-members are also lower in SSE mean and variance.

For the Beta process, we apply a concentration parameter $c = 1$ and $\rho_1 = 15$ which were fixed throughout the experiment on the basis that we wish to consider a large number of new bands but also apply a large weight to the bands used in previous pixels due to scene homogeneity. Figure 5.3 shows the prior or base distribution of the

Beta process. Fig. 5.5 shows the posterior Beta process at $50, 100, 1000, 10000$ pixels. Figure 5.7.2 shows the bands used at those particular pixel locations. The bands chosen are not always contiguous but are neighbours which implies that band correlation may not be a significant factor for unmixing performance, but shape information from absorption features maybe useful. There are no more than 24 bands used to describe each pixel in the entire synthetic image.

For the training set, using Rasmussen *et. al.*'s GPML toolbox used in [60], we specify a type 1 Gaussian process Matern kernel with a standard deviation of 0.1 and likelihood parameter to 30. This ensures that the function captures small absorption features even if the signal is non-smooth. The ground-truth data was down-sampled to 200 bands according to procedure proposed by Mittleman *et. al.* in [2], which were subsequently normalised to values between band numbers 1 and 224 using a linear mapping. The use of these widely spaced values as inputs to the Gaussian Process relative to original wavelengths ensures that the function is not over-sensitive and ignores drastic changes in spectra due to atmospheric noise. To accommodate for unknown noise sources and spectral variability, the likelihood value is adjusted to 0.8 and standard deviation to 5. The former change ensures a smoother function less sensitive to sporadic changes in the signal whilst lowering the likelihood parameter ensures that the function does not over-fit the training data.

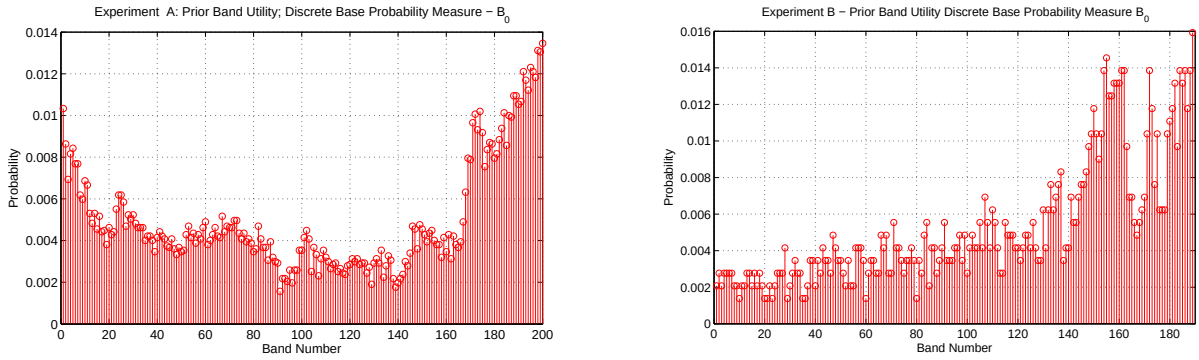
5.7.2 Experiment B

We assume knowledge of the potential materials in the scene and obtain ground-truth signatures (Alunite, Montmollonite, Sphene and Kaolinite) from the 2006 USGS database [1], removing bands $330 - 360$ through visual inspection on evidence of noisy absorption features. We down-sample AVIRIS-Cuprite dataset to 189 bands, eliminating atmospheric absorption and low SNR bands $1 - 2, 110 - 120, 151 - 170, 222 - 224$ [2]. Like the synthetic case we create some training data with these signatures using the same procedure and estimate a base distribution describing prior band utility. We use the openly available 80×80 sub-image used by Mittelman et.al. [2] with added $10dB$

Gaussian noise, to estimate the end-member extraction accuracy in terms of SSE mean. The mean is computed over all 6400 pixels and compared with the SSE mean computed by Mittelman *et. al.* for SCU, VCA and BLU across 20 different initialisations. The posterior endmember estimates are more accurate in terms of SSE than existing methods across all end-members. Figure 5.8 shows the difference between posterior and ground-truth endmembers used for an arbitrarily chosen pixel. In Fig. 5.8 a) although endmember estimates are not smooth they accurately capture the shape of the original endmember, which results in better endmember estimation accuracy than methods compared.

Figure 5.7 shows the abundance maps across four minerals. The abundance maps in a broad sense are well-matched with prominent regions in the visual image. In a broad sense the segmentation of material classes are quite distinct and match-up well to the visual image in fig. 5.7(a). Furthermore, the degree of sparsity is evident in the high abundances of certain materials and low values for the remainder at the same location. In fig. 5.7(a), the unique pink off-white region in the original image at top-left hand corner has a high abundance of 0.7 in fig. 5.7(e). In the same figure, the diagonal regions between locations $\{30, 32\}$ and $\{15, 15\}$ matches the distinct white region in-between the two green from fig. 5.7(a). The green regions have a smaller abundance than the white. The reverse is evident in fig. 5.7 f), where the green regions have the highest abundance relative to the rest of the image. The presence of Kaolinite is distinct in fig. 5.7(b), fig. 5.7(c) whilst the distinct brown region at the bottom of fig. 5.7(a) is evident in fig. 5.7(f). In fig. 5.7 d) parts of the road are visible around co-ordinates $\{30, 60\}$ to $\{70, 10\}$ whilst fig. 5.7 e) the sparsity of Sphene in certain regions is quite evident.

There are no more than 15 bands that are used throughout the entire scene in comparison to 24 bands used in the synthetic scene. The variation can be attributed to a larger difference in prior probabilities between bands greater than 140 to those less than 50 in comparison with Experiment A. This reduces the number of possible new band locations from which we sample from which results in a more homogenous band set required for the scene. Increasing the concentration parameter, c and weight, ρ_1, can



(a) (b)

Figure 5.3. The figure displays the base probability measure used for the Beta process, B_0 in experiment A and experiment B, where the prior probabilities indicate band utility. For both experiments, bands between $170 - 200$ are more useful than the remainder.

Table 5.1. Experiment A: Best-Case Abundance SSE for Synthetic Data: RSBS - Tech. A, noBandSel - Tech. A, B vs SCU, VCA, BLU

Endmem. No.	VCA	BLU	SCU	noBandSel - Tech. A	noBandSel - Tech B.	RSBS - Tech. A
1	68.28	39.51	14.07	**1.71**	**4.9**	**2.03**
2	108.97	55.14	18.23	**4.71**	**5.35**	**5.5**
3	83.24	18.38	13.28	**1.15**	**2.2**	**1.19**
4	49.13	22.41	9.75	**1.30**	**0.79**	**1.17**
5	67.59	41.15	18.16	**1.53**	**0.738**	**1.68**

account for potentially unknown materials and a non-homogenous scene but at a cost of using more bands to measure the scene.

5.8 Discussion

Both the abundance estimation accuracy and endmember estimation accuracy is better the state-of-the art methods tested using significantly fewer number of bands. The Beta process can be tuned to use fewer bands, we consider that option as part of future work. The proposed method involved tuning hyperparameters for both Gaussian and Beta processes, which can be carried out offline, these hyperparameters have a larger

Table 5.2. Experiment B: Endmember SSE against USGS ground-truth. Mean and Standard Deviation with added Gaussian Noise at 10dB SNR: RSBS vs SCU, VCA, BLU

SSE Mean - 10dB				
Endmem.	VCA	BLU	SCU	RSBS - Tech. A
Kaolin	1.34	1.23	1.02	**0.0761**
Kaolin 2	1.97	3.15	2.45	**0.076**
Alunite	11.87	9.97	8.42	**0.0759**
Mont.	2.59	3.62	2.98	**0.0755**
Sphene.	3	0.96	1.23	**0.0755**
SSE Std. Dev.				
Endmem.	VCA	BLU	SCU	RSBS - Tech. A
Kaolin.	0.84	0.25	0.26	**0.0081**
Kaolin. 2	0.3	0.35	0.29	**0.0081**
Alunite	0.3	0.35	0.29	**0.008**
Mont.	3.98	1.81	1.45	**0.008**
Sphene	0.5	0.44	0.34	**0.0083**

bearing on the result. Computational complexity is an issue with the Naive Gibbs approach, where $L > 100$ using both Techniques A and B throughout all experiments. For scenarios where there are $K >> 5$ endmembers in the library, $L >> 100$ significantly affecting algorithm performance. Further work is required to device strategies to reduce the number of Gibbs Sampling iterations. Some alternate approaches include Rao-Blackwellisation [27] or individual sampling of abundance co-ordinates through non-linear transformation of the Gamma random variables to enable sum-to-one constraints be applied implicitly as was the case in Technique B. Such an approach also ensures that the sampling occurs from the same Normal-Gamma distribution without the need for any normalisation achieved via the use of the Dirichelet even though

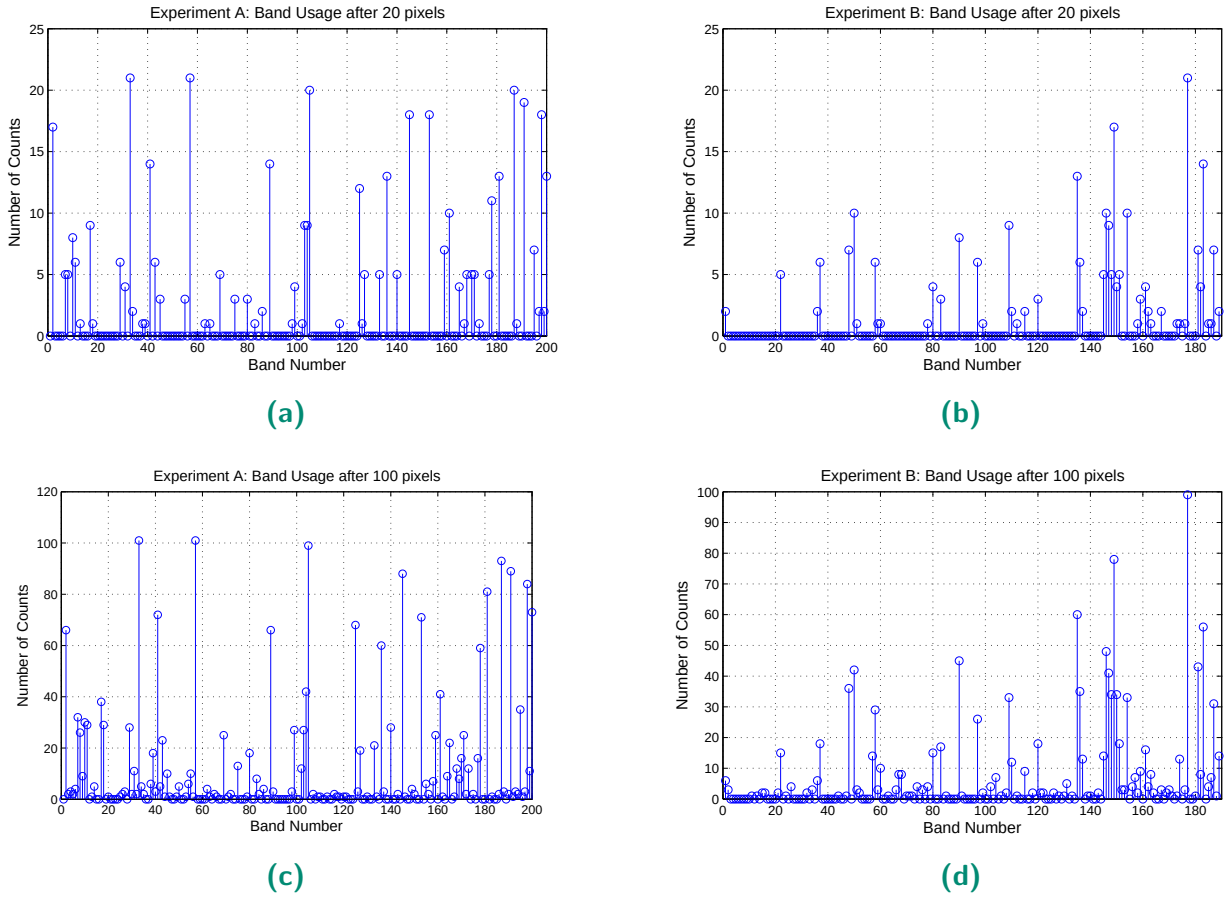


Figure 5.4. Experiment A a), c): Posterior Beta process measured after 20, 100 pixels, where the concentration parameter $c = 1$ and $\gamma = 15$. Experiment B b), d): Posterior Beta process measured after 20, 100 pixels, with the same hyper-parameters. Prior band utility is captured from the discrete base measures B_0 as is evident from the number of bands chosen after band 100. The small size of c ensures that bands used to describe previous pixels is captured. The size of the γ value ensures that a sufficient number of new bands are sampled as evident from bands with a small number of counts.

its unknown whether the same abundance estimation accuracy can be guaranteed especially when $K >> 5$. The utility of such a technique in a band selection scenario such as the one proposed in this paper is questionable due to the computational burden. The technique, however can be developed as a stand-alone abundance estimation technique as part of future work. We also propose to introduce and condition on a sparsity random variable to reduce computation time and improve accuracy when there are a large number of possible materials that are present in the scene. Dirichelet

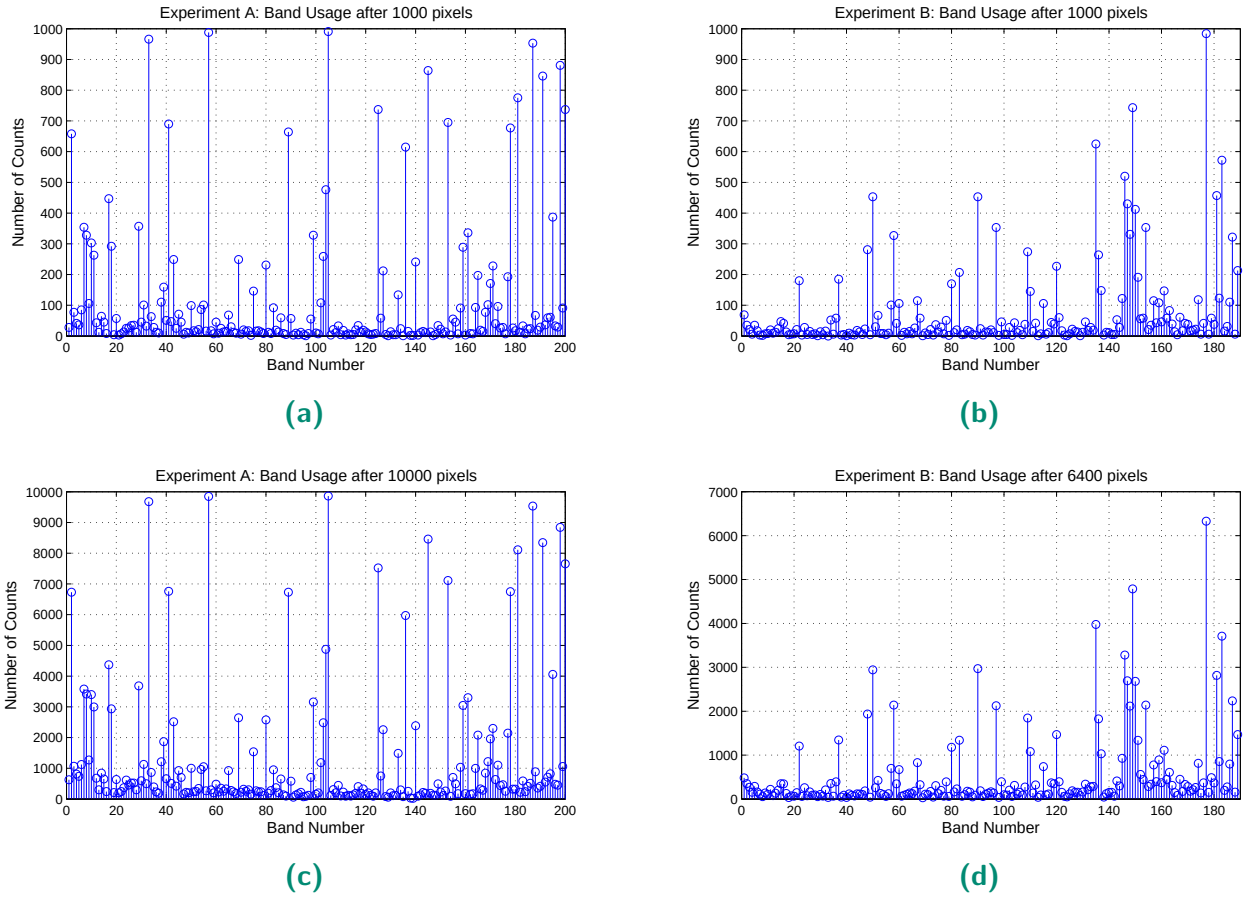


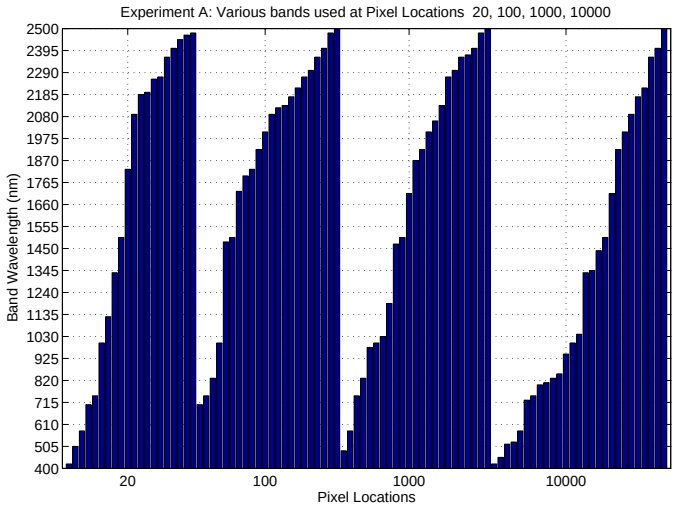
Figure 5.5. Experiment A a), c): Posterior Beta process measured after $1000, 10000$ pixels, where the concentration parameter $c = 1$ and $\gamma = 15$. Experiment B b), d): Posterior Beta process measured after $1000, 6400$ pixels, with the same hyper-parameters. Prior band utility is captured from the discrete base measures B_0 as is evident from the number of bands chosen after band 100. The small size of c ensures that bands used to describe previous pixels is captured. The size of the γ value ensures that a sufficient number of new bands are sampled as evident from bands with a small number of counts.

processes [27] provide an adequate mechanism to achieve this and enable the selection of M materials out of K total materials when $M << K$. The method is also applicable for surveillance scenarios where one seeks the material phenomenology to justify the presence of a material. As indicated in previous work [32], the bands used to describe each pixel can be referred to as critical bands whose locations may indicate the nature of material present in the pixel [32]. The band utility prior distribution, B_0 from both experiment A and B show a small cluster of neighbourhood bands that are more useful

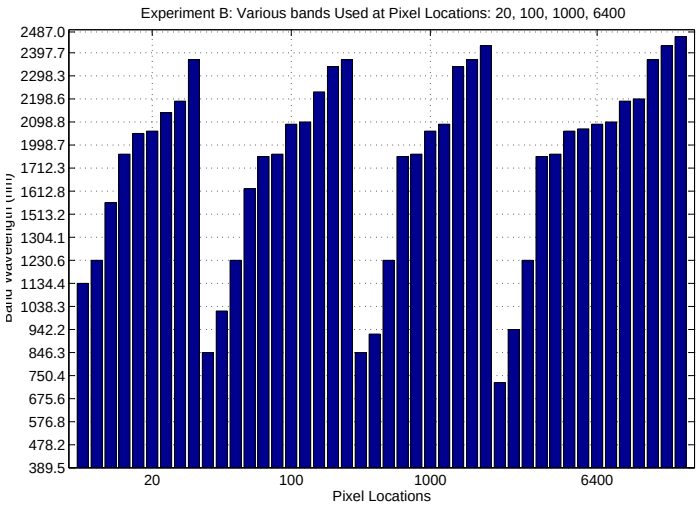
Table 5.3. Experiment A: Bands Used at different Pixel Locations

Pixel. No.	No. Bands Used	Bands used
20	20	2, 10, 17, 29, 33, 57, 69, 89, 105, 136, 161, 170, 171, 177, 178, 187, 191, 195, 197, 198
100	22	29, 33, 41, 57, 103, 105, 126, 133, 136, 145, 153, 161, 164, 165, 169, 173, 178, 181, 187, 191, 198, 200
1000	23	8, 17, 33, 41, 55, 57, 60, 75, 102, 105, 125, 140, 145, 153, 158, 165, 178, 181, 187, 188, 191, 198, 200
10000	27	2, 5, 11, 12, 17, 31, 33, 38, 39, 41, 43, 52, 57, 61, 89, 90, 99, 105, 125, 145, 153, 161, 169, 173, 187, 191, 200

than others which is also reflected in the band counts from the posterior Beta process in figure 5.5. Since this behaviour is not consistent across all bands, it can be said that sensor band correlation does not play a prominent role in unmixing performance. However, correlation due to material chemistry seems to have a positive impact that may contribute to improved unmixing performance. We note that critical bands quoted in Table 5.3 are band numbers and not wavelengths. Further details regarding exact wavelengths chosen are a subject for future study since the study is a proof-of-concept. It was also brought to our attention from a reviewer that the proposed RSBS method is applicable for any tunable detector in general used in pushbroom sensing which is more widely used from a surveillance standpoint. This broadens the applicability of this technique.

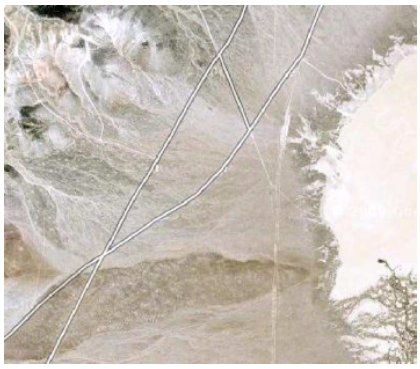


(a)

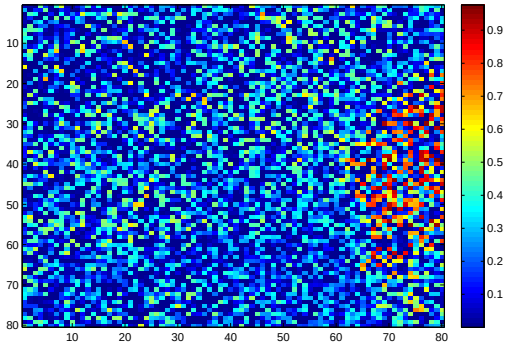


(b)

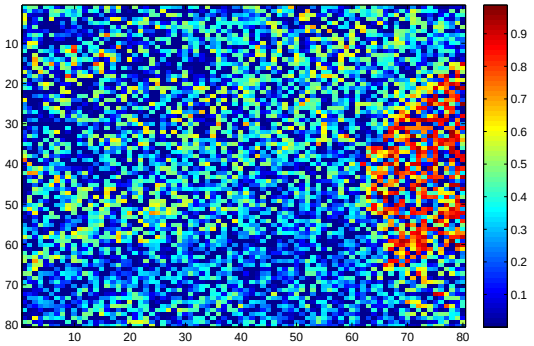
Figure 5.6. a) Experiment A: shows $20, 22, 23, 27$ bands used at pixel locations $20, 100$ and $1000, 10000$. Bands greater than 140 or $1500nm$ are used consistently across all 4 locations. This is agreeable with the base measure which contains larger probabilities at these locations. b) Experiment B: Only $10 - 14$ bands are used at the four locations $20, 100, 1000, 6400$ pixels. Bands greater than 1700 nm are used consistently across these locations.



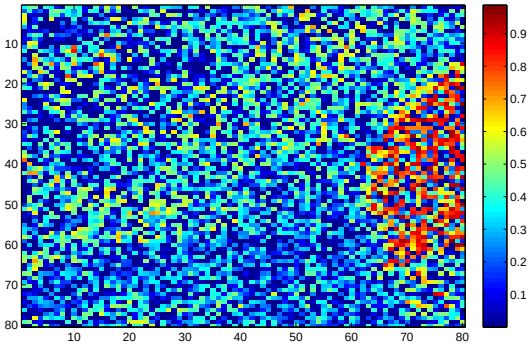
(a)



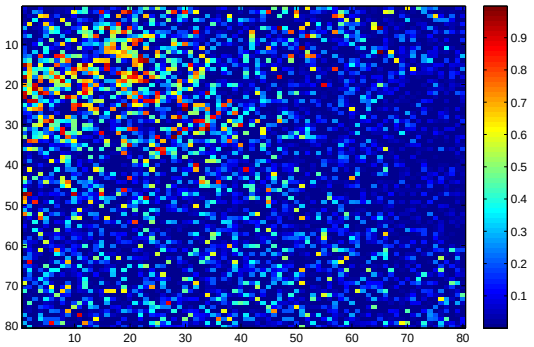
(b)



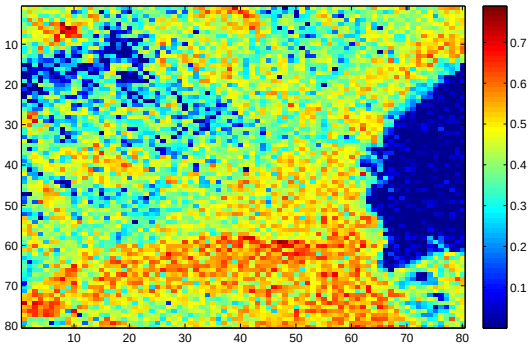
(c)



(d)

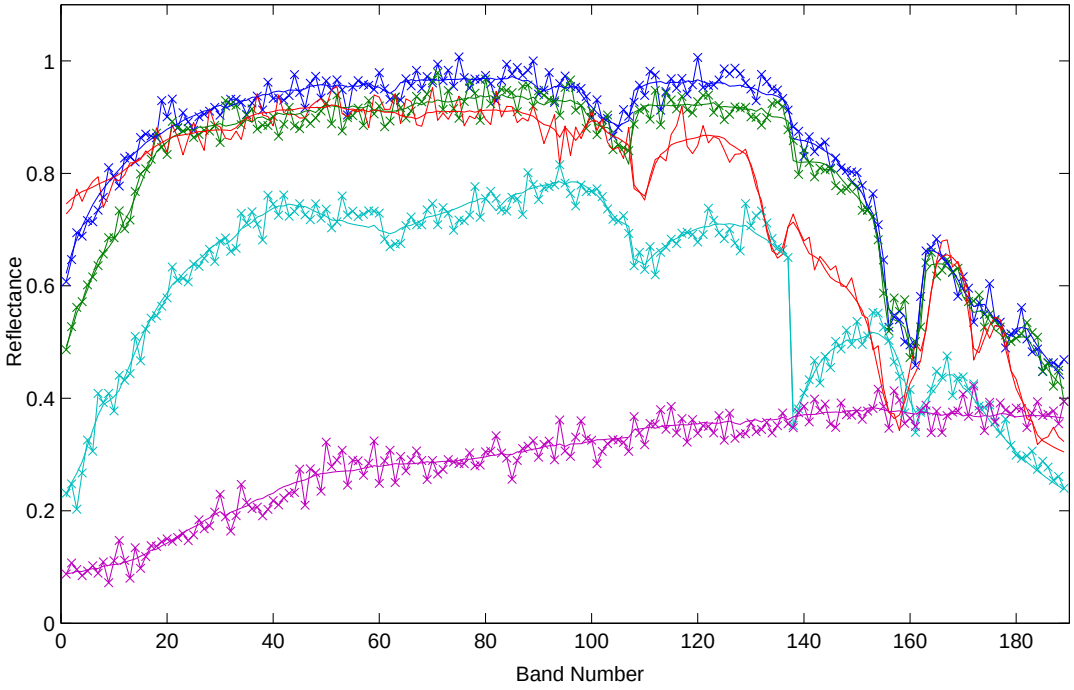


(e)



(f)

Figure 5.7. Experiment B: Abundance Map of AVIRIS-Cuprite image subset. a) Original Image (courtesy Mittleman *et. al.* [2]). b) Kaolnite 1 c) Kaolinite 2 d) Alunite e) Montmorollinite f) Sphene



(a)

Figure 5.8. Experiment B: a) Posterior endmember estimates at the first pixel of Kaolinite 1 (blue), Kaolinite 2 (green), Alunite (red), Montmorillonite (cyan) and Sphene (magenta). The signatures marked with crosses represent the estimate whilst those without any markers represent the true value.

5.9 Conclusion

We propose a method to carry out online band selection when the hyperspectral sensor is carrying an AFPA. The proposed method directly uses unmixing criteria to evaluate the utility of a selected subset of bands. Gaussian processes are used to preserve the natural variations of endmembers, the Normal-Gamma conjugacy and Dirichelet-Gamma relation offer a convenient posterior update of the abundance resulting in sparse abundance estimates and results that are significantly better in abundance and endmember estimation performance compared to state-of-the art Bayesian techniques.

Beta processes enable recursive online band selection through prior knowledge of possible materials and in the scene, where a prior discrete probability describing band utility is evaluated using a convex relaxation approach. The method applied demonstrates that only a small number of bands are required to achieve comparable unmixing performance as existing algorithms. Further work would entail the inclusion of Dirichelet processes under the same framework to handle large spectral libraries which handles scenarios where there is greater uncertainty over the possible endmembers in the scene.

5.10 Appendix

We derive the distribution of the Gamma abundance conditional on the multivariate Gaussian endmember based on Prado *et. al.*'s treatment in [64]. Since $x_{k,n} \sim Ga(\alpha_{k,n}/2, \beta_{k,n}/2)$. The prior distribution is given by,

$$P(x_{k,n}) = \frac{\beta_{k,n}^{\alpha_{k,n}/2}}{2^{\alpha_{k,n}/2}\Gamma(\frac{\alpha_{k,n}}{2})} x_{k,n}^{\frac{\alpha_{k,n}}{2}-1} \exp(-x_{k,n}\beta_{k,n}/2) \tag{5.31}$$

where, $\Gamma()$ refers to the gamma function. The conditional distribution of a univariate Gaussian given a univariate Gamma random variable is represented by, $\tilde{g}|x_{k,n} \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2 x_{k,n}^{-1})$, where $\tilde{g}, \mu, \sigma^2 \in \mathbb{R}$ are scalar. Thus, the joint density of $x_{k,n}, \tilde{g}$ is given by,

$$\begin{aligned} p(\tilde{g}, x_{k,n}) &= (\frac{x_{k,n}}{2\pi\sigma^2})^{1/2} \exp[\frac{-x_{k,n}(\tilde{g}-\tilde{\mu})^2}{2\sigma^2}] \\ &\times \frac{\beta_{k,n}^{\alpha_{k,n}/2}}{2^{\alpha_{k,n}/2}\Gamma(\frac{\alpha_{k,n}}{2})} x_{k,n}^{\frac{\alpha_{k,n}}{2}-1} \exp[-\frac{x_{k,n}\beta_{k,n}}{2}] \\ &\propto x_{k,n}^{(\frac{\alpha_{k,n}+1}{2})-1} \exp[-\frac{x_{k,n}}{2}\{\frac{(\tilde{g}-\tilde{\mu})^2}{\sigma_{k,n}^2} + \beta_{k,n}\}] \end{aligned} \tag{5.32}$$

where, the conditional probability $x_{k,n}$ given $\tilde{g}$,

$$\begin{aligned} p(x_{k,n}|\tilde{g}) &\propto x_{k,n}^{(\frac{\alpha_{k,n}+1}{2})-1} \exp[-\frac{x_{k,n}}{2}\{\frac{(\tilde{g}-\tilde{\mu})^2}{\sigma^2} + \beta_{k,n}\}] \\ x_{k,n}|\tilde{g} &\sim Ga(\frac{\tilde{\alpha}_{k,n}}{2}, \frac{\tilde{\beta}_{k,n}}{2}) \end{aligned} \tag{5.33}$$

where, $\tilde{\alpha}_{k,n} = \alpha_{k,n} + 1$ and $\tilde{\beta}_{k,n} = \beta_{k,n} + (\tilde{g}-\tilde{\mu})^2/\sigma^2$. For the multi-variate case, we are interested in the conditional probability of the abundance given the multivariate Gaussian endmember $g_{k,n}$. Thus, $p(x_{k,n}|g_{k,n}) \sim G(\tilde{\alpha}_{k,n}/2, \tilde{\beta}_{k,n}/2)$, where $P(g_{k,n}|x_{k,n}\forall k, x_{k',n}\forall k', y_n) \sim$

$\mathcal{N}(\mu_{k,n}^{G(l)}, \Sigma_{k,n}^{G(l)})$. From (5.33) the multivariate hyperparameters, $\tilde{\alpha}_{k,n}, \tilde{\beta}_{k,n}$ are given by,

$$\begin{aligned} \tilde{\alpha}_{k,n} &= \tilde{\alpha}_{k,n} + Q_n, \\ \tilde{\beta}_{k,n} &= \beta_{k,n} + (g_{k,n} - \mu_{k,n})^T \Sigma_{k,n}^{-1} (g_{k,n} - \mu_{k,n}) \end{aligned} \tag{5.34}$$

where Q_n is the degrees of freedom, $\mu_{k,n} \in \mathbb{R}_n^Q, \Sigma_{k,n} \in \mathbb{R}^{Q_n \times Q_n}$ refer to the endmember mean and covariance from the prior distribution, where, $g_{k,n} \sim \mathcal{N}(\mu_{k,n}^{G(l)}, \Sigma_{k,n}^{G(l)})$.

Chapter 6

Concluding Remarks

THIS chapter concludes the thesis by highlighting take home messages for sensor designers and analysts from results obtained in previous chapters. The chapter also highlights briefly further technical work to be completed to enhance the understanding and utility of the band selection work carried out in this thesis.

6.1 Conclusion

From the work conducted in this thesis it is evident that the band selection problem is dependent on the broader context for which it is used. If the aim is to improve model estimation accuracy reducing noise and improving data throughput, the work conducted in chapter 3 is relevant. The limitations are the prior knowledge required on the possible number of useful bands and assumption of structured local band correlation from overlapping sensor responses. For surveillance applications, where the aim might be to find anomalies and isolate critical band wavelengths that reveal the anomalies' salient chemical properties, chapter 4 is relevant. The algorithm and results derived in this chapter demonstrate that unknown subsets of contiguous narrow band wavelengths contribute the most in detecting anomalies, making these bands critical and subject to further improvement. The unsupervised nature of this algorithm enable the reduction of data complexity both spatially and spectrally reducing the processing and throughput burden on the image analyst. In this chapter we assume a locally correlated band structure and use a Gaussian mixture model when there is the possibility of sub-pixel anomalies which are ideally captured by compositional sub-pixel models. Finally, in scenarios where band selection is needed to be carried out online and some prior knowledge of possible target materials in the scene is known, chapter 5 is relevant. Band selection is carried out online, under a sub-pixel model without any assumptions on the band-correlation structure. The proposed framework also provides the ability to tune and influence band selection performance prior to data collection which provides a sense of robustness that is not available under methods proposed in previous chapters. For the scene tested, results demonstrate that it is the capacity of the proposed model and algorithm to estimate abundances in each pixel that has the largest bearing on the accuracy more so than the bands selected. Results indicate that the work carried out in the last chapter 5 is the way forward both in terms of spectral unmixing and unsupervised band selection both in terms of computational complexity in terms of number of bands used and the fact that fewer assumptions are made in the study than previous chapters. The caveat is that prior knowledge of possible endmembers in the scene is required. Band correlation inherent in the sensor as well as in the materials observed also does not seem to play a critical role for unmixing performance

since the bands selected are not necessarily contiguous unlike in 4. This highlights the importance of contiguous spectral bands for maximising divergence between anomaly and backgrounds and their lack of importance in blind source separation or un-mixing problems. Needless to say, this has serious implications for sensor design and reflects the importance of analysis techniques in sensor design.

6.2 Recommendations on Future Work

1. Dirchelet distributions provide inherent sparsity, through a simple hyperparameter update which results in sparse abundance estimates and is a key contributor in the abundance estimation accuracy. Studying some of the theoretical properties of this distribution and deriving update rules for hyperparameters that result in maximum likelihood estimates of the abundance would speed up algorithm convergence which is relevant for online applications.

2. Expanding the size of the spectral library is required to make the algorithm in Chapter 5 more realistic and practical. This can be enabled by applying the work of [57] who introduce non-parametric Bayesian methods to derive posterior estimates of a sparse number of signals from a spectral library that can be used to describe each pixel. Such a method would enable the use of large spectral library of signatures representing the possible endmembers in the scene.

3. Eigen-decomposition methods can be used to describe the behaviour of the iterative algorithm used to label measurements as outlier and partial backgrounds in Chapter 4. Such a technique would enable the description of the changes to outlier and partial background probability distributions in terms of rotation and scaling which provides an indicator as to what is occurring during this process.

References

[1] R. N. Clark, G. A. Swayze, R. Wise, E. Livo, T. Hoefen, R. Kokaly, and S. J. Sutley, *USGS digital spectral library splib06a*. US Geological Survey Denver, CO, 2007.

[2] R. Mittelman, N. Dobigeon, and A. O. Hero, "Hyperspectral image unmixing using a multiresolution sticky hdp," *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1656–1671, 2012.

[3] R. A. Schowengerdt, *Remote Sensing, Third Edition: Models and Methods for Image Processing*. Orlando, FL, USA: Academic Press, Inc., 2006.

[4] E. Puckrin, G. M. B. J. F. V. Turcotte, Caroline S., and M. Chamberland, "Airborne infrared hyperspectral imager for intelligence, surveillance, and reconnaissance applications," *Proc. SPIE*, vol. 8360, pp. 836 004–836 004–10, 2012.

[5] C. M. Stellman, G. Hazel, F. Bucholtz, J. V. Michalowicz, A. D. Stocker, and W. Schaaf, "Real-time hyperspectral detection and cuing," *Optical Engineering*, vol. 39, pp. 1928–1935, July 2000.

[6] F. van der Meer, "Analysis of spectral absorption features in hyperspectral imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 5, no. 1, pp. 55–68, Feb 2004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0303243403000382

[7] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, no. 1, pp. 65–74, 1988. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=3001

[8] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *Signal Processing Magazine, IEEE*, vol. 19, pp. 29–43, Jan. 2002.

[9] W. J. Gunning, J. L. Johnson, and J. F. DeNatale, "Lwir/mwir adaptive focal plane array," in *Proceedings of SPIE*, vol. 5612, 2004, p. 78.

[10] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58–69, Jan 2002.

[11] D. Marden and D. Manolakis, "Modeling hyperspectral imaging data," in *Proceedings of SPIE*, vol. 5093, 2003, p. 253.

[12] D. Manolakis, "Realistic matched filter performance prediction for hyperspectral target detection," *Optical Engineering*, vol. 44, no. 11, pp. 116 401–116 401, 2005.

[13] J. Theiler and C. Scovel, "Uncorrelated versus independent elliptically-contoured distributions for anomalous change detection in hyperspectral imagery," in *Proc. SPIE*, vol. 7246, 2009, p. 72460T.

[14] D. W. Stein, S. E. Stewart, G. D. Gilbert, and J. S. Schoonmaker, "Band selection for viewing underwater objects using hyperspectral sensors," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 3761, Oct. 1999, pp. 50–61.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[16] J. Kerekes and D. Snyder, "Unresolved target detection blind test project overview," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2010, pp. 769 521–769 521.

[17] J. Kerekes, "Rochester institute of technology target detection blind test project," 2008, http://dirsapps.cis.rit.edu/blindtest/information/.

[18] S. M. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1)," 1993.

[19] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for gaussian mixtures," *Neural computation*, vol. 8, no. 1, pp. 129–151, 1996.

[20] D. M. Titterington, A. F. M. Smith, and U. E. Makov. Wiley, 1985.

[21] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[22] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368.

[23] S. R. R. Salakhutdinov and Z. Ghahramani, "Optimization with em and expectation-conjugate-gradient," in *Intl. Conf. on Machine Learning (ICML (2003)*, 2003, pp. 672–679.

[24] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, pp. 2664–2677, 1994.

[25] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 1, pp. 161–173, 2009.

[26] S. M. Ross, *A course in simulation*. Prentice Hall PTR, 1990.

[27] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Cambridge, MA, USA, 2006, aAI0809973.

[28] M. I. Jordan, "Hierarchical models, nested models and completely random measures," *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger. New York: Springer*, 2010.

[29] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, vol. 11, 2007, pp. 564–571.

[30] B.-C. Gao, M. J. Montes, C. O. Davis, and A. F. Goetz, "Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean," *Remote Sensing of Environment*, vol. 113, pp. S17–S24, 2009.

[31] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 522–526, Oct 2006.

[32] G. Balasubramanian, V. K. Shettigara, S. Angeli, and G. A. Fowler, "Band selection using support vector machines for improving target detection in hyperspectral images," in *Proceedings of the 9th Biennial Conference on Digital Image Computing Techniques and Applications (DICTA07)*. IEEE, 2007, pp. 446–453.

[33] H. Du, H. Qi, X. Wang, R. Ramanath, and W. E. Snyder, "Band selection using independent components analysis for hyperspectral image processing," 2003, pp. 93–98.

[34] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 1552–1565, Jul 2004. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1315839

[35] F. Sawo, D. Brunn, and U. Hanebeck, "Parameterized joint densities with gaussian mixture marginals and their potential use in nonlinear robust estimation," in *International Conference on Control Applications*. IEEE, 2006, pp. 301–306.

[36] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 4, pp. 564–568, 2008.

[37] M. Graham and D. Miller, "Unsupervised learning of parsimonious mixtures on large spaces with integrated frature and component selection," *Transactions on Signal Processing, IEEE*, vol. 54, pp. 1289–1303, Apr. 2006.

[38] M. Law, M. Figueiredo, and A. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, Sep 2004. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1316850

[39] V. Roth and T. Lange, "Feature-selection in clustering problems," in *Proc. IEEE International Workshop on Microelectromechanical Systems (MEMS'97)*, Nagoya, Japan, Jan. 1997, pp. 290–294.

[40] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 155–176, 1996.

[41] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *Journal of the American Statistical Association*, vol. 89, pp. 1255–1270, 1993.

[42] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[43] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, Feb 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4663892

[44] A. Berger, V. Pietra, and S. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[45] G. Healey and D. Slater, "Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions," *IEEE Trans Geosc and Remote Sensing*, vol. 37, no. 6, pp. 2706–2717, 1999.

[46] T. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, July 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021039

[47] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx/, Apr. 2011.

[48] K. Kampa, E. Hasanbelliu, and J. Principe, "Closed-form cauchy-schwarz pdf divergence for mixture of gaussians," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2011, pp. 2578–2585.

[49] D. Snyder, J. Kerekes, I. Fairweather, R. Crabtree, J. Shive, and S. Hager, "Development of a web-based application to evaluate target finding algorithms," in *Geoscience and Remote Sensing Symposium, (IGARSS)*, vol. 2. IEEE, 2008, pp. II–915.

[50] T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields, "The hymap(tm) airborne hyperspectral sensor: The system, calibration and performance," 1998, pp. 37–42.

[51] J. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 2, pp. 354–379, 2012.

[52] M. D. Farrell Jr and R. M. Mersereau, "On the impact of pca dimension reduction for hyperspectral detection of difficult targets," *Geoscience and Remote Sensing Letters, IEEE*, vol. 2, no. 2, pp. 192–195, 2005.

[53] A. Zare and P. Gader, "Sparsity promoting iterated constrained endmember detection in hyperspectral imagery," *Geoscience and Remote Sensing Letters, IEEE*, vol. 4, no. 3, pp. 446–450, 2007.

[54] J. M. Nascimento and J. B. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, 2005.

[55] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse unmixing of hyperspectral data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. IEEE, 2012, pp. 7488–7491.

[56] A. S. Charles, B. A. Olshausen, and C. J. Rozell, "Learning sparse codes for hyperspectral imagery," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 5, pp. 963–978, 2011.

[57] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images," *Image Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 130–144, 2012.

[58] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. application to hyperspectral imagery," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1403–1413, 2010.

[59] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *Signal Processing, IEEE Transactions on*, vol. 57, no. 11, pp. 4355–4368, 2009.

[60] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 1.

[61] J. M. Bernardo, J. Berger, A. P. Dawid, and A. Smith, *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*. Oxford University Press, USA, 1999, vol. 6.

[62] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Sharing features among dynamical systems with beta processes," *Advances in Neural Information Processing Systems*, vol. 22, pp. 549–557, 2009.

[63] G. Welch and G. Bishop, "An introduction to the kalman filter."

[64] R. Prado and M. West, *Time Series: Modeling, Computation, and Inference*. Chapman & Hall, 2010.

[65] O. Eches, N. Dobigeon, and J.-Y. Tourneret, "Enhancing hyperspectral image unmixing with spatial correlations," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 11, pp. 4239–4247, 2011.

[66] J. M. Nascimento and J. B. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, 2005.