

Mumford–Shah type models for unsupervised hyperspectral image segmentation

Von der Fakultät für Mathematik und Naturwissenschaften der RWTH Aachen
University zur Erlangung des akademischen Grades eines Doktors der
Naturwissenschaften genehmigte Dissertation

vorgelegt von

Master of Science

Jan-Christopher Cohrs

aus

Uelzen.

Berichter: Prof. Dr. Benjamin Berkels
Prof. Dr. Lars Grasedyck

Tag der mündlichen Prüfung: 08.12.2025

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Abstract

Hyperspectral sensors provide images that are rich in information by accurately resolving the incoming spectra with a high sampling rate. The idea is that, by resolving the spectra with a sufficiently high level of detail, each constituent in the captured scene or specimen can be identified by its unique spectral fingerprint, the so-called spectral signature. This allows for a differentiation of the contributing constituents on a pixel level and, thus, leads to high quality image segmentations. However, some special characteristic properties of hyperspectral images like the strong noise and the spectral variability adversely affect the quality of the segmentations of such images. Several approaches for hyperspectral image segmentation have been proposed. A shortcoming of supervised methods is that the process of generating labeled training data is expensive and time-consuming, making unsupervised methods an important tool to solve the hyperspectral segmentation task. While different methods for unsupervised hyperspectral image segmentation have been introduced, an accurate description of the data despite the spectral variability and noise remains a decisive challenge.

In this thesis, we propose and investigate a novel segmentation framework and three new models for unsupervised hyperspectral image segmentation. The framework comprises a preprocessing for noise and dimensionality reduction by the minimum noise fraction transform, an alternating optimization approach to minimize the objective functional and a stopping criterion. The three introduced models, ϵ AMS, ϕ AMS and kMS, are based on the Mumford–Shah segmentation functional. The models ϵ AMS and ϕ AMS aim to directly model the spectra with first- and second-order statistics, while kMS maps the spectra into a higher-dimensional Hilbert space to describe the data there. To ensure that ϵ AMS and ϕ AMS remain always feasible, we propose a regularization of the covariance matrices for each of the models. Furthermore, we prove the existence of minimizers for both models. Additionally, we investigate the importance of the regularization parameter of ϵ AMS rigorously: we prove the Γ -convergence of the corresponding functional to a Γ -limit and see that we lose the guarantee of a minimizer in the limit with a counterexample. We solve the problem of the lack of closed-form solutions for the model-specific parameters of ϵ AMS and ϕ AMS for optimization by introducing fixed point iteration schemes. Furthermore, we derive a closed-form solution for the model-specific parameters of kMS. Extensive numerical experiments on four publicly available datasets show the great potential of all three methods. In particular, ϵ AMS and ϕ AMS show consistently the best performances and provide segmentations of the highest qualities among all competing methods, which include state-of-the-art methods for unsupervised hyperspectral image segmentation. An evaluation of the effect of the preprocessing by the minimum noise fraction transform on the segmentation results shows that the preprocessing has a clear positive effect but the main contribution comes from the models themselves. Finally, we test ϵ AMS also on multispectral Sentinel-2 data taken over the Arctic region and find out that the model is able to derive complex sea ice states from these images.

The models presented in this thesis are able to produce segmentations of high quality and have a great potential to enhance techniques used in practice that apply hyperspectral segmentation methods. Additionally, the theoretical analysis of the models provides a solid understanding of them and yields starting points for further improvements.

Zusammenfassung

Hyperspektralsensoren liefern Bilder, die durch genaues Auflösen der einfallenden Spektren mit einer hohen Abtastrate reich an Informationen sind. Die Idee ist, dass durch eine Auflösung der Spektren mit einem hinreichend hohen Detailgrad jeder Bestandteil der aufgenommenen Szene oder Probe durch seinen eindeutigen spektralen Fingerabdruck, die sogenannte spektrale Signatur, identifiziert werden kann. Dies erlaubt eine Unterscheidung der beitragenden Bestandteile auf Pixelebene und führt folglich zu qualitativ hochwertigen Bildsegmentierungen. Einige spezielle Charakteristika von Hyperspektralbildern wie starkes Rauschen und die spektrale Variabilität haben allerdings einen negativen Einfluss auf die Qualität der Segmentierung solcher Bilder. Diverse Methoden zur Segmentierung von Hyperspektralbildern wurden vorgeschlagen. Eine Schwachstelle von überwachten Methoden ist, dass der Prozess der Generierung von gekennzeichneten Trainingsdaten teuer und zeitaufwendig ist, was unüberwachte Methoden zu einem wichtigen Werkzeug zur Lösung des Problems der hyperspektralen Segmentierung macht. Auch wenn unterschiedliche Methoden zur unüberwachten hyperspektralen Bildsegmentierung vorgeschlagen wurden, bleibt die genaue Beschreibung der Hyperspektraldaten aufgrund der spektralen Variabilität und des Rauschens eine entscheidende Herausforderung.

In dieser Dissertation werden ein neuer Segmentierungsframework sowie drei neue Modelle zur unüberwachten hyperspektralen Bildsegmentierung vorgeschlagen und untersucht. Der Framework umfasst eine Vorverarbeitung zur Rausch- und Dimensionsreduktion durch die Minimum-Noise-Fraction-Transformation, einen alternierenden Optimierungsansatz zur Minimierung des Zielfunktional und ein Abbruchkriterium. Die drei eingeführten Modelle, ϵ AMS, ϕ AMS und kMS, basieren auf dem Mumford–Shah-Segmentierungsfunktional.

Die Modelle ϵ AMS und ϕ AMS zielen darauf ab, die Spektren direkt mit Statistiken erster und zweiter Ordnung zu modellieren, während kMS die Spektren in einen höherdimensionalen Hilbertraum abbildet, um die Daten dort zu beschreiben. Um sicherzustellen, dass ϵ AMS und ϕ AMS stets durchführbar sind, schlagen wir für jedes der beiden Modelle eine Regularisierung der Kovarianzmatrizen vor. Des Weiteren beweisen wir für beide Modelle die Existenz von Minimierern. Zusätzlich führen wir eine rigorose Untersuchung der Bedeutung des Regularisierungsparameters des Modells ϵ AMS durch: Wir beweisen die Γ -Konvergenz des zugehörigen Funktional gegen einen Γ -Grenzwert und sehen mithilfe eines Gegenbeispiels, dass wir die Garantie eines Minimierers im Grenzwert verlieren. Um das Fehlen geschlossener Lösungsformeln zur Optimierung der modellspezifischen Parameter von ϵ AMS und ϕ AMS zu lösen, schlagen wir Fixpunktiterationen vor. Außerdem leiten wir eine geschlossene Lösungsformel für die modellspezifischen Parameter von kMS her. Umfangreiche numerische Experimente auf vier öffentlich verfügbaren Datensätzen zeigen das große Potential dieser drei Methoden. Insbesondere ϵ AMS und ϕ AMS zeigen durchgehend die besten Ergebnisse und liefern Segmentierungen von der höchsten Qualität unter allen verglichenen Methoden, die auch State-of-the-art-Methoden für unüberwachte hyperspektrale Bildsegmentierung enthalten. Eine Evaluierung des Einflusses der Vorverarbeitung durch die Minimum-Noise-Fraction-Transformation auf die Segmentierungsergebnisse zeigt, dass die Vorverarbeitung einen klaren positiven Effekt hat, der Hauptbeitrag zu den Segmentierungsergebnissen allerdings von den Modellen selbst kommt. Abschließend testen wir ϵ AMS auch auf multispektralen Sentinel-2-Daten, die über der Arktisregion aufgenommen wurden, und finden heraus, dass das Modell fähig ist, aus diesen Bildern komplexe Meereiszustände abzuleiten.

Die in dieser Dissertation präsentierten Modelle sind in der Lage, Segmentierungen von hoher Qualität zu produzieren, und haben das große Potential, in der Praxis genutzte Techniken, die hyperspektrale Segmentierungsmethoden anwenden, zu verbessern. Zusätzlich liefert die theoretische Analyse ein solides Verständnis dieser Modelle und bietet Anhaltspunkte für weitere Verbesserungen.

Acknowledgments

I would like to express my deepest gratitude to all the people who helped and supported me getting through this challenging but also enjoyable journey of this project.

Most importantly, I owe my greatest thanks to Prof. Dr. Benjamin Berkels. He was a tremendous support; always reliable and available for discussions, sharing his extensive knowledge about mathematics and computer science and providing valuable assistance and guidance in any kind of scientific problems. He was always patient with me and allowed me to investigate my own ideas. I am deeply thankful for the confidence and trust in my abilities he has shown me. When he was asked, he also shared his opinion and knowledge about other things beside the scientific work and gave helpful advices. I really learned a lot from you.

Furthermore, I want to sincerely thank Prof. Chandrajit Bajaj, Ph.D., who contributed valuable and innovative ideas to the project and gave new and exciting perspectives on the problem. His way of thinking and drawing connections between different research fields is inspiring.

I also thank Prof. Dr. Lars Grasedyck who was so kind to be the second reviewer of my thesis and gave me the chance to talk about my work in the IGPM seminar.

In addition, I have to thank Prof. Ekaterina Kim, Ph.D., for coming up with the idea to apply our segmentation framework to Sentinel-2 data, the fruitful collaboration and the presentation of our results on a conference. The curiosity and excitement about new results were contagious. I got interesting insights in ice research and enjoyed the collaboration very much.

I am also very much in debt to my brother Kai-Hendrik Cohrs for sharing his thoughts and ideas, being always available for discussions, proofreading, and, most importantly, for being always there for me and keeping me motivated during my entire studies and life. Thank you so much.

Moreover, I have to thank my close friends Karla Pizzi and Richard Kramer with whom I fought through the Bachelor and Master studies together. I would not have finished the studies successfully without you. Thanks for sharing the same enthusiasm about maths that I have, being patient and understanding with me and also for

proofreading the thesis, Karla.

I would like to express my gratitude to Dr. Franz-Xaver Hipp, Albrecht Koschützke and Dr. Dietmar Dirmoser who listened to my concerns and were crucial factors encouraging me to take this position and start the project.

I am much indebted to my sister and her family as well as to my parents and my grandmother for being always ready to help whenever I needed it, especially, for taking care of our kids, and also for constantly encouraging me to carry on and believing in me that I can successfully finish this project.

Finally, a special and hugest thanks goes to my family. I am so grateful to my wife Leonina Koschützke for her never-ending support in any kind of problems, helping me going through this challenging time and for enduring me for several years now, and to my children Jonte and Tjark for reminding me of what matters most. There is nothing better than watching you growing up and having you around. This project was not my project alone, it was a family project.

Contents

Abstract	iii
Zusammenfassung	v
Acknowledgments	vii
Introduction	xiii
1 Chapter: Hyperspectral image segmentation	1
1.1 Hyperspectral imaging	1
1.2 Image segmentation	7
1.3 Related works	12
1.4 Notation	15
1.5 Dimensionality and noise reduction: MNF	17
1.5.1 Mathematical description	18
1.5.2 Estimation of covariances of signal and noise	20
2 Chapter: Variants of the Mumford–Shah model	23
2.1 The Mumford–Shah segmentation functional	23
2.2 The ϵ -regularized anisotropic Mumford–Shah model	33
2.2.1 Definition of indicator function	34
2.2.2 Existence of minimizers	41
2.2.3 Γ -convergence	55
2.2.4 Importance of ϵ -regularization	62
2.2.5 Lower bound for eigenvalues of covariance matrix	67
2.3 The ϕ -regularized anisotropic Mumford–Shah model	71
2.3.1 Model description and lower bound of eigenvalues	72
2.3.2 Existence of minimizers	76
2.4 The kernelized Mumford–Shah model	79
2.4.1 Model description	80

2.4.2	Kernels for hyperspectral imaging	84
3	Chapter: Optimization	89
3.1	Optimization framework	89
3.1.1	Convexification	90
3.1.2	Discretization	91
3.1.3	Optimization algorithm	93
3.1.4	Initialization	93
3.1.5	Stopping criterion	94
3.2	Optimization with respect to the labeling function	95
3.2.1	The PDHG method	96
3.2.2	Application of PDHG method to MS model	101
3.3	Optimization with respect to parameters of indicator functions	107
3.3.1	Parameters of ϵ AMS	108
3.3.2	Parameters of ϕ AMS	113
3.3.3	Parameters of kMS	126
3.4	Computational complexity of ϵ AMS	129
4	Chapter: Numerical results	131
4.1	Experimental setup	132
4.1.1	Datasets	132
4.1.2	Performance metrics	135
4.1.3	Settings	140
4.2	Results	147
4.2.1	Comparison of segmentation results	147
4.2.2	Effect of MNF	165
4.2.3	Convergence of alternating optimization	166
4.2.4	Effects of early stopping in PDHG algorithm	169
4.3	Derivation of complex sea ice states from multispectral images	170
4.3.1	Experimental setup	171
4.3.2	Results	173
	Conclusions	179
	A Mathematical foundations and results	183
	Bibliography	191
	List of Abbreviations	203
	List of Symbols	205

List of Figures	211
List of Tables	213

Introduction

Computers and algorithms are more and more used to automate tasks and simplify our everyday lives. Important resources of information in this development are images that provide rich visual representations of scenes or objects and allow computers to *see* and infer information about the image content. Hence, many technical devices are nowadays equipped with cameras to enhance their abilities. For example, images are used in industry to detect defects in components, in medical imaging to support the doctor's diagnostic investigations and in robotics to let the robot explore its surroundings.

The simplest forms of images are grayscale and RGB images that can be acquired with a normal photo camera. The former only uses different brightness values to store the information from the captured scene, while the latter extends this and represents the scene with the intensities of red, green and blue light. Hence, both base their representations on aspects that are visible to the human eye. However, there are more sophisticated imaging techniques, such as magnetic resonance imaging or electron microscopy, that allow us to even visually represent scenes and objects that are usually hidden from the human eye. Another example of such imaging techniques is hyperspectral imaging done by hyperspectral sensors that significantly extend the recorded wavelength range compared to normal photo cameras and also capture wavelengths in the invisible spectrum, resolving them with a high level of detail [59]. The highly detailed representation of the spectra shall yield different spectra for each constituent in the image that can be understood as fingerprints of the constituents. Hence, even if the human eye cannot see a difference between two materials because they have the same color and, therefore, reflect the same spectrum in the visible regime, hyperspectral sensors register differences in the reflected spectra in the invisible regime and, thus, hold necessary information to distinguish such objects. This makes hyperspectral imaging a powerful tool that provides extremely rich visual representations of scenes. Nowadays, it is applied in many different environments, such as monitoring vegetation health [90], monitoring food quality [67], sorting and recycling waste [85] or detecting diseases [68].

Having rich and useful information representations is only one aspect of an

effective exploitation of images. It is of equal importance to have powerful algorithms to successfully extract the information relevant for further processing. Such algorithms ideally also account for special properties of the considered data type.

This thesis deals with the problem of unsupervised hyperspectral image segmentation. Segmentation describes a computer vision task in which one aims to divide an image into meaningful regions to distinguish between constituents in the scene represented by the image. If a segmentation algorithm obtains only the image as input without any prior knowledge about the present constituents, the algorithm is unsupervised [93]. Hyperspectral images seem to be a suitable source of information for image segmentation due to the detailed and comprehensive data about the scene they are carrying. However, in contrast to other image types, hyperspectral images suffer from strong so-called spectral variability [59], a specific effect in this data type that significantly hampers the quality of its processing and, therefore, a segmentation algorithm has to account for. The overall goal of this thesis is to propose new approaches to the problem of unsupervised hyperspectral image segmentation that are able to handle the spectral variability and provide high-quality segmentations of hyperspectral images.

One well-established and popular model for image segmentation was proposed by Mumford and Shah in 1989 [76]. We will propose three models that are all based on the Mumford–Shah functional, but equipped with data terms that are adapted for hyperspectral images. Since we obtain segmentations from the models by finding minimizers, a question that naturally arises is if a minimizer of the corresponding functional actually exists. We will prove an existence result for minimizers for two of the models. Since the third functional has the same structure as the classical Mumford–Shah functional, for which an existence proof exists, it is straightforward to prove an existence result also for that functional following the reasoning in the existence proof for the classical Mumford–Shah functional. A proof for the existence of minimizers of the third proposed functional is, therefore, not part of this thesis. Another question that comes along with the minimization of functionals is how we can effectively find such minimizers. For two of the models, we will introduce fixed point iteration schemes to optimize them with respect to the parameters in the data term. For the data term of the third functional, we will derive a closed-form solution to compute the minimizing data term parameter choice. One of the models includes a regularization parameter needed to ensure the feasibility of the model. Since this parameter is chosen to be small, we will devote a part of the thesis to the question of what happens if this parameter tends to zero and prove the Γ -convergence of the functional to a Γ -limit.

This thesis is structured as follows: Chapter 1 introduces hyperspectral imaging and the image segmentation task. It explains the special properties of hyperspectral data and gives definitions for the relevant terms of mathematical image processing.

Furthermore, it gives a brief introduction to the applied preprocessing technique since it plays an important role in our segmentation framework. Then, we introduce the Mumford–Shah segmentation functional and the proposed variants of it in Chapter 2. In particular, we prove the existence results and the Γ -convergence. Afterward, we explain in Chapter 3 how we optimize the models, starting with a general overview of the used optimization framework and then presenting the fixed point iteration schemes as well as the applied optimization algorithms proposed by other authors. In Chapter 4, we present the results of numerical experiments performed with the models. Necessary definitions and results that are not directly linked to the topic of the thesis but are needed as mathematical background in definitions or proofs are stated in the appendix (cf. Appendix A).

Parts of the presented material have already been published: The proposed ϵ -regularized Mumford–Shah model, the optimization framework including the fixed point iteration scheme and stopping criterion as well as the numerical experiments with this model have been presented in [26]. The results of the application of that model to Sentinel-2 data for the derivation of complex sea ice states have been published in [28]. Finally, the proof for the existence of minimizers of this model and its Γ -convergence proof have been made public in [27]. The code for the three developed segmentation methods, which are described in this thesis, is publicly available at <https://github.com/berkels/msiplib>.

Chapter 1

Hyperspectral image segmentation

The overall goal of this thesis is to propose models for hyperspectral image segmentation. The term *hyperspectral image segmentation* is a composition of the two terms *hyperspectral imaging* and *image segmentation*. Therefore, to obtain an understanding of the problem statement, it is necessary to get an understanding of these two terms. This chapter is devoted to the introduction and explanation of the underlying concepts and necessary background information. We will start with introducing hyperspectral imaging in Section 1.1, where we will also give a formal definition of an image in a mathematical sense. Then, we describe image segmentation in Section 1.2. Afterward, we will review existing literature and related works in Section 1.3. It follows an overview of the mathematical notation that is used throughout the thesis in Section 1.4. The chapter ends with a description of the minimum noise fraction transform (MNF) as a dimensionality reduction technique for hyperspectral images in Section 1.5, which is an important component of our segmentation workflow.

1.1 Hyperspectral imaging

The term *hyperspectral imaging* refers to an imaging technique that yields special image types that contain much more information about the captured scene than a usual color image. The following information to explain the term mainly builds upon [59].

In general, image acquisition is done with a spectral imaging system, like digital color cameras or spectrometers. Digital color cameras are used to generate one of the most common image types, which is the *RGB image* where *RGB* stands for red (R), green (G) and blue (B). The camera captures a scene by storing the intensities of the wavelengths corresponding to these three colors in the reflected light at each spatial position in the scene. The tuple of a spatial position plus its recorded spectrum then forms a *pixel* of the image. The result of the acquisition process is a three-

dimensional vector for each spatial position in the captured scene, containing the intensities of red, green and blue light in the reflectance of that location. Hence, images contain two types of information: *spatial* information, telling us at which position in the image a pixel is located, and *spectral* information, giving us specific information about the recorded reflectance of a pixel. One refers to the spectral measurements as spectral *bands* or *channels* [93]. When acquiring an RGB image, the camera analyzes the reflectance only in the wavelength range of visible light, i.e., from approximately 380 nm to 750 nm, and takes only three intensity values from this rather large wavelength interval of the electromagnetic spectrum. Since this is merely a coarse representation of the spectrum, the *spectral resolution* of RGB images is low.

Multispectral sensors analyze a larger part of the electromagnetic spectrum than digital color cameras and sample it with a higher but often irregularly spaced spectral resolution. The result is a *multispectral image* with a number of spectral channels in the order of ten. Please note that in some contexts, an image is already called a multispectral image if it has more than one channel [93]. In this sense, RGB images are considered as multispectral images. However, we distinguish between RGB and multispectral images. An example of a multispectral sensor is the sensor the Sentinel-2 satellites are equipped with. It takes 13 measurements from the wavelength range from approximately 443 nm to 2190 nm [34].

Hyperspectral sensors enhance this even further. They consider a very large and continuous part of the analyzed spectrum and sample this with an extremely small step size, resulting in a high spectral resolution. At each pixel position, one obtains a discretized version of the underlying reflected continuous spectrum. Hence, compared to RGB and multispectral images, the only difference is that, at each pixel position, the spectrum carries much more information about the reflected. The high sampling rate and continuity of the original spectrum introduce high correlation between neighboring channels, making the information in the recorded spectra partially redundant. The resulting data is sometimes referred to as *hyperspectral data cube* because the spatial position is given by two dimensions and the spectral information adds a third one that is (almost) continuous.

Three main differences between multispectral and hyperspectral images should be pointed out: (1) the spectral resolution is much higher in the case of hyperspectral images, (2) as is the number of channels per pixel and (3) hyperspectral spectra consist of contiguous, narrow, evenly-spaced spectral bands. An example of a hyperspectral sensor is the reflective optics system imaging spectrometer (ROSIS), with a wavelength range from 430 nm to 850 nm for airborne as well as from 400 nm to 1040 nm for space-borne sensors [61], covering the visible and near-infrared (NIR) regime. Two other examples are the hyperspectral digital imagery collection experiment (HYDICE) sensor and airborne visible/infrared imaging spectrometer (AVIRIS), both recording wavelengths from 400 nm to 2500 nm but with different spectral resolutions [94, 47].

Hence, they cover the visible, NIR and mid-infrared (MIR) regime. These three examples are from the field of remote sensing, where nowadays, hyperspectral imaging is often applied. However, there are also other types of hyperspectral images like Fourier-transform infrared spectroscopy (FT-IR) images, where the infrared regime is measured at each pixel position [49], or electron energy-loss spectroscopy (EELS) images, that do not analyze the electromagnetic spectrum but the energy distribution of electrons after these had contact with a specimen [36]. Therefore, it should be stressed that hyperspectral imaging is not restricted to an analysis of the electromagnetic spectrum. Fig. 1.1 visualizes for exemplary RGB, multispectral and hyperspectral sensors the parts of the electromagnetic spectrum that these sensors analyze and at which wavelengths of the spectrum the sensors take measurements. The first

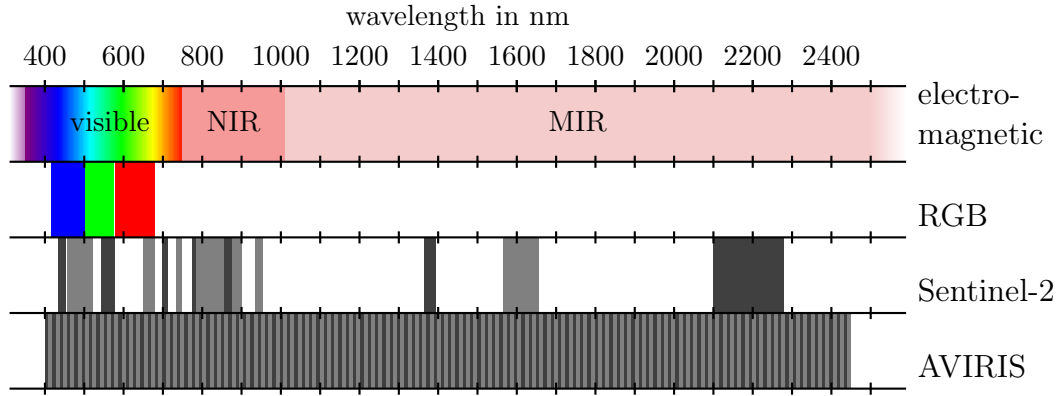


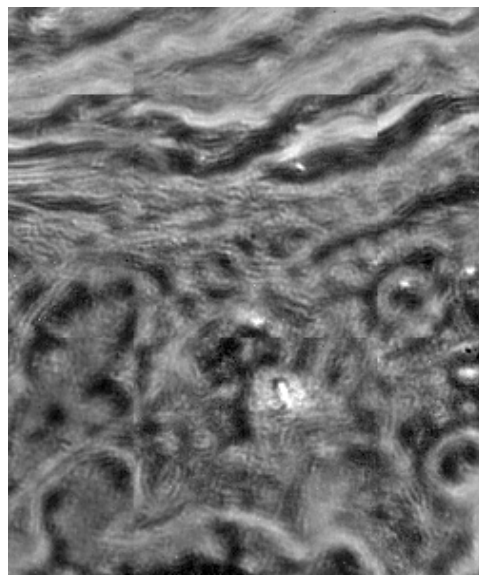
Figure 1.1: Visualization of spectral bands used by a typical RGB camera, the Sentinel-2 multispectral sensor and the AVIRIS hyperspectral sensor.

row shows the part of the electromagnetic spectrum from approximately 350 nm to 2500 nm, from left to right starting with the visible light, then NIR and MIR regime. The boxes in the other rows mark the wavelength ranges that are recorded at each position in the image to yield the intensity values for the channels of the pixels. The second row shows the three spectral bands of an RGB camera [16]. In the third row, the spectral bands of the Sentinel-2 multispectral sensor are shown [34]. The color coding for this as well as the last row is only used for better visibility of the different bands and differentiation between neighboring bands, the colors do not have any further meaning. The reason for the uneven distribution of the bands is that the bands are chosen to serve special purposes. For example, the band at 940 nm is designed for the correction of water vapor effects and the band at 1375 nm for the detection of cirrus clouds [34]. Please note that two bands (called band 8 and 8b at about 800 nm to 900 nm) in the NIR range overlap but are recorded at different

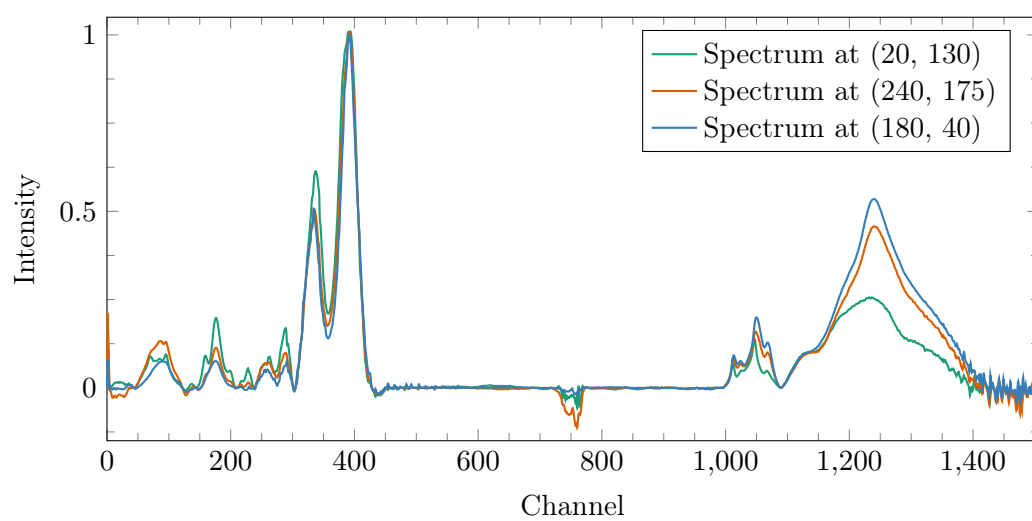
spatial resolutions. The spectral bands of an AVIRIS sensor are shown in the last row [93].

Practical applications of hyperspectral imaging are, e.g., surveillance, mineralogy, environmental management and agriculture in the field of remote sensing [86], but also biomedical imaging [49] and in material science to investigate atoms and molecules [36]. Fig. 1.2 shows an FT-IR image of human tissue and the spectra of three of its pixels. Each pixel has 1506 channels carrying information about the reflectance of the tissue at that position. In Fig. 1.2a, we see the image after compressing the spectral information to one value per pixel by computing the pixel-wise mean over all available channels. The plots of the spectra in Fig. 1.2b have the channels on the x -axis and the intensities on the y -axis. One can observe the correlation of neighboring channels, meaning that the intensity value of the i -th channel is dependent on the intensities in the channels $i - 1$ and $i + 1$ and the other way around. For instance, if the intensity in channel i is high, then the intensities of the channels $i - 1$ and $i + 1$ also attain values that are close to the intensity of channel i and cannot jump to a low value.

Hyperspectral imaging is motivated by the observation that, under laboratory conditions, pure materials reflect unique spectra, so-called *spectral signatures*, which can be understood as their fingerprints, making it possible to distinguish between materials only based on their reflectance. Since hyperspectral sensors make use of the full spectral information that is available in a certain spectral range, this finding has led to the assumption that the recorded spectra of hyperspectral images carry enough information to distinguish between the present constituents in a specimen or scene only based on their reflectances. In a practical environment, however, this is not true. Different effects like noise, the angle of illumination and the relative viewing angle with respect to the surface let the reflected spectra vary and introduce strong so-called *spectral variability*, which is characteristic for hyperspectral images. It describes that spectra, even reflected by the same material or constituent of the specimen or scene, vary strongly with different strengths in the different channels around the spectral signature of that material or constituent. Depending on the imaging modality, there can be more specific influences that strengthen the variability. For example, in remote sensing also atmospheric effects like scattering have to be taken into account. Another source of variability can be a low spatial resolution. If a pixel covers a large area of the specimen or scene, it can happen that more than one constituent contributes to the reflectance of that pixel resulting in mixed pixels. The discipline of *hyperspectral unmixing* deals with this effect. Its goal is to separate the spectra into the spectral signatures (called *endmembers*) of the contributing constituents and the percentages (called *abundances*), telling how much each constituent contributed to the reflectances of the pixels [11]. The spectral variability is visualized in Fig. 1.3. To create the figure, the spectra of the pixels in the FT-IR image corresponding to one tissue type in Fig. 1.2 were projected onto the first two principal components using the principal



(a) Pixel-wise mean over the channels.



(b) Spectra of three pixels.

Figure 1.2: Exemplary FT-IR image; data courtesy of Rohit Bhargava (University of Illinois Urbana Champaign).

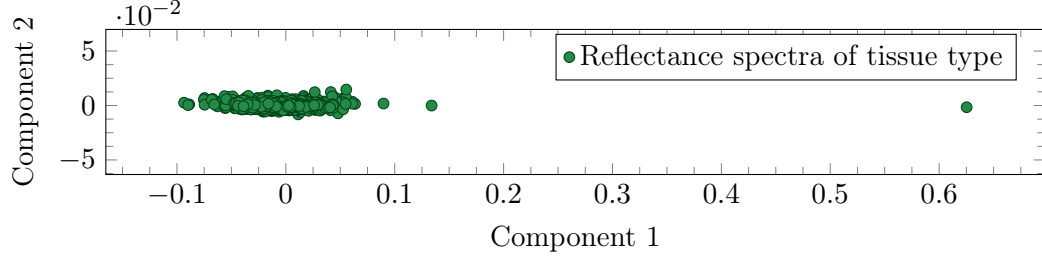


Figure 1.3: Distribution of spectra from one tissue type in the FT-IR image; data courtesy of Rohit Bhargava (University of Illinois Urbana Champaign).

components analysis (PCA) [88] by discarding the spatial positions of the pixels and considering the corresponding spectra as a set of vectors to which the PCA is then applied. The information which pixels belong to this specific tissue type was given by expert knowledge. One can see in the figure that the spectra show a strong variation in the direction of the first component, while in the direction of the second component, there is merely a weak variation. Since the PCA gives directions that are ordered by the amount of variance in the data they explain, the variation in the remaining directions is even weaker. Although the figure suggests that the spectral distribution of the spectra could obey a normal distribution, this is not the case in general. The spectral distribution of the constituents in an image is unknown and can follow arbitrary distributions. Spectral variability is sometimes also referred to as *intra-class variability* [86].

The spectral variability is one of the main characteristic challenges that come along with hyperspectral data, a second challenge poses noise in the data, and the third is the correlation between neighboring channels, which introduces redundancy in the data and, therefore, leads to a high dimensionality of the data without adding more information. The high dimensionality of the data poses further challenges as computations are expensive and visual inspection of the data is difficult. Methods applied to hyperspectral data have to take these challenges into account and be able to deal with them. The problems of high dimensionality and redundancy in the data are usually tackled with dimensionality reduction techniques. As mentioned in the introduction of the chapter, the technique that is used in this thesis to reduce the dimensionality of and the noise in the data is the MNF that we describe in Section 1.5. The techniques developed and explained in this thesis mainly tackle the spectral variability.

1.2 Image segmentation

Image processing is the discipline that deals with the automated processing of images. For general information about image processing, we refer the interested reader to [15], which also serves as a reference for the information given in this section.

Image processing comprises several basic tasks, including denoising, deblurring, segmentation, registration or inpainting. Denoising aims to correct erroneous information (the noise) introduced during the data acquisition process, as noise can have negative effects on the quality of further image analysis. It is crucial for denoising techniques that they preserve the true underlying information and structure of the image. Deblurring is concerned with the removal of blurring effects in an image. Registration is about aligning several representations of an object given by images with each other. The images showing the object can be acquired with different imaging modalities. Inpainting has the goal to meaningfully reconstruct or fill corrupted parts of an image. In this thesis, we focus on the segmentation task. Image segmentation aims to partition an image into meaningful regions or objects. The notion of the term *meaningful* depends on the setting the segmentation is applied to, i.e., what type of image is processed and which information one wants to obtain from the data. The simplest case is the separation of an image into two segments, e.g., foreground and background, referred to as binary segmentation. In the more general case, that we consider in this thesis, more than two segments are sought. This is called *multiphase segmentation*. An image and an exemplary segmentation of it are shown in Figs. 1.4a and 1.4b.

To introduce the segmentation problem in a mathematical setting, we first need to define an image in a mathematical sense. Let $d, L \in \mathbb{N}$ and let $\Omega \subseteq \mathbb{R}^d$ be a domain, i.e., non-empty, open and connected, that is additionally bounded (cf. Definition A.8), and $\mathcal{N} \subseteq \mathbb{R}^d$ a d -dimensional finite regular grid, i.e., a product of finite equidistant one-dimensional grids. For theoretical analysis of an image, we define it in a continuous setting:

Definition 1.1 (Mathematical image).

An image is a function g that maps from a domain Ω , called image domain, to a value range $W \subseteq \mathbb{R}^L$, i.e.,

$$g: \Omega \rightarrow W.$$

When we want to work with images in practice, we need to discretize them first, which leads us to the notion of a digital image:

Definition 1.2 (Digital image).

A digital image is a function g that maps from a finite regular grid $\mathcal{N}$, the image

domain, to a value range $W \subseteq \mathbb{R}^L$, i.e.,

$$g: \mathcal{N} \rightarrow W.$$

Hence, an image assigns a vector in the value range to each position in the image domain. The assigned vector contains the information we have about that location in the image. Relating it to the examples of images given in Section 1.1, this vector carries the spectral information. In particular, L describes the number of channels of the image. Images with $L = 1$ have only one brightness value per pixel and are called *grayscale images*. For RGB images, we have $L = 3$. For general multispectral and hyperspectral images, L depends on the sensor and its spectral resolution. As mentioned above, L is in the order of ten for multispectral images. In the case of hyperspectral images, we assume $L \gg 1$. The FT-IR image shown in Fig. 1.2 has $L = 1506$, an AVIRIS image $L = 224$ [47]. The image domain in the continuous and discrete case is often chosen as $\Omega = (0, 1)^d$ and $\mathcal{N} = \{1, \dots, N_1\} \times \dots \times \{1, \dots, N_d\}$, respectively, where $N_1, \dots, N_d \in \mathbb{N}$. For most of the image types, we have $d = 2$, like RGB images, but also images with $d = 3$ are used in practice, like in magnetic resonance imaging (MRI) [13]. For the theoretical analysis in Chapter 2, we will consider general d -dimensional image domains; however, the images considered in the numerical experiments in Chapter 4 are two-dimensional, hence we can restrict ourselves to $d = 2$ there. The *pixel* in a mathematical sense is then given as the tuple $(x, g(x)) \in \Omega \times W$ or $((i_1, \dots, i_d), g(i_1, \dots, i_d)) \in \mathcal{N} \times W$ of a location in the image domain and its associated spectral information. Please note that in the literature, this tuple is called *pixel* for $d = 2$ and *voxel* for $d = 3$. However, in this thesis, we consider images with domains of arbitrary finite dimension and will consistently refer to the tuple of spatial and spectral information as *pixel*, regardless of the value of d .

In the following, we will only consider the continuous case for ease of notation when giving a mathematical description of segmentation. However, it also applies to the discrete case. Let $k \in \mathbb{N}$ be the number of segments that we assume to be fixed. Mathematically, the segmentation task aims to find a partition $(\mathcal{O}_1, \dots, \mathcal{O}_k)$ of Ω into k subsets, i.e., $\mathcal{O}_1, \dots, \mathcal{O}_k \subseteq \Omega$ are pairwise disjoint sets and

$$\bigcup_{l=1}^k \mathcal{O}_l = \Omega,$$

such that the pixels in each $\mathcal{O}_l$ share certain characteristics. The sets $\mathcal{O}_1, \dots, \mathcal{O}_k$ are the segments of the image. In particular, this means that any pixel is always assigned to exactly one segment. Please note that the given partition condition is slightly stronger than what is usually required, namely the equality of the closure of Ω and

the union of the closures of the $\mathcal{O}_1, \dots, \mathcal{O}_k$:

$$\bigcup_{l=1}^k \overline{\mathcal{O}_l} = \overline{\Omega}.$$

However, our condition on the sets $\mathcal{O}_1, \dots, \mathcal{O}_k$ is weaker since the sets are usually required to be open. The slight deviation from the usual conditions is needed for the proofs of the existence of minimizers for two of our models in Chapter 2. In the proofs, we work with characteristic functions of the subsets of Ω instead of the subsets themselves and the former condition implies that the characteristic functions of all subsets sum up to 1 at each $x \in \Omega$, which is not the case for the latter condition that involves the closures. Segmenting an image means grouping the pixels by finding a partition of Ω with respect to a suitable notion of homogeneity. It is a different way of saying what we described at the beginning of the section: We are looking for *meaningful* objects in an image. This meaningfulness has now been specified by the notion of homogeneity of the pixels. Which notion of homogeneity is appropriate mainly depends on the application. For example, if one can assume structure in the image, like in images of crystal lattices or a certain shape of objects to be detected, this can be taken into account when defining the sense of homogeneity. The homogeneity can be quantified in different ways, e.g., with a similarity measure that computes the difference between spectra or by comparing the spectra with some representative like a mean spectrum of a segment. As described above, we assume for hyperspectral images that constituents reflect unique spectral signatures. Therefore, a natural idea is to use only the spectral information of the pixels and discard the spatial information when quantifying the homogeneity, potentially after some preprocessing. In two of our methods that we propose in this thesis, we will go this way and quantify the homogeneity of pixels only based on their spectra (cf. Chapter 2). However, since we are dealing with images, there is more information that can be exploited. The image domain contains the spatial information of an image and allows for a definition of spatial neighborhood relations of pixels. Therefore, one could also quantify the homogeneity of pixels based on the pixel plus its neighborhood. We will do this in the third model that we propose in this thesis. A meaningful assumption that one can make and incorporate into a segmentation model is that neighboring pixels have a high probability to belong to the same segment.

A partition of the image domain is a grouping of spatial positions, which implies a grouping of the pixels, which in turn implies a grouping of the associated spectra, and the other way around. Hence, we can go back and forth between the segments and the corresponding groupings of the spectra (called *spectral clusters*) as these are connected. In Fig. 1.4, we see an example of an RGB image, a possible segmentation of that image into three segments and the corresponding spectral clusters that are given by the segmentation. Also in Fig. 1.4c, the spectra were projected onto the

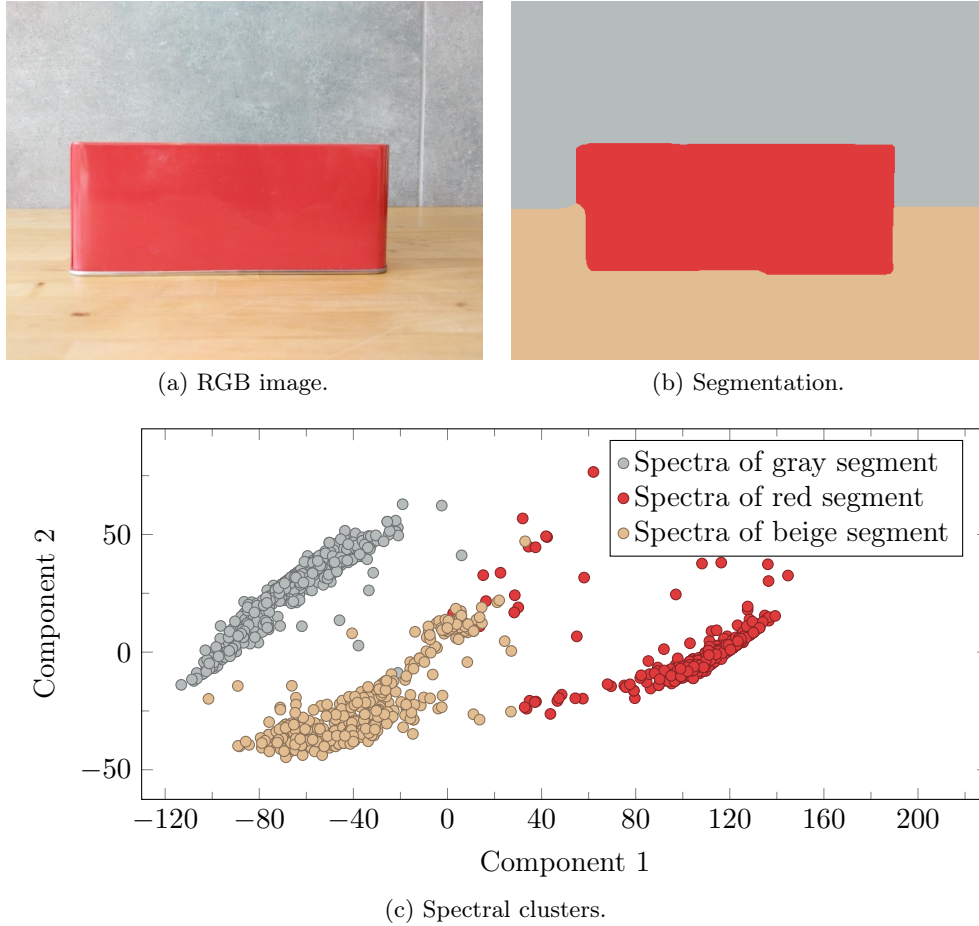


Figure 1.4: An exemplary RGB image showing a red box on a wooden surface with a gray background, a segmentation of it into three segments and the corresponding spectral clusters.

first two principal components of all spectra given by the PCA for better visibility. One can see that even in the case of RGB images, spectral variability due to different illumination conditions and noise plays a role. Nevertheless, the spectral distributions are quite well separated. Some of the red spectra are close to the beige cluster. Hence, from a spectral perspective, they should be classified as beige segment. However, spatial information can cause them to be put into the red segment.

An image does not have a unique segmentation. Depending on the application and, therefore, the chosen notion of homogeneity of the pixels, different segmentations are possible. Moreover, segments do not have to be connected regions and can be

separated in general. We have fixed the number of segments k to a natural number and assume that this number is given as prior knowledge. There are approaches that try to estimate the number of segments in an image. However, since the number of segments that are sought can also depend on the application, we will always assume that k is a priori given.

Image segmentation is closely related to the task of pixel classification. In [79], Nalepa et al. describe segmentation as the task of finding boundaries of objects in an image, while classification is described as an assignment of a class label to each pixel. They emphasize that a segmentation also implies a classification of the pixels. The remote sensing community refers to a pixel-wise classification of a hyperspectral image as *hyperspectral image classification* [42]. However, one has to be careful with the terminology here since in other communities, the term *image classification* usually refers to the task of assigning one class label from a set of given classes to an entire image [92].

Furthermore, one can either tackle the segmentation and classification tasks *supervised* or *unsupervised*. The following information about supervised and unsupervised methods builds upon [93]. In a supervised setting, so-called *training data* is given as prior knowledge. This data consists of samples plus segment or class labels for that samples that are considered to be the true labels. The samples and their labels are fed into a segmentation or classification method. The model parameters are then adjusted in such a way that they ideally output for each sample in the training data the corresponding true label. Once the model parameters have been adjusted based on the training data, new data samples can be input into the model and are then segmented or classified with these adjusted parameters. An unsupervised setting is a situation in which no prior knowledge about the target segment or class labels of any data samples is given. All parameters of the model are estimated from the (unlabeled) data that one wants to segment. Both settings have advantages and disadvantages. The generation of labeled training data for supervised classification and segmentation methods is very time-consuming. Usually, an expert is needed who manually labels the pixels for training, which is referred to as *a priori* labeling. Furthermore, supervised methods need a significant amount of training time, the time to adjust the model parameters based on the training data, to be ready for classification. Once the model is trained and the parameters adjusted, however, the classification of new data is usually fast. An additional drawback is that a classifier that is trained on an image cannot be expected to generalize to other images in other environments or applications. The assigned class labels in the supervised setting, however, carry semantic information about the class. In the case of unsupervised methods, no a priori labeling of training data is needed, which saves time. However, as opposed to supervised methods, an unsupervised segmentation or classification gives only symbolic labels to the samples, and these labels are unrelated to the meaning of

the shared characteristics of the samples in the segments. After the segmentation is obtained, an expert may assign semantic labels to the segments if the meanings are needed. This is called *a posteriori* labeling. It is much cheaper than in the supervised case because the analyst only needs to label the segments of interest and can do this with the help of the shapes of the found segments. Additionally, merely the segment obtains the label which then applies to all the pixels included in the segment. Unsupervised methods can also be applied directly in different environments and applications. The time needed for the segmentation of an image or classification of the pixels, however, is rather long compared to the time supervised methods need for classification since it involves fitting all model parameters which is usually done by the optimization of an objective function. The methods developed in this thesis are unsupervised segmentation methods. Since segmentation and (pixel-wise) classification are so closely related and the segments can be considered as classes, we will use the terms *segment* and *class* interchangeably.

In this thesis, we will tackle the hyperspectral segmentation problem with unsupervised approaches. We will introduce and analyze variants of the Mumford–Shah (MS) segmentation functional for that, which are variational approaches to the problem. In particular, we will propose three new data terms for the functional. Hence, we obtain segmentations of images by using these variants of the MS functional as the objective function and finding minimizers of it. For two variants, we will also prove the existence of minimizers.

1.3 Related works

The field of hyperspectral image classification and segmentation is an active field, in particular, in the remote sensing community. In principle, also general purpose image segmentation methods can be applied for hyperspectral image segmentation. However, worse results than the ones from special methods for hyperspectral data should be expected since the general methods do not take into account the special properties of hyperspectral data. For this reason, we limit the following literature review of segmentation and classification methods to models that are explicitly designed for hyperspectral images.

One can essentially divide the approaches into two categories: spectral approaches, which rely only on the spectral information of an image to obtain the segmentation or classification, and spectral-spatial approaches, which combine spectral and spatial information to enhance the results. Many authors try to solve the task with supervised methods. Because of the shortage of labeled training data for hyperspectral image segmentation in many fields, authors often use a part of the pixels of an image that are labeled as training data in order to classify the remaining pixels of the very same image with the trained classifier [79]. This, however, is quite an impractical approach

because we cannot expect a classifier that is trained on some pixels of an image to generalize to other images. The research on supervised methods mainly focuses on introducing models that achieve high classification accuracies while reducing the number of necessary training samples to make the labeling process cheaper. However, the *Hughes phenomenon* [56] introduces a severe difficulty when performing supervised hyperspectral classification. It describes the behavior that with increasing dimensionality of the data, i.e., with increasing number of spectral channels, the classification accuracy drops because the larger the dimensionality, the more labeled samples are needed to describe the data distribution and, consequently, to reliably estimate the model parameters of the supervised classifier [93]. Since obtaining labeled training data is remarkably expensive in terms of time and labor and there is very little labeled data available, people try to investigate methods to mitigate that effect, e.g., by also incorporating unlabeled data samples in the model parameter estimation [100], exploring dimensionality reduction techniques [2] or selecting the most descriptive bands [102].

The majority of articles about hyperspectral image segmentation and classification deals with supervised methods. Supervised methods that use only spectral information are usually variants of the support vector machine (SVM) [41, 24], but also works based on neural networks exist [74]. More recently, supervised spectral-spatial approaches have become more common. Many of them leverage neural networks as classifiers [97, 98, 109, 101].

Much fewer works concentrate on the challenging task of unsupervised hyperspectral image segmentation and classification. One class of such methods aims to determine adaptive neighborhoods of pixels, named *superpixels*, which can help to enhance the results when solving other analysis tasks on the data [43, 75]. Other approaches to the segmentation task not only try to find the segmentation of the image but also to estimate the number of segments in the image [38, 104, 20, 7]. Since general clustering techniques can be directly applied to cluster the spectra of hyperspectral images and obtain segmentations, models like, e.g., k -means, Gaussian mixture model (GMM) or Bayesian Gaussian mixture model (BGM) can be counted towards this category. However, it also includes more sophisticated models that are designed for hyperspectral data. Spectral clustering in the sense of a spectral analysis of the similarity or adjacency matrix of a graph is adopted in several approaches: in [53] on a bipartite graph representation where a radial basis function kernel is used to compute the data affinity matrix, in [105, 115] on a constructed anchor-based graph, and in [111] on the adjacency matrix obtained from a spectral-spatial sparse subspace clustering. Hassanzadeh et al. [54] apply spectral clustering in their developed multi-manifold learning framework that uses contractive autoencoders. In [1], Acito et al. introduce a segmentation method that projects the spectra of the pixels onto a set of significant directions, fits a GMM to each of these projections and fuses the

segmentation results afterwards. Murphy and Maggioni [77] apply density estimation and geometry learning with a diffusion distance to detect the modes of the spectral distributions of the segments. A manifold matrix factorization that simultaneously performs dimensionality reduction and unsupervised classification is proposed in [112]. Nalepa et al. [78] introduce a 3D convolutional autoencoder that also performs dimensionality reduction and clustering simultaneously by mapping the spectra into a latent space defined by the bottleneck layer and disentangling them there. The approach by Obeid et al. [83] is a modification of the deep embedded clustering model [108], extending it to deal with hyperspectral images and data imbalance. The work of Bench et al. goes in a similar direction as they also apply deep convolutional autoencoders to perform end-to-end dimensionality reduction and clustering; however, their approach is a spectral-spatial one, considering patches of pixels [9]. Liu et al. investigate the application of a hierarchical clustering algorithm to hyperspectral image segmentation [66]. One work closely related to the models proposed in this thesis is the one by Zhu et al. [116]. The authors also target unsupervised segmentation of hyperspectral images and develop a modification of the MS model with a novel nonlocal total variation (TV) semi-norm. Two data terms described in this thesis use the Mahalanobis distance [70], which measures the distance of a point to the mean of a distribution while accounting for the strengths of variation of the distribution (standard deviations) in the different main directions of variation. This distance has already been applied in the hyperspectral imaging context, e.g., for target [57] and anomaly detection [114], for citrus damage detection [44], supervised reduction of the spectral dimensionality [65] and supervised classification [64]. These methods have in common that they apply the squared Mahalanobis distance, which is opposed to our data term that uses the non-squared version. Moreover, the mentioned methods adjust the model parameters based on labeled training data, while we estimate all parameters of our methods only from the data without any labeled training data.

The MS functional [76] is applied in several image processing contexts, e.g., for inpainting [37, 51], image restoration [6, 58] or deblurring [52]. Moreover, it is particularly popular in image segmentation. The following literature review of variants of the MS segmentation functional and existence results for minimizers of these variants is not limited to hyperspectral imaging. The original formulation of the functional comprises a squared distance as the data fidelity term plus two regularizing terms and aims to find piecewise smooth approximations to the input image. Moreover, there exists a variant that has stronger requirements on the approximations as it merely allows piecewise constant functions for that. A detailed description of the functional will be given in Section 2.1. Different data terms are proposed for the MS functional. Tsai et al. [103] introduce a weighting factor in the data term that depends on the position in the image and allows giving a low or no weight to corrupted or noisy pixels. In [91], Ramlau and Ring use the piecewise constant MS functional with a

squared L^2 -distance of the Radon transform of an unknown density function and the data for X-ray tomography. Additionally, for some variants of the MS functional the existence of minimizers has been proven. David shows that for its original formulation, the MS functional admits a unique minimizer if the set of segment boundaries is fixed, using a convexity argument [32, Prop. 3]. De Giorgi et al. [33, Th. 1.1] for the original and Bar et al. [6, Th. 2] for a slight modification show the existence of minimizers if the set of discontinuities of the approximating function replaces the set of segment boundaries in the functional. Rondi and Santosa introduce a data term for the MS functional to apply it in electrical impedance tomography, where one takes measurements at the domain boundary to reconstruct information about the conductivity distribution inside the domain. The data term consists of the difference of the approximating function after the application of an operator and the data, measured in the norm of the space of bounded, linear operators that have the space of traces of $L^2(\Omega)$ on $\partial\Omega$ with zero mean as domain and co-domain [96]. They show its performance for electrical impedance tomography and prove existence results for a variant of the MS model with a general data term that satisfies a continuity condition over two function classes. The MS variant that is introduced by Fornasier et al. [39] has a similar structure as the one of Rondi and Santosa: Here, a linear, non-invertible operator is applied to the approximation and compared to the data in the squared L^2 -norm. Furthermore, they prove two results for different classes of linear operators that guarantee the existence of minimizers but also show counterexamples where minimizers do not exist in general. Cai et al. [18] propose a modification of the MS functional with a data term being the squared difference and including a blurring operator; the regularization of the discontinuity set, however, is omitted. Moreover, they prove the existence of a unique minimizer. This model is later extended to deal with corrupted images that are affected by Poison noise and a result that a unique minimizer exists is proven [17]. For the MS functional restricted to piecewise constant functions, Mumford and Shah show the existence of minimizers when only continuous input images are allowed [76, Th. 5.1].

1.4 Notation

In this section, we will introduce symbols and notation used throughout this thesis. While most follow usual conventions, we list them here to prevent ambiguity or confusion. Any deviations from these conventions will be explicitly indicated.

Please note that the symbols d , k , L , Ω , $\mathcal{N}$, g and W , we introduce in the following, have already been used for the introduction of images in Section 1.2 and are therefore strongly related to the imaging context throughout this thesis. This is intended and the symbols keep these meanings, i.e., d , k and L are the dimension of the image domain, the number of segments and the number of channels of the

image, respectively, Ω is the image domain in the continuous, $\mathcal{N}$ the domain in the discrete setting, g the image with value range W . Nevertheless, we want to point out that they can also appear in a more general mathematical context in which they only have their mathematical meaning. For instance, Ω is used in the definition of the Sobolev and L^p spaces where it is only considered as a bounded domain with Lipschitz boundary that is the domain of the functions contained in the spaces.

We denote the natural numbers with $\mathbb{N}$, starting at 1. The symbol $\mathbb{N}_0$ stands for the natural numbers including 0, i.e., $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. The real numbers are denoted by $\mathbb{R}$, the positive real numbers by $\mathbb{R}_{>0} := \{x \in \mathbb{R} \mid x > 0\}$ and the extended real numbers by $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. We only consider real-valued vector spaces.

The symbols d , k and L are assumed to be natural numbers, hence $d, k, L \in \mathbb{N}$. The set W is always supposed to be a subset of $\mathbb{R}^L$. The boundary of a set $A \subseteq \mathbb{R}^d$ is described by ∂A . When dealing with the measurability of sets, it is always meant with respect to the Borel σ -algebra. If $A \subseteq \mathbb{R}^d$ is measurable, then $|A|$ denotes the Lebesgue measure of A . The $(d-1)$ -dimensional Hausdorff measure is described by $\mathcal{H}^{d-1}$. For the characteristic function of a set $A \subseteq \mathbb{R}^d$, we use the notation

$$\chi_A: \mathbb{R}^d \rightarrow \{0, 1\}, \quad x \mapsto \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases} \quad (1.1)$$

We further suppose $N_1, \dots, N_d \in \mathbb{N}$. Moreover, we let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with Lipschitz boundary (cf. Definition A.8) and $\mathcal{N} := \{1, \dots, N_1\} \times \dots \times \{1, \dots, N_d\}$.

We denote the p -norm for $1 \leq p \leq \infty$ in finite-dimensional spaces with $\|\cdot\|_p$ and the Frobenius norm with $\|\cdot\|_F$. The space of m -times continuously differentiable functions for $m \in \mathbb{N}$ mapping from Ω to $\mathbb{R}^n$, $n \in \mathbb{N}$, is described by $C^m(\Omega, \mathbb{R}^n)$ and the subspace of $C^m(\Omega, \mathbb{R}^n)$ containing functions with compact support by $C_c^m(\Omega, \mathbb{R}^n)$ (cf. Definition A.9). In the case where $n = 1$, we simplify these notations to $C^m(\Omega)$ and $C_c^m(\Omega)$. We denote the Lebesgue spaces or L^p spaces of functions mapping from Ω to $\mathbb{R}^n$, $n \in \mathbb{N}$, with $L^p(\Omega, \mathbb{R}^n)$ for $1 \leq p \leq \infty$ and its norm with $\|\cdot\|_{L^p(\Omega, \mathbb{R}^n)}$ (cf. Definition A.10). If $n = 1$, we write $L^p(\Omega)$. The Sobolev space for $m \in \mathbb{N}$ and $1 \leq p \leq \infty$ of functions mapping from Ω to $\mathbb{R}$ and its norm are described by $H^{m,p}(\Omega)$ and $\|\cdot\|_{H^{m,p}(\Omega)}$, respectively, (cf. Definition A.11). We always assume that g is a function with $g \in L^\infty(\Omega, \mathbb{R}^L)$.

The identity matrix is denoted by I , its dimension is clear from the context. The transpose of a vector or matrix is indicated with a superscript T . For two vectors $v, w \in \mathbb{R}^d$, the Euclidean inner product is written as $v^T w$. The orthogonal group in dimension $n \in \mathbb{N}$ is denoted with $O(n)$. Let $e_i := (\delta_{1i}, \dots, \delta_{ni})^T \in \mathbb{R}^n$ be the i -th canonical unit vector, $n \in \mathbb{N}$ and $i \in \{1, \dots, n\}$, where δ_{jl} for $j, l \in \{1, \dots, n\}$ is the Kronecker delta. The trace of a matrix M is denoted by $\text{tr } M$. The (open) ball with radius $r > 0$ around $x \in \mathbb{R}^d$ with respect to the norm $\|\cdot\|$ is expressed by $B_r^{\|\cdot\|}(x)$. We

use the symbols $\rightarrow$ for strong convergence, $\rightharpoonup$ for weak convergence and $\xrightarrow{*}$ for weak* convergence.

We denote the partial derivative of a differentiable function f with respect to the variable x_i with $\frac{\partial f}{\partial x_i}$ or $\partial_{x_i} f$ and the gradient with ∇f . The operator div describes the divergence of a function, defined for a function $\varphi \in C^1(\Omega, \mathbb{R}^d)$ as

$$\operatorname{div} \varphi = \sum_{i=1}^d \partial_{x_i} \varphi_i. \quad (1.2)$$

In the case of a function $f: \mathbb{R} \rightarrow \mathbb{R}$, we denote the first and second derivative with f' and f'' , respectively. The restriction of a function f to a subset A of its domain is denoted by $f|_A$. If the arguments of a function are finite-dimensional, we use round brackets, e.g., $f(x)$ where x is from a finite-dimensional space, if they are infinite-dimensional, we use rectangular brackets, e.g., $E[u]$ where u is contained in an infinite-dimensional space.

1.5 Dimensionality and noise reduction: MNF

An important part of our segmentation workflow is a preprocessing by the minimum noise fraction transform. Therefore, an explanation of the transformation is given in the following. The information about the MNF is following [46, 63, 10].

The MNF was introduced in [46]. While originally proposed for multispectral data, it is now also a popular tool in hyperspectral image processing. The MNF is closely related to the PCA [88]. The PCA is a linear transformation that, given a set of vectors as input data, returns an orthonormal basis for the space in which the data resides such that the basis vectors describe the directions in which the data shows the highest variance. The basis vectors are called principal components (PCs) and are ordered such that the variance they account for decreases. To each PC, the PCA also returns a non-negative scalar that quantifies the variance of the data in the corresponding direction. The PCs and variances are obtained by computing an eigendecomposition of the covariance matrix of the data, which is estimated in practice as the empirical covariance (cf. Eq. (3.36)). By projecting the data onto the subspace spanned by a few PCs that account for most of the variance, a dimensionality reduction is achieved while keeping most of the information in the data. However, the PCs are not ordered by their described noise. Hence, it can happen that PCs that are less noisy are discarded, while ones containing much noise are kept. Another drawback is that the result of the PCA is sensitive to the scaling of the data. For instance, if the range of the intensities in one channel is doubled, the variance in this channel increases by a factor of four, giving that channel more weight when it comes to the computation of the PCs. The doubling of the intensities could stem

from a noisier signal and the PCA would react on that by fitting this higher noise content. However, we would want a transformation to ideally lead to consistent results, regardless of the scaling in the data. These problems are tackled by the MNF. The method is also a linear transformation of the data but returns a basis for the data space such that the basis vectors describe the directions with the highest signal quality, measured in terms of the signal-to-noise ratio (SNR), and are ordered such that the signal quality decreases. The SNR is defined as signal variance divided by noise variance below in (1.4). Sorting the basis vectors such that their SNRs decrease is equivalent to sorting them such that the noise fractions, given as the fraction of noise variance divided by total variance [50], of the basis vectors increase. This fact gives the transformation the name *minimum noise fraction*. In particular, both measures SNR and noise fraction lead to the same basis vectors. Please note that the basis vectors are sometimes ordered in reverse order, i.e., the noise fraction decreases and SNR increases. In that case, the transformation is called maximum noise fraction. Moreover, the MNF implies a natural scaling of the data that makes the noise range equal and uncorrelated across all channels, as we will see in the following section.

1.5.1 Mathematical description

For a mathematical description of the MNF, we consider the discrete setting, i.e., the image domain is $\mathcal{N}$. We start with a noisy image

$$\tilde{g}: \mathcal{N} \rightarrow \tilde{W},$$

where $\tilde{W} \subseteq \mathbb{R}^{\tilde{L}}$ is the value range with $\tilde{L} \in \mathbb{N}$ channels. The MNF assumes that the noisy image equals the sum of a signal $s: \mathcal{N} \rightarrow \mathbb{R}^{\tilde{L}}$ and noise $n: \mathcal{N} \rightarrow \mathbb{R}^{\tilde{L}}$, i.e.,

$$\tilde{g} = s + n.$$

Let $\Sigma_T \in \mathbb{R}^{\tilde{L} \times \tilde{L}}$, $\Sigma_S \in \mathbb{R}^{\tilde{L} \times \tilde{L}}$ and $\Sigma_N \in \mathbb{R}^{\tilde{L} \times \tilde{L}}$ be the covariance matrices of the spectra of the image $\tilde{g}$ (the index T in Σ_T stands for *total covariance*), the signal s and the noise n , respectively. Furthermore, it is usually assumed that there is no correlation between signal and noise. That implies

$$\Sigma_T = \Sigma_S + \Sigma_N. \quad (1.3)$$

The MNF determines uncorrelated vectors $v_1, \dots, v_{\tilde{L}} \in \mathbb{R}^{\tilde{L}}$ such that the SNRs, given for the i -th vector for $i \in \{1, \dots, \tilde{L}\}$ by

$$\text{SNR}_i = \frac{v_i^T \Sigma_S v_i}{v_i^T \Sigma_N v_i}, \quad (1.4)$$

are successively maximized. Obviously, a scaling of v_i does not affect SNR_i . Therefore, the vectors are usually normed by

$$v_i^T \Sigma_N v_i = 1 \quad (1.5)$$

for all $i \in \{1, \dots, \tilde{L}\}$, meaning that we set the noise variance in each MNF channel to 1. Moreover, we ensure that there is no correlation between successive MNF components by imposing

$$v_j^T \Sigma_N v_i = 0 \quad (1.6)$$

for all $i \in \{1, \dots, \tilde{L} - 1\}$ and $j \in \{i + 1, \dots, \tilde{L}\}$. The successive maximization of the SNRs in (1.4) over the MNF components subject to the constraints in Eqs. (1.5) and (1.6) is the same as a simultaneous diagonalization of Σ_S and Σ_N [10]. That is, we are seeking an invertible matrix $V \in \mathbb{R}^{\tilde{L} \times \tilde{L}}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{\tilde{L} \times \tilde{L}}$ with non-negative diagonal entries $\lambda_1, \dots, \lambda_{\tilde{L}}$ in decreasing order such that

$$V^T \Sigma_S V = \Lambda, \quad V^T \Sigma_N V = I. \quad (1.7)$$

The columns of V are the vectors $v_1, \dots, v_{\tilde{L}}$ from above. The diagonal entries $\lambda_1, \dots, \lambda_{\tilde{L}}$ equal the corresponding SNRs, i.e., $\lambda_i = \text{SNR}_i$ for $i \in \{1, \dots, \tilde{L}\}$. If Σ_N is positive definite, then the feasibility of the simultaneous diagonalization is guaranteed (cf. Proposition A.3). Please note that one can also equivalently compute the *generalized eigendecomposition* of Σ_S and Σ_N . When Σ_N is the identity matrix, the MNF equals the PCA of the signal. In that case, Eq. (1.6) imposes an orthogonality constraint, making the PCA an orthogonal transformation. This is not the case for the MNF since V is not necessarily an orthogonal matrix. In particular, one can show that the MNF is actually a two-stage transformation that first transforms the data in such a way that the noise covariance matrix is the identity matrix, i.e., unit variance of the noise in every channel and no correlation between noise in different channels, and applies the PCA afterward. It is, therefore, sometimes also called *noise-adjusted principal components transform*.

Once the matrices V and Λ are found, similar to the PCA, the spectra of all pixels are projected onto the $L \leq \tilde{L}$ MNF components $v_1, \dots, v_L$ showing the highest SNRs. The result is a denoised and dimensionality-reduced image

$$g: \mathcal{N} \rightarrow W \subseteq \mathbb{R}^L.$$

Hence, outside of Section 1.5, we will always consider denoised images with reduced dimensionality (cf. Definitions 1.1 and 1.2).

1.5.2 Estimation of covariances of signal and noise

The covariance matrices of the signal Σ_S and the noise Σ_N are usually inaccessible in practice. Hence, the question arises of how to estimate them to apply the MNF to the data.

To estimate the signal covariance under the assumption that the signal and noise are uncorrelated, as is done in this thesis, we can use (1.3). Together with (1.7), we find

$$V^T \Sigma_T V = V^T \Sigma_S V + V^T \Sigma_N V = \Lambda + I, \quad V^T \Sigma_N V = I.$$

Consequently, instead of simultaneously diagonalizing Σ_S and Σ_N , we can also simultaneously diagonalize Σ_T and Σ_N since the eigenvectors (MNF components) are the same and the eigenvalues $\mu_1, \dots, \mu_{\tilde{L}} \in \mathbb{R}$ connected with $\lambda_1, \dots, \lambda_{\tilde{L}}$ by

$$\mu_i = \lambda_i + 1$$

for every $i \in \{1, \dots, \tilde{L}\}$. An estimate for Σ_T can be obtained by computing the empirical covariance of all spectra in the image (computed similar to Eq. (3.36)).

To estimate the noise covariance, we use the so-called *shift-difference*. The assumption is that the signal is locally constant. We explain the idea for images with two-dimensional domains, i.e., $d = 2$, but it naturally generalizes to higher dimensions. Hence, by computing the difference of the spectrum at a pixel to the spectrum of the lower right neighbor, we obtain at that pixel an estimate for the noise. If we do that for each pixel, we have a noise estimate for each position in the image from which we can compute the noise covariance matrix as the empirical covariance (computed similar to (3.36)). Only for the pixels at the lower and the right boundary, we do not get a noise estimate since they do not have lower right neighbors and are, therefore, ignored in the computation of the noise covariance. Bessel's correction (see, e.g., [12, (1.59)]) is used in both computations of the empirical covariances for the image and the noise such that we obtain unbiased estimators of the true covariances Σ_T and Σ_N . Of course, it is also possible to consider the difference to a neighbor in another direction. However, one should expect similar results. In formulas, the noise at position $(i_1, i_2) \in \mathcal{N} \setminus \{(j_1, j_2) \in \mathcal{N} \mid j_1 = N_1 \vee j_2 = N_2\}$ is estimated as

$$n(i_1, i_2) \approx \tilde{g}(i_1 + 1, i_2 + 1) - \tilde{g}(i_1, i_2).$$

Please recall that $N_1 \in \mathbb{N}$ and $N_2 \in \mathbb{N}$ are the indices of the lower and the right boundary of the grid $\mathcal{N}$, respectively. Another approach to estimate the noise at each pixel, also often used in the context of the MNF, is presented in [80], where the spectrum of each pixel is compared to the spectra in a local spatial neighborhood. However, in principle, any valid noise estimation procedure that yields a reasonable noise estimate for each pixel can be used.

To learn more about the MNF, we refer the interested reader to [10] where Berman et al. present some more properties of the MNF that go beyond what is needed for this thesis.

Chapter 2

Variants of the Mumford–Shah model

The Mumford–Shah segmentation functional is a well-established and popular model for image processing tasks, introduced by Mumford and Shah in 1989 [76]. It is the fundament of all models that are proposed and described in this thesis. Therefore, we will see a detailed description of the MS model in Section 2.1. Afterward, we will introduce and analyze our own variants of the model in Sections 2.2 to 2.4. Whenever possible, we will consider the continuous case and only switch to the discrete case when necessary.

2.1 The Mumford–Shah segmentation functional

Let $g \in L^\infty(\Omega, \mathbb{R}^L)$ be an image. Images are not continuous functions in general. If we imagine an image showing a colored object in front of a background with a different color, the image will have a discontinuity at the object boundaries. Intuitively, we would expect a segmentation method to detect the object and the background and use one segment from the partition of the image domain for the object and one for the background. In particular, this means that we expect the model to detect the discontinuities of the image at the object boundaries. Hence, from a mathematical perspective, we are dealing with a free discontinuity problem. Mumford and Shah phrase this problem as searching for a set of discontinuities $\Gamma \subseteq \Omega$, we can think of as containing the boundaries of the segments, which consists of the points where the image *jumps*, such that, first, the variation of g is low and/or smooth inside $\Omega \setminus \Gamma$, and, second, the variation of g is large and/or discontinuous across most of the points in Γ . Finding a decomposition of the image domain by detecting the discontinuities boils down to approximating the image g with piecewise smooth functions in the

sense that they are differentiable when restricted to $\Omega \setminus \Gamma$. Let $\alpha, \beta > 0$. To measure the quality of such a piecewise smooth approximation f , Mumford and Shah define the functional

$$J_{\text{sMS}}[f, \Gamma] = \alpha^2 \int_{\Omega} \|g(x) - f(x)\|_2^2 \, dx + \int_{\Omega \setminus \Gamma} \|Df(x)\|_F^2 \, dx + \beta \text{Length}(\Gamma). \quad (2.1)$$

The operator Length gives the length of the arcs in Γ that together describe the segment or object boundaries later. The first term is the data fidelity term that takes care that the approximation is as close to the measured image g as possible. The second term is a regularizer that keeps the variation of f outside of Γ small. The third summand aims to keep the length of the discontinuities short. Please note that this formulation is slightly more general than the original one since Mumford and Shah considered only grayscale images, i.e., $L = 1$; however, we also allow vector-valued images here by using the 2-norm inside the integral of the first term. All three terms are necessary to avoid trivial solutions: Dropping the first term, J_{sMS} attains the minimum by choosing f as a constant function and $\Gamma = \emptyset$. If we omit the second term, we reach the minimum with $f = g$ and $\Gamma = \emptyset$. If we drop the third term, we can choose $\Gamma = \Omega$ and $f = g$.

Based on J_{sMS} , Mumford and Shah also propose to approximate g only with piecewise constant functions by restricting f to this function class, i.e., we require f to be constant on the regions of Ω that are given by the discontinuities Γ . This implies a decomposition of Ω into pairwise disjoint sets $\mathcal{O}_1, \dots, \mathcal{O}_k \subseteq \Omega$, i.e.,

$$\Omega = \bigcup_{l=1}^k \mathcal{O}_l,$$

such that the approximating function f is constant on these sets, which gives us a partition of Ω into $\mathcal{O}_1, \dots, \mathcal{O}_k$ based on g . This partition is a segmentation of the image g . For the set of discontinuities Γ , we then have

$$\Gamma = \Omega \cap \bigcup_{l=1}^k \partial \mathcal{O}_l. \quad (2.2)$$

This variant of J_{sMS} will be the basis of the segmentation models proposed in this thesis. In this case, f is given for every $x \in \Omega$ as

$$f(x) := \sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l,$$

where $\mu_l \in \mathbb{R}^L$ and $\chi_{\mathcal{O}_l}$ is the characteristic function (cf. (1.1)) of $\mathcal{O}_l$ for $l \in \{1, \dots, k\}$. One can think of μ_l as the mean spectrum inside $\mathcal{O}_l$, i.e., the channel-wise mean over

the pixels contained in $\mathcal{O}_l$. Plugging that into (2.1) and using (2.2) leads to

$$J_{\text{sMS}}[f, \Gamma] = \alpha^2 \int_{\Omega} \left\| g(x) - \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right) \right\|_2^2 dx + \beta \text{Length}(\Gamma). \quad (2.3)$$

The second term of J_{sMS} equals 0 in this case because the derivative of constant functions vanishes. To derive the formulation of the functional, we will use throughout this thesis, we divide (2.3) by α^2 , set $\beta_0 := \beta/\alpha^2$ and reformulate J_{sMS} as follows: first, we note that for any $l \in \{1, \dots, k\}$ it holds

$$(\chi_{\mathcal{O}_l}(x) \mu_l)^T \left(\sum_{j=1}^k \chi_{\mathcal{O}_j}(x) \mu_j \right) = (\chi_{\mathcal{O}_l}(x) \mu_l)^T (\chi_{\mathcal{O}_l}(x) \mu_l),$$

by the definition of the characteristic function and because $\mathcal{O}_1, \dots, \mathcal{O}_k$ are pairwise disjoint. This implies

$$\begin{aligned} \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right)^T \left(\sum_{j=1}^k \chi_{\mathcal{O}_j}(x) \mu_j \right) &= \sum_{l=1}^k \left[(\chi_{\mathcal{O}_l}(x) \mu_l)^T \left(\sum_{j=1}^k \chi_{\mathcal{O}_j}(x) \mu_j \right) \right] \\ &= \sum_{l=1}^k (\chi_{\mathcal{O}_l}(x) \mu_l)^T (\chi_{\mathcal{O}_l}(x) \mu_l). \end{aligned}$$

With this equality, we obtain from (2.3)

$$\begin{aligned} &\alpha^{-2} J_{\text{sMS}}[f, \Gamma] \\ &= \int_{\Omega} \left\| g(x) - \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right) \right\|_2^2 dx + \beta_0 \text{Length}(\Gamma) \\ &= \int_{\Omega} g(x)^T g(x) - 2g(x)^T \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right) + \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right)^T \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right) dx \\ &\quad + \beta_0 \text{Length}(\Gamma) \\ &= \int_{\Omega} g(x)^T g(x) - 2g(x)^T \left(\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) \mu_l \right) + \sum_{l=1}^k (\chi_{\mathcal{O}_l}(x) \mu_l)^T (\chi_{\mathcal{O}_l}(x) \mu_l) dx \\ &\quad + \beta_0 \text{Length}(\Gamma) \\ &= \int_{\Omega} \sum_{l=1}^k \chi_{\mathcal{O}_l}(x) (g(x)^T g(x) - 2g(x)^T \mu_l + \mu_l^T \mu_l) dx + \beta_0 \text{Length}(\Gamma) \end{aligned}$$

$$= \sum_{l=1}^k \int_{\mathcal{O}_l} \|g(x) - \mu_l\|_2^2 \, dx + \beta_0 \text{Length}(\Gamma). \quad (2.4)$$

In the second last equality, it was used that

$$\sum_{l=1}^k \chi_{\mathcal{O}_l}(x) = 1$$

for almost every $x \in \Omega$ because of the partition property of $\mathcal{O}_1, \dots, \mathcal{O}_k$ and $\chi_{\mathcal{O}_l}^2 = \chi_{\mathcal{O}_l}$ for every $l \in \{1, \dots, k\}$. The last equality needs the linearity of the integral. If $\mathcal{O}_1, \dots, \mathcal{O}_k$ are fixed, one can show that μ_l has to be chosen as

$$\bar{\mu}_l(\mathcal{O}_l) := \frac{\int_{\mathcal{O}_l} g(x) \, dx}{|\mathcal{O}_l|}, \quad (2.5)$$

for every $l \in \{1, \dots, k\}$ to minimize the expression in (2.4) [76]. The integral is meant component-wise. Therefore, we can replace μ_l with $\bar{\mu}_l(\mathcal{O}_l)$ and, then, have to optimize only with respect to Γ . We call the formulation of the piecewise constant MS functional given in (2.4) J_{cMS} , i.e.,

$$J_{\text{cMS}}[\Gamma] := \sum_{l=1}^k \int_{\mathcal{O}_l} \|g(x) - \bar{\mu}_l(\mathcal{O}_l)\|_2^2 \, dx + \beta_0 \text{Length}(\Gamma). \quad (2.6)$$

Please note that also for J_{cMS} the minimizer is not trivial. Imagine, for instance, an image g with a black background and only a white circle in the foreground. If we perform a segmentation into $k = 2$ segments with J_{cMS} , we will obtain an approximating function f that detects both segments and their mean spectra only approximately. The reason is that $\text{Length}(\Gamma)$ equals, in this case, the perimeter of the circle and is trying to minimize it. Due to the different scaling of the two terms in the functional, it is beneficial for the minimization to make the segment of the circle slightly smaller than the circle in the original image (how much smaller depends on the chosen value of β_0). The circle segment can be placed anywhere inside the region of the original circle. Consequently, the minimizer in this example is neither trivial nor unique.

We will use J_{cMS} as the basis of our models but in a more general and slightly different formulation. In Section 1.2, we described the image segmentation task and, in particular, that one needs to define a notion of homogeneity suitable for the practical application. This part is taken over by the integrands in J_{cMS} , which are called *indicator functions* and denoted by

$$f_l: \Omega \rightarrow \mathbb{R}$$

for $l \in \{1, \dots, k\}$. It is possible to choose different indicator functions for different segments. If the indicator functions have parameters as in J_{CMS} , one also has to optimize them. Since it is difficult to optimize over the set of segment boundaries Γ , we reformulate the functional to make it depend on the partition of Ω itself and not on the boundaries of the segments. To this end, let $\mathcal{O} \subseteq \mathbb{R}^d$ be measurable. We denote the *perimeter of $\mathcal{O}$ in Ω* with $\text{Per}(\mathcal{O}, \Omega)$. We will see a mathematical definition later in Definition 2.6 and think of it for the moment as the perimeter or length of the boundary of $\mathcal{O}$ in Ω . Furthermore, we define the set

$$\mathcal{A} := \left\{ (\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \left| \begin{array}{l} \forall l \in \{1, \dots, k\} : \mathcal{O}_l \subseteq \Omega \text{ measurable,} \\ \text{Per}(\mathcal{O}_l, \Omega) < \infty, \forall m \neq l : \mathcal{O}_l \cap \mathcal{O}_m = \emptyset, \bigcup_{l=1}^k \mathcal{O}_l = \Omega \end{array} \right. \right\}, \quad (2.7)$$

which contains all partitions of Ω into k measurable subsets with finite perimeter in Ω . In particular, an element of $\mathcal{A}$ is considered to be a valid segmentation. A partition of a set into sets with finite perimeter is also called a *Caccioppoli partition*. For more information about the theory and results on Caccioppoli partitions, we refer to [5]. As it was already commented on in Section 1.2, we require the union of the segments $\mathcal{O}_l$ to be equal to Ω , which is a stronger condition than requiring the equality of the closures, since this means that the sum of the corresponding characteristic functions $\chi_{\mathcal{O}_l}$ at each $x \in \Omega$ equals one and is needed for the proofs of the existence of minimizers later. Please also note that we do not require the segments to be non-empty. Before stating the used formulation of the MS model, we observe that the equality

$$\text{Length}(\Gamma) = \frac{1}{2} \sum_{l=1}^k \text{Per}(\mathcal{O}_l, \Omega) \quad (2.8)$$

holds because Γ contains the segment boundaries in Ω , i.e., $\Gamma = \Omega \cap \left(\bigcup_{l=1}^k \partial \mathcal{O}_l \right)$. We have to multiply by $\frac{1}{2}$ because when considering the perimeters of the segments in Ω , we consider each segment boundary twice since two adjacent segments share a boundary, which counts towards the perimeters of both segments. We use the MS functional in the following form

$$J_{\text{MS}} : \mathcal{A} \rightarrow \mathbb{R},$$

$$(\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \mapsto \sum_{l=1}^k \left[\int_{\mathcal{O}_l} f_l(x) \, dx + \lambda \text{Per}(\mathcal{O}_l, \Omega) \right], \quad (2.9)$$

which we directly obtain from (2.6) by replacing the integrands by the generic indicator functions f_l , which do not need to depend on the $\mathcal{O}_l$ as in (2.6) in general, and using (2.8). The sum of the integrals over $\mathcal{O}_l$ is the data term and the sum of the perimeters of the segments a regularizer. Both terms are balanced by a regularization parameter $\lambda := \beta_0/2 > 0$. As mentioned, the indicator functions define the notion of homogeneity of pixels. The indicator functions presented in this thesis are designed for hyperspectral image data. Since we assume that each constituent in the image reflects its own spectral signature, we will only use spectral information in the indicator functions. However, we would like to point out that it would also be possible to use spatial information in the indicator functions by, e.g., considering patches of pixels. The regularizer takes care of spatial neighborhood relations of pixels. By penalizing large perimeters of segments, we take into account that neighboring pixels have a high probability to belong to the same segment because each pixel having a neighboring pixel from another segment contributes to the perimeter. If the model puts the two neighboring pixels into the same segment, the perimeter becomes smaller, which is beneficial in terms of the functional value. For ease of notation, we define the symbol

$$\mathcal{O} := (\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \in \mathcal{A}.$$

We will not optimize the sets $\mathcal{O}_1, \dots, \mathcal{O}_k$ directly but will equivalently optimize their characteristic functions, i.e., $\chi_{\mathcal{O}_1}, \dots, \chi_{\mathcal{O}_k}$, instead. Please note that from now on, we will switch between the two equivalent representations of a set without explicitly mentioning it. To optimize characteristic functions, we need a function space that allows jumps in functions. The advantage of optimizing over a function space with vector space structure is that one can relax the optimization problem and then exploit the structure to find minimizers, which we will do in our case. In fact, there is a strong connection between sets of finite perimeter and functions of bounded variation [5], and instead of optimizing with respect to $\mathcal{O}_1, \dots, \mathcal{O}_k$, we will optimize with respect to their characteristic functions over the space of functions of bounded variation (BV), as this function space meets the aforementioned requirements. Moreover, this also enables us to make use of the extensive theory about BV functions to prove the existence results. In order to establish this connection, we need to define the TV and the space of BV functions. Then, we will see that the TV of the characteristic function of a set, in fact, coincides with the perimeter of this set. We consider in the following only the necessary theory about BV functions, tailored for our purposes as a study of BV functions is out of the scope of this thesis. For a general introduction and extensive theory about BV functions, we refer to [5]. Please have in mind that we assume throughout the thesis that $\Omega \subseteq \mathbb{R}^d$ is a bounded domain with Lipschitz boundary (cf. Definition A.8).

Definition 2.1 (Total variation [5, Def. 3.4], BV functions [5, Def. 3.1, Prop. 3.6]). Let $u \in L^1(\Omega)$. We define the total variation $|u|_{\text{BV}(\Omega)}$ of u as

$$|u|_{\text{BV}(\Omega)} := \sup \left\{ \int_{\Omega} u(x) \operatorname{div} \varphi(x) \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^d), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\},$$

and call u a function of bounded variation if $|u|_{\text{BV}(\Omega)} < \infty$. Moreover, we define the space of functions of bounded variation $\text{BV}(\Omega)$ as

$$\text{BV}(\Omega) := \left\{ u \in L^1(\Omega) \mid |u|_{\text{BV}(\Omega)} < \infty \right\}$$

and equip it with the norm

$$\|u\|_{\text{BV}(\Omega)} := \|u\|_{L^1(\Omega)} + |u|_{\text{BV}(\Omega)}.$$

One can show that $(\text{BV}(\Omega), \|\cdot\|_{\text{BV}(\Omega)})$ is a Banach space [5]. Moreover, the requirement on the functions φ to lie in $C_c^\infty(\Omega, \mathbb{R}^d)$ can be replaced by $C_c^1(\Omega, \mathbb{R}^d)$. Even though this seems to be more restrictive, it leads to an equivalent definition, which can be shown with a smoothing argument [5].

Lemma 2.2 (Convexity of TV [22]).

The function $|\cdot|_{\text{BV}(\Omega)}$, defined in Definition 2.1, is convex.

The convexity (cf. Definition A.5) of the TV follows because the integral expression in the supremum of its definition is linear (and therefore convex) in u and the supremum preserves the inequality for convexity.

Remark 2.3 (Definition of BV with distributional derivatives).

One can also introduce $\text{BV}(\Omega)$ using distributional derivatives. For $u \in L^1(\Omega)$, it can be shown (cf. [5, Prop. 3.6]) that $u \in \text{BV}(\Omega)$ if and only if there is a finite vector-valued Radon measure (cf. [5, Def. 1.40 (b)]) in Ω , denoted by $Du = (D_1u, \dots, D_du)$, such that

$$\int_{\Omega} u(x) \partial_{x_i} \varphi(x) \, dx = - \int_{\Omega} \varphi(x) \, dD_i u(x)$$

for all $\varphi \in C_c^\infty(\Omega)$ and all $i \in \{1, \dots, d\}$. That is, the distributional derivative can be represented by Du in Ω . Then, it holds for the TV of u

$$|u|_{\text{BV}(\Omega)} = |Du|(\Omega),$$

where $|Du|(A)$ for a measurable set $A \subseteq \Omega$ is given as

$$|Du|(A) = \sup \left\{ \sum_{j=1}^{\infty} |Du(A_j)| \mid \forall j \in \mathbb{N} : A_j \text{ measurable}, \right. \\ \left. \forall m \neq j : A_j \cap A_m = \emptyset, A = \bigcup_{j=1}^{\infty} A_j \right\}.$$

We call $|Du|(\Omega)$ the total variation of the measure Du .

Please note that this remark is given for sake of completeness because weak* convergence in BV is defined using the finite Radon measure (cf. Definition 2.16). However, we will not directly work with it but only use the weak* compactness result in BV (cf. Theorem 2.17), which implies strong convergence in $L^1(\Omega)$. Hence, we will not give any more details and explanations about this definition of BV .

The next remark shows that the TV coincides with the norm of the gradient for differentiable functions.

Remark 2.4 (TV of differentiable functions).

Let $u \in H^{1,1}(\Omega)$ and $\varphi \in C_c^\infty(\Omega, \mathbb{R}^d)$ with $\|\varphi(x)\|_2 \leq 1$ for all $x \in \Omega$. Then, we have for $i \in \{1, \dots, d\}$ that $\varphi_i \in C_c^\infty(\Omega)$. We obtain $\varphi_i \in H^{1,\infty}(\Omega)$ for any $i \in \{1, \dots, d\}$ since $C_c^\infty(\Omega) \subseteq H^{1,\infty}(\Omega)$ (cf. Proposition A.14). This allows us to apply Gauß's theorem for Sobolev functions (cf. Theorem A.15 (2)) which gives us

$$\begin{aligned}
 \int_{\Omega} u(x) \operatorname{div} \varphi(x) \, dx &= \int_{\Omega} u(x) \sum_{i=1}^d \partial_{x_i} \varphi_i(x) \, dx \\
 &= \sum_{i=1}^d \int_{\Omega} u(x) \partial_{x_i} \varphi_i(x) \, dx \\
 &= \sum_{i=1}^d \left[- \int_{\Omega} \partial_{x_i} u(x) \varphi_i(x) \, dx + \underbrace{\int_{\partial\Omega} u(x) \varphi_i(x) \nu_i(x) \, d\mathcal{H}^{d-1}(x)}_{=0} \right] \\
 &= - \int_{\Omega} \sum_{i=1}^d \partial_{x_i} u(x) \varphi_i(x) \, dx \\
 &= \int_{\Omega} -\nabla u(x)^T \varphi(x) \, dx.
 \end{aligned} \tag{2.10}$$

The outer normal of $\partial\Omega$ is denoted with ν . The integral over $\partial\Omega$ is 0 because φ_i has compact support in Ω and is therefore equal to 0 on $\partial\Omega$. Taking the supremum over all admissible φ yields

$$\begin{aligned}
 |u|_{BV(\Omega)} &= \sup \left\{ \int_{\Omega} u(x) \operatorname{div} \varphi(x) \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^d), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\} \\
 &= \sup \left\{ \int_{\Omega} -\nabla u(x)^T \varphi(x) \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^d), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\} \\
 &= \int_{\Omega} \|\nabla u(x)\|_2 \, dx
 \end{aligned} \tag{2.11}$$

because the integrand in the penultimate line is maximized when $\varphi(x)$ has length 1 and points in the opposite direction of $\nabla u(x)$ for each $x \in \Omega$, i.e., when $\varphi(x) = -\nabla u(x) / \|\nabla u(x)\|_2$ and extending it by setting it to 0 if $\nabla u(x) = 0$. This also needs a denseness argument, using that $C_c^\infty(\Omega)$ is dense in $L^p(\Omega)$ for $1 \leq p < \infty$ (cf. Theorem A.13), to be allowed to switch the supremum with the integral since the normalized gradient is not an admissible choice for φ in general. Hence, in the sense of Eq. (2.11), the TV can be understood as the norm of the gradient, but it requires an even weaker concept of differentiability than Sobolev spaces, namely the distributional derivative (cf. Remark 2.3), and is therefore applicable to a broader class of functions.

The identity in (2.10) will later help to compute the total variation numerically when it comes to numerical optimization in Chapter 3.

So far, the perimeter of a set has not been defined mathematically. However, as already mentioned, there is a strong connection between sets of finite perimeter and BV functions. In particular, the next remark shows that the TV of a characteristic function equals the length of the boundary of the corresponding set. This will then motivate to define $\text{Per}(\mathcal{O}, \Omega)$ as the TV of the corresponding characteristic function (cf. Definition 2.6).

Remark 2.5 (TV of characteristic functions).

As an example, consider the ball $B_r^{\|\cdot\|_2}(0) \subseteq \mathbb{R}^2$ with respect to $\|\cdot\|_2$ with radius $r > 0$ around 0 and let $\varphi \in C_c^\infty(\mathbb{R}^2, \mathbb{R}^2)$ such that it holds $\|\varphi(x)\|_2 \leq 1$ for all $x \in \mathbb{R}^2$. First, note that $B_r^{\|\cdot\|_2}(0)$ is a bounded domain with Lipschitz boundary. For $i \in \{1, 2\}$, we have that $\varphi_i \in C_c^\infty(\mathbb{R}^2)$. Since $C_c^\infty(\mathbb{R}^2) \subseteq H^{1,1}(B_r^{\|\cdot\|_2}(0))$ (cf. Proposition A.14), it follows $\varphi_i \in H^{1,1}(B_r^{\|\cdot\|_2}(0))$. Thus, we are allowed to apply Gauß's theorem (cf. Theorem A.15 (1)) to φ_i for $i \in \{1, 2\}$. We find with the linearity of the integral

$$\begin{aligned}
 \int_{\mathbb{R}^2} \chi_{B_r^{\|\cdot\|_2}(0)} \operatorname{div} \varphi(x) \, dx &= \int_{B_r^{\|\cdot\|_2}(0)} \operatorname{div} \varphi(x) \, dx \\
 &= \sum_{i=1}^2 \int_{B_r^{\|\cdot\|_2}(0)} \partial_{x_i} \varphi_i(x) \, dx \\
 &= \sum_{i=1}^2 \int_{\partial B_r^{\|\cdot\|_2}(0)} \varphi_i(x) \nu_i(x) \, d\mathcal{H}^1(x) \\
 &= \int_{\partial B_r^{\|\cdot\|_2}(0)} \varphi(x)^T \nu(x) \, d\mathcal{H}^1(x) \\
 &\leq \int_{\partial B_r^{\|\cdot\|_2}(0)} |\varphi(x)^T \nu(x)| \, d\mathcal{H}^1(x) \\
 &\leq \int_{\partial B_r^{\|\cdot\|_2}(0)} \|\varphi(x)\|_2 \|\nu(x)\|_2 \, d\mathcal{H}^1(x)
 \end{aligned} \tag{2.12}$$

$$\begin{aligned}
&\leq \int_{\partial B_r^{\|\cdot\|_2}(0)} d\mathcal{H}^1(x) \\
&= 2\pi r.
\end{aligned}$$

The vector $\nu(x) \in \mathbb{R}^2$ describes the outer normal of $\partial B_r^{\|\cdot\|_2}(0)$ at $x \in \partial B_r^{\|\cdot\|_2}(0)$. The third equality is a consequence of Theorem A.15 (1). The second inequality holds due to the Cauchy-Schwarz inequality and the third because $\|\nu(x)\|_2 = 1$ for every $x \in \partial B_r^{\|\cdot\|_2}(0)$ and $\|\varphi(x)\|_2 \leq 1$ for every $x \in \mathbb{R}^2$. Hence, we have found an upper bound for the TV of $\chi_{B_r^{\|\cdot\|_2}(0)}$, which is the length of the boundary of $B_r^{\|\cdot\|_2}(0)$. Moreover, it is possible to find an admissible φ such that the expression above we started with equals the upper bound. The outer normal of $\partial B_r^{\|\cdot\|_2}(0)$ at $x \in \partial B_r^{\|\cdot\|_2}(0)$ is given as $\nu(x) = \frac{x}{\|x\|_2}$. An admissible φ with $\varphi = \nu$ on $\partial B_r^{\|\cdot\|_2}(0)$ can be found with a proper extension in normal direction. For such a φ , the integrand in (2.12) is

$$\varphi(x)^T \nu(x) = \nu(x)^T \nu(x) = \|\nu(x)\|_2^2 = 1$$

at every $x \in \partial B_r^{\|\cdot\|_2}(0)$ and the inequalities are equalities in this case. As a consequence, $\left| \chi_{B_r^{\|\cdot\|_2}(0)} \right|_{\text{BV}(\mathbb{R}^2)} = 2\pi r$, which coincides with the perimeter of a circle and, therefore, with $\text{Per}(B_r^{\|\cdot\|_2}(0), \mathbb{R}^2)$. Hence, in this example, the TV of the characteristic function coincides with the length of the boundary or the perimeter of the set. One can show in general that for a measurable, bounded domain $\mathcal{O} \subseteq \mathbb{R}^d$ with Lipschitz boundary and $\overline{\mathcal{O}} \subseteq \Omega$ the TV of the corresponding characteristic function equals the length of $\partial \mathcal{O}$ in Ω [22].

Hence, we will use the following definition of $\text{Per}(\mathcal{O}, \Omega)$:

Definition 2.6 (Sets of finite perimeter [5, Def. 3.35]).

For a measurable set $\mathcal{O} \subseteq \mathbb{R}^d$, we define $\text{Per}(\mathcal{O}, \Omega)$, the perimeter of $\mathcal{O}$ in Ω , as the TV (cf. Definition 2.1) of its characteristic function, i.e.,

$$\begin{aligned}
\text{Per}(\mathcal{O}, \Omega) &:= |\chi_{\mathcal{O}}|_{\text{BV}(\Omega)} \\
&= \sup \left\{ \int_{\mathcal{O}} \text{div } \varphi \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^d), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\}.
\end{aligned}$$

If $\text{Per}(\mathcal{O}, \Omega) < \infty$, the set $\mathcal{O}$ is said to be a set of finite perimeter in Ω .

This definition yields the following equivalence:

Proposition 2.7.

Let $\mathcal{O} \subseteq \mathbb{R}^d$ be a measurable set. Then, it holds $\text{Per}(\mathcal{O}, \Omega) < \infty$ if and only if $\chi_{\mathcal{O}} \in \text{BV}(\Omega)$.

Proof. Since Ω is bounded throughout this thesis, we always obtain that $|\mathcal{O} \cap \Omega| \leq |\Omega| < \infty$, which implies

$$\|\chi_{\mathcal{O}}\|_{L^1(\Omega)} = \int_{\Omega} |\chi_{\mathcal{O}}(x)| \, dx = \int_{\mathcal{O} \cap \Omega} dx = |\mathcal{O} \cap \Omega| < \infty,$$

and, hence, we have that $\chi_{\mathcal{O}} \in L^1(\Omega)$. It follows that $\text{Per}(\mathcal{O}, \Omega) < \infty$ if and only if $\chi_{\mathcal{O}} \in \text{BV}(\Omega)$ because $|\chi_{\mathcal{O}}|_{\text{BV}(\Omega)} = \text{Per}(\mathcal{O}, \Omega)$ (cf. Definition 2.6). $\square$

In this section, we have introduced the MS segmentation functional in the form given in (2.9), which is our tool of choice to obtain segmentations of hyperspectral images. The variable of the functional is a partition of the image domain Ω of subsets with finite perimeter. We established the connection between sets of finite perimeter and BV functions. It allows us to use the extensive theory about BV functions when proving theoretical results about the functional. The indicator functions f_l in J_{MS} remain abstract functions up to here. The next sections in this chapter are devoted to introducing and analyzing indicator functions that are designed to handle hyperspectral images.

2.2 The ϵ -regularized anisotropic Mumford–Shah model

The squared Euclidean norm that was originally used as an indicator function in J_{cMS} has two main drawbacks: first, it is isotropic and, hence, assumes the same strength of variation in each channel and, second, it is highly sensitive to outliers through the squaring, which will have a strong effect on the estimates of the mean spectra. The mean spectra can be dragged out of the region where the spectra are densely distributed and where one, therefore, would expect the mean spectra to be. In this section, we will develop and analyze an alternative indicator function for J_{MS} . It generalizes the Euclidean norm by making it anisotropic, which allows us to model the different strengths of variation in the channels caused by the spectral variability. It is done by also taking into account second-order statistics, i.e., an estimation of the covariance matrix, leading to a variant of the Mahalanobis distance [70]. The covariance matrix, which is positive semidefinite and symmetric by definition, appears inverted in the indicator function. Hence, we have to require it to be positive definite, which implies its invertibility. To achieve that this requirement is satisfied, we propose a regularization of the covariance matrix. We choose a small constant $\epsilon > 0$ and set every eigenvalue that is smaller than $\epsilon^2 > 0$ to ϵ^2 . The robustness against outliers is obtained by using the non-squared anisotropic norm as indicator function, yielding more robust estimates for the segments' mean spectra. The resulting model is called ϵAMS , standing for *ϵ -regularized anisotropic Mumford–Shah model*.

The section starts with a detailed description of the indicator function in Section 2.2.1. Since the goal is to find minimizers of J_{MS} , the next step will be to prove the existence of minimizers for J_{MS} equipped with the indicator function, shown in Section 2.2.2. The model uses a regularization with a parameter $\epsilon > 0$ that is supposed to be small. To answer the question what happens if this parameter tends to 0, we will prove the Γ -convergence of the functional to a Γ -limit in Section 2.2.3. In particular, it will turn out that we lose the guarantee for a minimizer in the Γ -limit because the functional domain is not closed anymore. The material presented in Section 2.2.1 was already published in [26]. The contents of Sections 2.2.2 and 2.2.3 and the part about the derivation of the lower bound in Section 2.2.5 were presented in [27] and are reproduced with permission from Springer Nature.

2.2.1 Definition of indicator function

As mentioned above, we will now describe an indicator function that tackles the two main drawbacks of the squared Euclidean norm when used as indicator function and additionally describe a regularization to ensure the feasibility of the model. The idea of how to make the indicator function anisotropic is visualized in Fig. 2.1. The plot

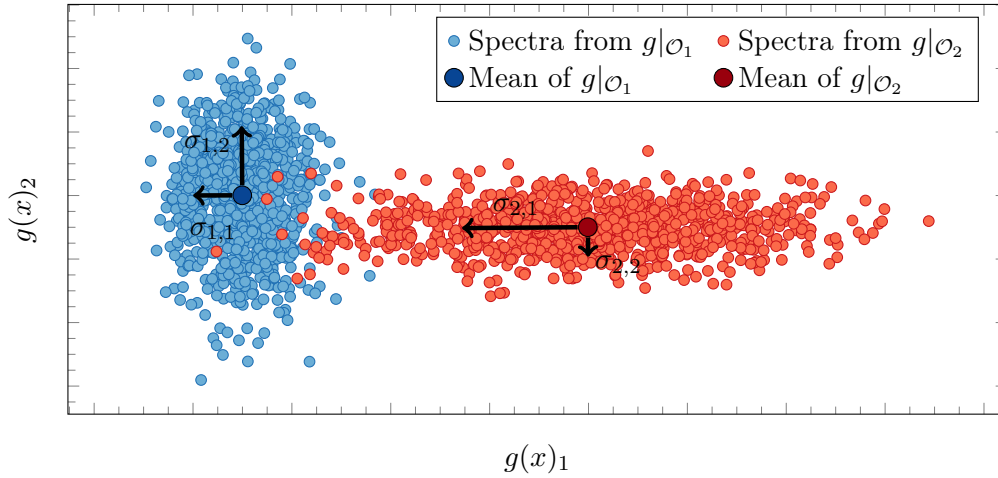


Figure 2.1: Sampled spectra from an image g with a two-dimensional value range ($L = 2$) and two spectral clusters. The arrows depict the principal components of the spectral clusters, scaled by their standard deviations.

shows sampled spectra from two spectral distributions described by $g|_{\mathcal{O}_1}$ and $g|_{\mathcal{O}_2}$ and corresponding to two segments $\mathcal{O}_1$ and $\mathcal{O}_2$ of an image $g: \Omega \rightarrow W \subseteq \mathbb{R}^2$ with a two-dimensional value range. We assume that $\mathcal{O}_1$ and $\mathcal{O}_2$ correspond to a binary segmentation of g that minimizes J_{MS} with an indicator function that is suitable for

the task to be solved. Here, we see the problem of an isotropic indicator function like the Euclidean norm. If we consider a spectrum from the segment depicted in red that lies at the left end of the red spectral cluster, it will have a smaller distance to the mean of the blue cluster when measured in the Euclidean norm. Hence, the potential to be misclassified as a spectrum from the blue cluster and, therefore, to be counted towards the wrong segment is high. However, if we had an estimate for the covariances of the two spectral clusters, we would know that the red cluster has a strong variation, quantified by the standard deviation $\sigma_{2,1} > 0$, in the direction of the horizontal principal component, depicted by the horizontal arrow, and we would also have a quantification of the variation of the blue cluster in the direction of the horizontal principal component given by the standard deviation in that direction $\sigma_{1,1}$. With this information, we can make it cheaper in terms of the objective function to deviate from the mean in directions where we know that the spectra vary strongly and make it more expensive to deviate in directions where the spectral cluster shows a weak variation. Hence, the idea of the indicator function for each segment is to measure the difference of the spectra to the mean in the Euclidean norm, but after scaling the difference with the reciprocals of the standard deviations in direction of the corresponding principal components of the spectral clusters. In Fig. 2.1, the strengths of variation, i.e., the standard deviations, are indicated by $\sigma_{1,1}$, $\sigma_{1,2}$, $\sigma_{2,1}$ and $\sigma_{2,2}$. In the concrete example, we would get that since the standard deviation $\sigma_{1,1}$ of the blue spectral cluster is small, it would result in a large factor $1/\sigma_{1,1}$ which makes it particularly expensive to deviate from the mean of the blue cluster in that direction. On the other hand, the red cluster has a high standard deviation $\sigma_{2,1}$ in that direction with a consequence that we scale in that direction with a small number $1/\sigma_{2,1}$ and make it, therefore, cheap to deviate from the red mean in that direction. Hence, a spectrum of the red cluster has a good chance to be closer to the red mean when measured in this anisotropic indicator function.

The difference between squaring and not squaring the Euclidean norm is shown in Fig. 2.2. It shows samples from the spectral distribution corresponding to a segment $\mathcal{O} \subseteq \Omega$ of an image $g: \Omega \rightarrow W \subseteq \mathbb{R}^2$, again with a two-dimensional value range. For the samples, the mean was computed, once with respect to the squared Euclidean norm $\|\cdot\|_2^2$, which is the classical indicator function of the MS functional, and with respect to the non-squared Euclidean norm $\|\cdot\|_2$. We observe that in the former case, the mean is dragged away from the center of the region where the spectra are densely distributed, which is caused by the sparsely distributed spectra to the left of the dense region. The latter is much more robust against these sparsely distributed spectra and much closer to the center of the dense region where we would expect the mean to be.

Let the set of symmetric and positive definite matrices of dimension $L \times L$ be denoted as $\mathcal{P}$, i.e.,

$$\mathcal{P} := \{M \in \mathbb{R}^{L \times L} \mid M^T = M, \forall z \in \mathbb{R}^L \setminus \{0\} : z^T M z > 0\}. \quad (2.13)$$

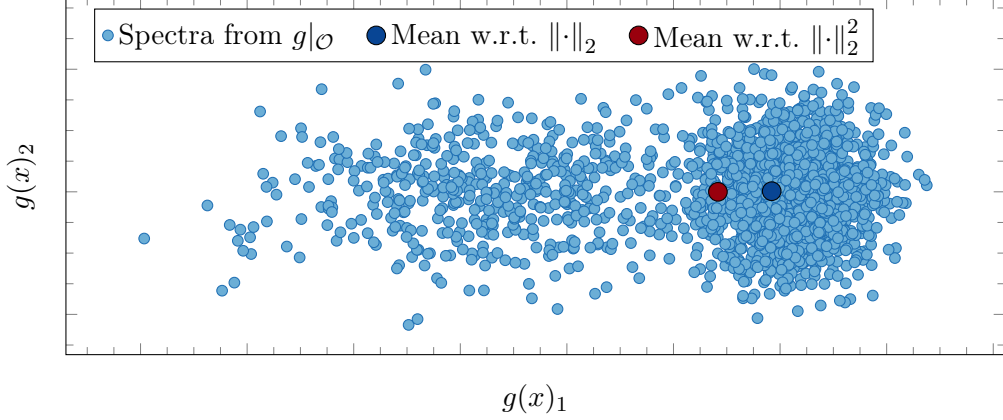


Figure 2.2: Samples from a spectral distribution corresponding to a segment $\mathcal{O}$ of an image g with a two-dimensional value range ($L = 2$) and the means computed with respect to the squared $\|\cdot\|_2^2$ and non-squared $\|\cdot\|_2$ Euclidean norm.

For $l \in \{1, \dots, k\}$, we define the indicator function for segment l at any $x \in \Omega$ as

$$f_l(x; \mu_l, \Sigma_l) := \sqrt{(g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l)} + \log \det \Sigma_l, \quad (2.14)$$

with $\mu_l \in \mathbb{R}^L$ and $\Sigma_l \in \mathcal{P}$ being estimates for the mean and the covariance matrix of the segment, respectively. So far, the $\log \det \Sigma_l$ term has not been discussed but will be motivated later. This definition requires that Σ_l is invertible and Σ_l^{-1} positive semidefinite, which is equivalent to Σ_l being positive definite, i.e., $\Sigma_l \in \mathcal{P}$. From this requirement, it then follows the feasibility of the logarithm and the feasibility of the square root term. A covariance matrix is symmetric and positive semidefinite by definition. Hence, we only need to achieve positive definiteness of the matrix. We assume for the moment that $\Sigma_l \in \mathcal{P}$ and show below how we ensure this in practice. The first summand in (2.14) is the anisotropic Euclidean distance of the spectrum at $x \in \Omega$ to μ_l . By the spectral theorem for symmetric real matrices, we can orthogonally diagonalize Σ_l with an orthogonal matrix $V_l \in O(L)$ and a diagonal matrix $D_l^2 \in \mathbb{R}^{L \times L}$ with positive diagonal elements $\sigma_{l,i}^2 > 0$ for $i \in \{1, \dots, L\}$ such that $\Sigma_l = V_l D_l^2 V_l^T$. The eigenvectors $v_{l,i} \in \mathbb{R}^L$ of Σ_l are given as the columns of V_l and correspond to the principal directions of variation in the segment's spectral distribution. The diagonal elements of D_l^2 are the variances of the spectral distribution along the $v_{l,i}$ for $i \in \{1, \dots, L\}$. By taking the square root of D_l^2 , we obtain a diagonal matrix $D_l \in \mathbb{R}^{L \times L}$ with the square roots of the diagonal elements of D_l^2 on its diagonal, which are the standard deviations $\sigma_{l,1}, \dots, \sigma_{l,L}$. It holds $D_l^2 = D_l D_l$. This relation between variance and standard deviation, i.e., that the square root of the

variance equals the standard deviation, motivates to denote the diagonal matrix D_l^2 in the eigendecomposition of Σ_l as the square of another matrix D_l . In particular, the eigendecomposition of Σ_l^{-1} is given as $\Sigma_l^{-1} = V_l(D_l^2)^{-1}V_l^T$ with $(D_l^2)^{-1}$ having the reciprocals of $\sigma_1^2, \dots, \sigma_L^2$ on its diagonal. Moreover, we have $\Sigma_l^{-1} \in \mathcal{P}$ as the inverse of a symmetric matrix is also symmetric and positive definite since its eigenvalues are the reciprocals of the eigenvalues of Σ_l , which are positive. We can now write (2.14) as follows:

$$\begin{aligned} f_l(x; \mu_l, \Sigma_l) &= \sqrt{(g(x) - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g(x) - \mu_l)} + \log \det \Sigma_l \\ &= \sqrt{(g(x) - \mu_l)^T V_l D_l^{-1} D_l^{-1} V_l^T (g(x) - \mu_l)} + \log \left(\prod_{i=1}^L \sigma_{l,i}^2 \right) \\ &= \sqrt{[D_l^{-1} V_l^T (g(x) - \mu_l)]^T [D_l^{-1} V_l^T (g(x) - \mu_l)]} + \sum_{i=1}^L \log \sigma_{l,i}^2 \\ &= \|D_l^{-1} V_l^T (g(x) - \mu_l)\|_2 + \sum_{i=1}^L \log \sigma_{l,i}^2, \end{aligned}$$

where we used that the determinant of a matrix equals the product of its eigenvalues. As planned, the indicator function coincides with the Euclidean norm where the difference of the spectrum $g(x)$ and the mean μ_l is measured after transforming to the eigenvector basis of Σ_l and scaling by the reciprocals of the standard deviations. This is done by projecting $g(x) - \mu_l$ onto the eigenvectors of Σ_l , i.e., by multiplication with V_l^T , and then multiplying the entries with $1/\sigma_1, \dots, 1/\sigma_L$, i.e., by a multiplication with D_l^{-1} . From this formulation of the indicator function, one can easily see that the first summand (the anisotropic Euclidean norm) of the indicator function cannot be used alone. The resulting functional that we obtain when using f_l as the indicator functions in J_{MS} also has $\mu_1, \dots, \mu_k$ and $\Sigma_1, \dots, \Sigma_k$ as unknowns. Hence, when optimizing the functional, we also have to optimize with respect to the estimates for mean and covariance of the segments. However, if we had only the first summand in f_l , we could make the indicator function come arbitrarily close to 0 by choosing Σ_l such that $\sigma_1^2, \dots, \sigma_L^2$ are very large because this makes the scaling factors $1/\sigma_1, \dots, 1/\sigma_L$ approach 0. Hence, a second term is needed that benefits from making $\sigma_1^2, \dots, \sigma_L^2$ small. The term $\log \det \Sigma_l$ takes over this role. Since it is equal to the sum over the logarithms of the eigenvalues of Σ_l and the logarithm is monotonically increasing, this term tries to make the variances small. In this sense, the two terms balance each other and therefore, both are needed. The reason to take the term $\log \det \Sigma_l$ and not any other term that benefits from small variances is that our approach models the spectral variability of the spectral signature of the constituent represented by segment $\mathcal{O}_l$ with first- and second-order statistics, i.e., with mean and covariance, which is similar

to modeling the spectral distribution of segment l with a Gaussian distribution. If we assume that $f^G(\cdot; \mu_l, \Sigma_l)$ describes the density of a Gaussian distribution with mean μ_l and covariance Σ_l , then this means that this density assigns the mass

$$f^G(g(x); \mu_l, \Sigma_l) = \frac{1}{\sqrt{(2\pi)^L \det \Sigma_l}} \exp \left[-\frac{1}{2} (g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l) \right] \quad (2.15)$$

to a spectrum $g(x)$ for $x \in \Omega$. To obtain maximum likelihood estimates for μ_l and Σ_l , one can equivalently consider $\log f^G(g(x); \mu_l, \Sigma_l)$; we arrive at

$$\begin{aligned} \log f^G(g(x); \mu_l, \Sigma_l) &= -\log \sqrt{(2\pi)^L \det \Sigma_l} - \frac{1}{2} (g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l) \\ &= -\frac{L}{2} \log(2\pi) - \frac{1}{2} \log \det \Sigma_l - \frac{1}{2} (g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l). \end{aligned}$$

The first term is a constant and hence, does not influence mean and covariance estimate in a maximum likelihood estimation. We see that $\log \det \Sigma_l$ appears here together with the squared anisotropic Euclidean norm, which justifies our choice for the term in our indicator function that keeps the variances small.

One can also interpret the $\log \det \Sigma_l$ term geometrically: The eigenvalues of Σ_l describe how much the spectral cluster of the l -th segment extends in the directions of $v_{l,1}, \dots, v_{l,L}$ and their determinant can therefore be understood as a quantification of the volume of the spectral cluster. By keeping these quantities small, we try to keep the spectral cluster of the segment small and the term is, therefore, like a built-in spectral volume regularization. This also helps to balance the data term of J_{MS} with the regularizing perimeter. Since the perimeter considers the spatial neighborhood relations, it is possible that pixels that are spatially close are put into the same segment although they are spectrally far apart. This increases the volume of the estimated spectral cluster. By the $\log \det \Sigma_l$ term, such a behavior is hampered. One can further observe that for fixed Σ_l , the term reduces to a constant shift of the indicator function. Please note that the first summand in Eq. (2.14) is, in fact, the Mahalanobis distance [70] with respect to the covariance matrix.

The unknowns μ_l and Σ_l of f_l have to be estimated from the data. This will be done by optimizing the functional also with respect to them with a derivative-based procedure, described in Section 3.3.1, which uses the gradient with respect to μ_l and Σ_l . Hence, we need differentiability of f_l in μ_l and Σ_l . Since the square root is not differentiable at 0, we add a small constant $\eta > 0$ to the term under the square root and define the regularized indicator function $f_{\eta,l}$ as

$$f_{\eta,l}(x; \mu_l, \Sigma_l) := \sqrt{(g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l) + \eta} + \log \det \Sigma_l. \quad (2.16)$$

Because we have assumed that $\Sigma_l \in \mathcal{P}$, the vector-matrix-vector product under the square root is non-negative. By adding η , the argument of the square root is always

positive and we achieve differentiability with respect to μ_l and Σ_l . This is the indicator function we will work with in the following.

Remark 2.8.

It is easy to show that for $M \in \mathcal{P}$, the function

$$\begin{aligned} \langle \cdot, \cdot \rangle_M : \mathbb{R}^L \times \mathbb{R}^L &\rightarrow \mathbb{R}, \\ (y, z) &\mapsto y^T M z, \end{aligned}$$

defines an inner product on $\mathbb{R}^L$ and that $\|z\|_M := \sqrt{\langle z, z \rangle_M}$ for any $z \in \mathbb{R}^L$ is a norm on $\mathbb{R}^L$. Hence, the first summand in (2.14) is a norm on $\mathbb{R}^L$. Since we regularize this norm with η , the first term in our indicator function (cf. (2.16)) is a regularized norm, which motivates the following notation that we will use throughout the thesis:

$$\|z\|_{\Sigma_l^{-1}, \eta} := \sqrt{z^T \Sigma_l^{-1} z + \eta} \quad (2.17)$$

for any $z \in \mathbb{R}^L$. This allows us to write $f_{\eta, l}$ as

$$f_{\eta, l}(x; \mu_l, \Sigma_l) = \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l.$$

For the considerations above, we have assumed that $\Sigma_l \in \mathcal{P}$ to obtain its invertibility and the positive semidefiniteness of its inverse. However, a covariance matrix is only positive semidefinite in general. Therefore, the question arises how to handle the case where the covariance estimate Σ_l is symmetric but only positive semidefinite. In that case, we propose to use a regularization for Σ_l to ensure positive definiteness. Let $\epsilon > 0$ be fixed. We define the mapping that returns the eigenvalues of a symmetric matrix in decreasing order as

$$\sigma^2 : \{M \in \mathbb{R}^{L \times L} \mid M^T = M\} \rightarrow \mathbb{R}^L, \quad (2.18)$$

i.e., $\sigma^2(M)_1 \geq \sigma^2(M)_2 \geq \dots \geq \sigma^2(M)_L$. Furthermore, we define the set of symmetric and positive definite matrices with eigenvalues greater or equal to ϵ^2 as

$$\mathcal{P}_\epsilon := \{M \in \mathcal{P} \mid \forall i \in \{1, \dots, L\} : \sigma^2(M)_i \geq \epsilon^2\}. \quad (2.19)$$

We have $\mathcal{P}_\epsilon \subseteq \mathcal{P}$. Let $\Sigma_l = V_l D_l^2 V_l^T$ be the eigendecomposition of the covariance matrix with $V_l \in O(L)$ and the diagonal matrix $D_l^2 \in \mathbb{R}^{L \times L}$, as above, but now having only non-negative diagonal elements $\sigma_{l,i}^2 \geq 0$ for $i \in \{1, \dots, L\}$. Again, we have $D_l^2 = D_l D_l$. We define the matrix $D_{\epsilon, l} \in \mathbb{R}^{L \times L}$ as a diagonal matrix with the diagonal entries

$$\max\{\epsilon, \sigma_{l,1}\}, \dots, \max\{\epsilon, \sigma_{l,L}\}.$$

Hence, $D_{\epsilon,l}$ is a version of D_l where each element on the diagonal that is smaller than ϵ is replaced by ϵ . Furthermore, we define the matrix

$$\Sigma_{\epsilon,l} := V_l D_{\epsilon,l} D_{\epsilon,l} V_l^T = V_l D_{\epsilon,l}^2 V_l^T.$$

This matrix coincides with Σ_l except for the eigenvalues that are greater or equal to ϵ^2 . In particular, it holds that $\Sigma_{\epsilon,l} \in \mathcal{P}_\epsilon \subseteq \mathcal{P}$. If $\Sigma_l \in \mathcal{P}_\epsilon$, then we obtain $\Sigma_l = \Sigma_{\epsilon,l}$ because the regularization leaves matrices in $\mathcal{P}_\epsilon$ unchanged. It can, therefore, be understood as a projection onto $\mathcal{P}_\epsilon$. It turns out that the described ϵ -regularization has several more important roles in our segmentation framework, which will be discussed later. In particular, it is a key to show the existence of minimizers of the model, as we will see in Section 2.2.2.

Now, we can state our first variant of the MS functional that we call ϵ AMS. In preparation of that, we introduce two more symbols for ease of notation

$$\boldsymbol{\mu} := (\mu_1, \dots, \mu_k) \in (\mathbb{R}^L)^k, \quad \boldsymbol{\Sigma} := (\Sigma_1, \dots, \Sigma_k) \in \mathcal{P}^k.$$

Let $X_\epsilon := \mathcal{A} \times (\mathbb{R}^L)^k \times \mathcal{P}_\epsilon^k$. Then, we define ϵ AMS as

$$J_\epsilon: X_\epsilon \rightarrow \mathbb{R},$$

$$(\mathcal{O}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \right]. \quad (2.20)$$

The data term of the model has a certain similarity with the GMM, which assumes that the data distribution is a convex combination of Gaussian densities (cf. (2.15)). Both assume that the spectral cluster of a segment obeys a Gaussian distribution and, hence, model it with first- and second-order statistics. Nevertheless, there are two important differences between the models: (1) our model supposes equally likely clusters and (2) the Mahalanobis distance in our model is wrapped in a square root, in contrast to the Gaussian distribution in a GMM. Since we are not bound to fulfill the requirements for a probability density, we have generally more freedom regarding the design of our segmentation model.

We can also observe that the described regularization of the covariance estimates has a geometric interpretation. With the regularization the $\log \det \Sigma_l$ term becomes

$$\log \det \Sigma_{\epsilon,l} = \log \left(\prod_{i=1}^L \max \{ \epsilon^2, \sigma_{l,i}^2 \} \right).$$

When computing a segmentation of an image by minimizing J_ϵ , the algorithm aims to make $\log \det \Sigma_{\epsilon,l}$ small by minimizing the eigenvalues of $\Sigma_{\epsilon,l}$. However, once $\sigma_{l,i}^2 \leq \epsilon^2$ for all $i \in \{1, \dots, L\}$, we have

$$\log \det \Sigma_{\epsilon,l} = \log \epsilon^{2L}.$$

Hence, we find a lower bound for the spectral volume regularizer, which depends on ϵ . In this sense, ϵ defines a minimum volume of the spectral clusters of the estimated segments. This is of particular interest for piecewise constant images because these show no spectral variation and the Mahalanobis term can, therefore, become 0, allowing the $\log \det \Sigma_l$ term to bring the eigenvalues of Σ_l arbitrarily close to 0. Since, in hyperspectral imaging, one usually makes the assumption that constituents or materials in the captured scene reflect the pure spectral signatures of these constituents or materials, hyperspectral images are assumed to be images without spectral variability and, hence, supposed to be described by piecewise constant functions. Therefore, we expect our model to be able to handle such images.

Another effect of this regularization is that the conditions of the regularized covariance matrices are improved compared to the unregularized covariance estimates. This means that the regularized matrices are farther away from being singular and, therefore, their inverses can be computed with higher accuracy, which is interesting for solving the minimization problem numerically.

2.2.2 Existence of minimizers

Since we find segmentations of images by minimizing J_ϵ , a natural question is if minimizers actually exist and if it is therefore meaningful to search for a minimizer. This subsection is devoted to answer that question for J_ϵ for which we will prove an existence result.

The reasoning to show that minimizers of J_ϵ exist follows the *direct method in the calculus of variations*. Four steps have to be undertaken:

1. proof of a lower bound $\underline{J}_\epsilon \in \mathbb{R}$ of J_ϵ , i.e.,

$$\underline{J}_\epsilon := \inf_{(\mathcal{O}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \in X_\epsilon} J_\epsilon[\mathcal{O}, \boldsymbol{\mu}, \boldsymbol{\Sigma}] > -\infty,$$

2. selection of a minimizing sequence $(\mathcal{O}_n, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)_{n \in \mathbb{N}} \subseteq X_\epsilon$; for such a sequence, it holds $J_\epsilon[\mathcal{O}_n, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n] \xrightarrow{n \rightarrow \infty} \underline{J}_\epsilon$,
3. construction of a subsequence $(\mathcal{O}_{n_m}, \boldsymbol{\mu}_{n_m}, \boldsymbol{\Sigma}_{n_m})_{m \in \mathbb{N}}$ of the minimizing sequence and an element $(\mathcal{O}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \in X_\epsilon$ that satisfy

$$J_\epsilon[\mathcal{O}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*] \leq \liminf_{m \rightarrow \infty} J_\epsilon[\mathcal{O}_{n_m}, \boldsymbol{\mu}_{n_m}, \boldsymbol{\Sigma}_{n_m}], \quad (2.21)$$

4. conclusion that $(\mathcal{O}^*, \mu^*, \Sigma^*)$ minimizes J_ϵ because

$$\begin{aligned}
 \underline{J}_\epsilon &= \lim_{m \rightarrow \infty} J_\epsilon[\mathcal{O}_{n_m}, \mu_{n_m}, \Sigma_{n_m}] \\
 &= \liminf_{m \rightarrow \infty} J_\epsilon[\mathcal{O}_{n_m}, \mu_{n_m}, \Sigma_{n_m}] \\
 &\geq J_\epsilon[\mathcal{O}^*, \mu^*, \Sigma^*] \\
 &\geq \underline{J}_\epsilon.
 \end{aligned} \tag{2.22}$$

The first equality in (2.22) holds because the sequence $(J_\epsilon[\mathcal{O}_{n_m}, \mu_{n_m}, \Sigma_{n_m}])_{m \in \mathbb{N}}$ converges to $\underline{J}_\epsilon$ as $(\mathcal{O}_{n_m}, \mu_{n_m}, \Sigma_{n_m})_{m \in \mathbb{N}}$ is a subsequence of a minimizing sequence. The second equality is true because the limit and the limit inferior of a convergent sequence are equal. Since starting point and endpoint of (2.22) are the same, all involved inequalities have to be equalities and, thus, $(\mathcal{O}^*, \mu^*, \Sigma^*)$ minimizes the functional J_ϵ .

Please note that we deviate from the usual way of applying the direct method in the third step and only show the existence of one (not necessarily convergent) subsequence that fulfills (2.21) instead of showing the existence of a convergent subsequence and the lower semi-continuity of the functional. For the proof of the existence result, we need some definitions, lemmas and propositions that we will state first. First, we recall that the symmetric difference $A \triangle B$ of two sets A and B is computed as $A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$. As the notion of convergence for subsets of $\mathbb{R}^d$, we use *convergence in measure*.

Definition 2.9 (Convergence in measure).

Consider a sequence $(A_n)_{n \in \mathbb{N}}$ of measurable, bounded subsets of $\mathbb{R}^d$ and let $A \subseteq \mathbb{R}^d$ be a measurable, bounded set. Then, $(A_n)_{n \in \mathbb{N}}$ is said to converge to A in measure in Ω if the Lebesgue measure of the symmetric difference $|(A_n \triangle A) \cap \Omega|$ of A_n and A in Ω goes to 0, which means that

$$|(A_n \triangle A) \cap \Omega| \rightarrow 0$$

when $n \rightarrow \infty$. We write: $A_n \rightarrow A$ for $n \rightarrow \infty$.

Please note that since Ω is bounded, convergence in measure and local convergence in measure coincide [5, Rem. 3.37]. Convergence in measure is, in fact, equivalent to $L^1(\Omega)$ -convergence of the corresponding characteristic functions because of the following proposition.

Proposition 2.10.

For measurable sets $A, B \subseteq \mathbb{R}^d$, it holds

$$|(A \triangle B) \cap \Omega| = \|\chi_A - \chi_B\|_{L^1(\Omega)}.$$

Proof. Let $x \in \mathbb{R}^d$. Then, we have

$$\begin{aligned} & |\chi_A(x) - \chi_B(x)| = 1 \\ \Leftrightarrow & (x \in A \wedge x \notin B) \vee (x \notin A \wedge x \in B) \\ \Leftrightarrow & x \in A \triangle B \\ \Leftrightarrow & \chi_{A \triangle B}(x) = 1. \end{aligned}$$

Since $|\chi_A(\cdot) - \chi_B(\cdot)|$ and $\chi_{A \triangle B}$ only attain values in $\{0, 1\}$, it follows

$$|\chi_A(x) - \chi_B(x)| = \chi_{A \triangle B}(x)$$

for all $x \in \mathbb{R}^d$. That implies

$$|(A \triangle B) \cap \Omega| = \int_{\Omega} \chi_{A \triangle B}(x) \, dx = \int_{\Omega} |\chi_A(x) - \chi_B(x)| \, dx = \|\chi_A - \chi_B\|_{L^1(\Omega)}. \quad \square$$

We can therefore switch between a sequence $(\mathcal{O}_n)_{n \in \mathbb{N}} \subseteq \mathcal{A}$ and its corresponding sequence of characteristic functions. It will be shown next that if measurable and bounded sets converge in measure in Ω to a measurable and bounded set, also their Lebesgue measures restricted to Ω converge to the measure of the limit set restricted to Ω .

Proposition 2.11.

For a sequence $(A_n)_{n \in \mathbb{N}}$ of measurable, bounded subsets of $\mathbb{R}^d$ and a measurable, bounded set $A \subseteq \mathbb{R}^d$ such that $A_n \xrightarrow{n \rightarrow \infty} A$ in Ω , it holds that

$$|A_n \cap \Omega| \rightarrow |A \cap \Omega| \text{ when } n \rightarrow \infty,$$

which means that also the Lebesgue measures of the sets restricted to Ω converge.

Proof. It holds

$$\begin{aligned} ||A_n \cap \Omega| - |A \cap \Omega|| &= \left| \int_{\Omega} \chi_{A_n}(x) \, dx - \int_{\Omega} \chi_A(x) \, dx \right| \\ &\leq \int_{\Omega} |\chi_{A_n}(x) - \chi_A(x)| \, dx \\ &= \|\chi_{A_n} - \chi_A\|_{L^1(\Omega)} \\ &\stackrel{\text{Proposition 2.10}}{=} |(A_n \triangle A) \cap \Omega| \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad \square$$

What we will also need for the existence proof is that $\mathcal{P}_{\epsilon}$ is closed.

Proposition 2.12 (Continuity of eigenvalues [84, 3.1.2]).

The eigenvalues of a matrix vary continuously with the entries of the matrix.

Lemma 2.13.

For any fixed $\epsilon > 0$, we have that the set $\mathcal{P}_\epsilon$ is a closed subset of $\mathbb{R}^{L \times L}$.

Proof. By Proposition 2.12, we have continuous dependence of the eigenvalues on the matrix entries. Furthermore, the transpose is a linear function on a finite-dimensional space, which implies its continuity. Since equalities and non-strict inequalities remain true in the limit, the symmetry and inequality condition in the definition of $\mathcal{P}_\epsilon$ are preserved in the limit. The preservation of these two conditions additionally implies the positive definiteness in the limit because positive definiteness is equivalent to symmetry and positivity of the eigenvalues. We obtain the closedness of $\mathcal{P}_\epsilon$. $\square$

We list some further definitions and results on BV functions that we need for the proof.

Lemma 2.14.

Consider a symmetric and positive semidefinite matrix $M \in \mathbb{R}^{L \times L}$. Then, we have

$$\sigma^2(M)_L \|z\|_2^2 \leq z^T M z \leq \sigma^2(M)_1 \|z\|_2^2 \quad \text{for all } z \in \mathbb{R}^L,$$

where $\sigma^2(M)_1$ and $\sigma^2(M)_L$ are the largest and smallest eigenvalue of M , respectively.

Proof. Since M is symmetric, the spectral theorem for symmetric real matrices guarantees the existence of an orthogonal matrix $V \in O(L)$ and a diagonal matrix $D \in \mathbb{R}^{L \times L}$ with diagonal entries $\sigma^2(M)_1, \dots, \sigma^2(M)_L$ such that $M = V D V^T$. We define $w = V^T z$ for $z \in \mathbb{R}^L$. Then, we have $z = V w$ and obtain

$$\begin{aligned} z^T M z &= (V w)^T M V w = w^T D w = \sum_{i=1}^L \sigma^2(M)_i w_i^2 \\ &\leq \sigma^2(M)_1 \sum_{i=1}^L w_i^2 = \sigma^2(M)_1 w^T w = \sigma^2(M)_1 (V^T z)^T V^T z = \sigma^2(M)_1 z^T z. \end{aligned}$$

The lower bound can be obtained analogously. $\square$

The following proposition uses the notion of an extension domain. We will not give a formal definition of an extension domain since we do not work with it explicitly but only need to conclude that Ω is an extension domain, which is required to apply Theorem 2.17. A definition can be found, e.g., in [5, Def. 3.20].

Proposition 2.15 ([5, Prop. 3.21]).

Let $\Omega \subseteq \mathbb{R}^d$ be open. If Ω has a compact Lipschitz boundary, it is an extension domain.

Definition 2.16 (Weak* convergence in BV [5, Def. 3.11]).

Let $(u_n)_{n \in \mathbb{N}} \subseteq \text{BV}(\Omega)$ be a sequence and $u \in \text{BV}(\Omega)$. If $u_n \xrightarrow{n \rightarrow \infty} u$ in $L^1(\Omega)$ and the corresponding sequence of finite Radon measures $(Du_n)_{n \in \mathbb{N}}$ (cf. Remark 2.3) converges weakly* to Du in Ω , then $(u_n)_{n \in \mathbb{N}}$ is said to be weakly* convergent in $\text{BV}(\Omega)$ with limit u .

Also here, we avoid explicit formulas showing what is meant that the sequence of finite Radon measures converges weakly* because we do not explicitly work with it but only refer to [5, Def. 3.11] for further details.

Theorem 2.17 (Compactness in BV [5, Th. 3.23]).

Let Ω be a bounded extension domain and let $(u_n)_{n \in \mathbb{N}} \subseteq \text{BV}(\Omega)$ be a bounded sequence in $\text{BV}(\Omega)$. Then, there exists a subsequence $(u_{n_m})_{m \in \mathbb{N}}$ that converges weakly* to $u \in \text{BV}(\Omega)$.

Lemma 2.18 ([4, Lem. 3.22 (1)]).

Let $(f_n)_{n \in \mathbb{N}} \subseteq L^p(\Omega)$ be a sequence and $f \in L^p(\Omega)$ for $1 \leq p < \infty$. If $f_n \rightarrow f$ when $n \rightarrow \infty$ in $L^p(\Omega)$, then the existence of a subsequence $(f_{n_m})_{m \in \mathbb{N}}$ follows that satisfies $f_{n_m}(x) \rightarrow f(x)$ in $\mathbb{R}$ when $m \rightarrow \infty$ for almost every $x \in \Omega$.

Proposition 2.19 (Semi-continuity of perimeter [5, Rem. 3.37, Prop. 3.38 (b)]).

The mapping $\mathcal{O} \mapsto \text{Per}(\mathcal{O}, \Omega)$ is a lower semi-continuous function in regard to convergence in measure in Ω , i.e., for a sequence of measurable bounded subsets $(\mathcal{O}_n)_{n \in \mathbb{N}}$ of $\mathbb{R}^d$ and a measurable, bounded set $\mathcal{O} \subseteq \mathbb{R}^d$ such that $\mathcal{O}_n \rightarrow \mathcal{O}$ when $n \rightarrow \infty$, it holds

$$\text{Per}(\mathcal{O}, \Omega) \leq \liminf_{n \rightarrow \infty} \text{Per}(\mathcal{O}_n, \Omega).$$

Theorem 2.20 (Fatou's lemma [5, Th. 1.20]).

Let $(u_n)_{n \in \mathbb{N}}$ be a sequence of measurable functions with $u_n: \Omega \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ and $\underline{u} \in L^1(\Omega)$. If $u_n \geq \underline{u}$ almost everywhere for every $n \in \mathbb{N}$, it holds

$$\int_{\Omega} \liminf_{n \rightarrow \infty} u_n(x) \, dx \leq \liminf_{n \rightarrow \infty} \int_{\Omega} u_n(x) \, dx.$$

Theorem 2.21 (Existence of minimizers of ϵAMS).

For the functional J_{ϵ} , defined in (2.20), there exists a minimizer in X_{ϵ} .

Proof. We follow the steps of the direct method in the calculus of variations that were outlined above.

The first step is to show that J_ϵ has a lower bound. To see this, we reformulate J_ϵ as follows:

$$\begin{aligned}
 J_\epsilon[\mathcal{O}, \mu, \Sigma] &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \right] \\
 &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} \, dx + \log \det \Sigma_l \int_{\mathcal{O}_l} dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \right] \\
 &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} \, dx + |\mathcal{O}_l| \log \det \Sigma_l + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \right].
 \end{aligned} \tag{2.23}$$

It was used that $\log \det \Sigma_l$ is independent of $x \in \Omega$ and the linearity of the integral. The positive definiteness of Σ_l^{-1} , which follows from the positive definiteness of $\Sigma_l \in \mathcal{P}_\epsilon$, and $\eta > 0$ imply

$$\|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} = \sqrt{(g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l) + \eta} > 0 \tag{2.24}$$

for any $x \in \Omega$ and each $l \in \{1, \dots, k\}$. The second term in the inner sum in (2.23) is also bounded below for every $l \in \{1, \dots, k\}$ because

$$\begin{aligned}
 |\mathcal{O}_l| \log \det \Sigma_l &= |\mathcal{O}_l| \log \prod_{i=1}^L \sigma^2(\Sigma_l)_i \geq |\mathcal{O}_l| \log \prod_{i=1}^L \epsilon^2 = |\mathcal{O}_l| \log \epsilon^{2L} \\
 &\geq -|\Omega| |\log \epsilon^{2L}| > -\infty,
 \end{aligned} \tag{2.25}$$

which holds since $\log$ is a monotonically increasing function, $\Sigma_l \in \mathcal{P}_\epsilon$ and $\infty > |\Omega| \geq |\mathcal{O}_l| \geq 0$ by $\mathcal{O}_l \subseteq \Omega$ and $|\cdot|$ being a measure. Together with $\operatorname{Per}(\mathcal{O}_l, \Omega) = |\chi_{\mathcal{O}_l}|_{\operatorname{BV}(\Omega)} \geq 0$ and $\lambda > 0$, we deduce that there is a constant $\underline{c} := -k |\Omega| |\log \epsilon^{2L}| > -\infty$ such that

$$J_\epsilon[\mathcal{O}, \mu, \Sigma] \geq \underline{c} \text{ for any } (\mathcal{O}, \mu, \Sigma) \in X_\epsilon. \tag{2.26}$$

The existence of a lower bound of J_ϵ guarantees the existence of a minimizing sequence

$$(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X_\epsilon.$$

Since the sequence of evaluations $(J_\epsilon[\mathcal{O}_n, \mu_n, \Sigma_n])_{n \in \mathbb{N}}$ of the minimizing sequence is convergent in $\mathbb{R}$, it is bounded. Therefore, there exists a constant $\overline{C} < \infty$ such that

$$J_\epsilon[\mathcal{O}_n, \mu_n, \Sigma_n] \leq \overline{C} \text{ for all } n \in \mathbb{N}. \tag{2.27}$$

We will use $\overline{C}$ later to show the existence of convergent subsequences. To this end, we need that $\overline{C}$ also bounds the terms in the inner sum in (2.23) separately. In the case that $\epsilon < 1$, however, it is possible that $\log \det \Sigma_l$ is negative. Then, also $|\mathcal{O}_l| \log \det \Sigma_l$ would be negative and, therefore, an upper bound $\overline{C}$ for the whole expression, as in (2.27), might be too small to bound the individual terms in the inner sum since the terms $\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} dx$ and $\lambda \text{Per}(\mathcal{O}_l, \Omega)$ in (2.23) are non-negative, which would let them (partly) cancel with a negative $|\mathcal{O}_l| \log \det \Sigma_l$. Hence, $\overline{C}$ needs to be sufficiently large to bound the individual terms. This can be achieved, for instance, by using $\overline{C} + |\Omega| |\log \epsilon^{2L}|$ instead of $\overline{C}$ as the upper bound since $-|\Omega| |\log \epsilon^{2L}|$ is a lower bound of $|\mathcal{O}_l| \log \det \Sigma_l$ and would consequently always compensate a negative $|\mathcal{O}_l| \log \det \Sigma_l$ term. We assume that $\overline{C}$ is chosen large enough to be an upper bound also for the individual terms in the inner sum in (2.23).

The next step is the construction of a subsequence of the minimizing sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ and also an element $(\mathcal{O}^*, \mu^*, \Sigma^*) \in X_\epsilon$ such that the inequality in (2.21) is satisfied. The procedure to find the subsequence is the following: we go through

$$(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} = (\mathcal{O}_{n,1}, \dots, \mathcal{O}_{n,k}, \mu_{n,1}, \dots, \mu_{n,k}, \Sigma_{n,1}, \dots, \Sigma_{n,k})_{n \in \mathbb{N}}$$

component-wise and successively choose subsequences of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ such that the requirements on the sequences of the components are fulfilled. This means, we start with considering $(\mathcal{O}_{n,1})_{n \in \mathbb{N}}$ and select a subsequence of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ such that our requirements on $(\mathcal{O}_{n,1})_{n \in \mathbb{N}}$ are met. From that subsequence of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ we pick another subsequence such that the requirements on $(\mathcal{O}_{n,2})_{n \in \mathbb{N}}$ are satisfied. Then, another subsequence of the previous subsequence is chosen such that $(\mathcal{O}_{n,3})_{n \in \mathbb{N}}$ behaves as we wish, and so on. Once we arrive at $(\mathcal{O}_{n,k})_{n \in \mathbb{N}}$, we continue with $(\Sigma_{n,1})_{n \in \mathbb{N}}$ until $(\Sigma_{n,k})_{n \in \mathbb{N}}$ and then $(\mu_{n,1})_{n \in \mathbb{N}}$ until $(\mu_{n,k})_{n \in \mathbb{N}}$ in the same manner. Since we always successively choose subsequences of the full sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$, we achieve that in the end, the sequence of every component meets the requirements. Please note that we denote every subsequence of the starting sequence again by $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ to simplify the notation.

When applying the direct method, the element $(\mathcal{O}^*, \mu^*, \Sigma^*)$ is usually obtained by finding convergent subsequences of the minimizing sequence and using their limits as this element. However, in our case, this will not be possible for every component. Therefore, we differ slightly from the usual way how the direct method is applied. If it is not possible to choose a convergent subsequence, we will only select a subsequence and an element such that the equation in (2.21) is satisfied.

To establish the inequality, we proceed as described and first show that for $l \in \{1, \dots, k\}$ the sequence $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ is bounded since it is a part of the minimizing sequence and therefore admits the selection of a convergent subsequence (in the notion of Definition 2.16). Instead of $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$, we equivalently work with the sequence of

characteristic functions

$$(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}} \subseteq \text{BV}(\Omega).$$

The sequence is in $\text{BV}(\Omega)$ because $\mathcal{O}_n = (\mathcal{O}_{n,1}, \dots, \mathcal{O}_{n,k}) \in \mathcal{A}$ for every $n \in \mathbb{N}$ implying $\text{Per}(\mathcal{O}_{n,l}, \Omega) < \infty$ and by Proposition 2.7. We observe that $\partial\Omega$ is a compact set because of its boundedness since Ω is bounded, and, additionally, its closedness because it can be written as

$$\partial\Omega = \overline{\Omega} \cap \overline{\mathbb{R}^d \setminus \Omega}$$

and is, therefore, equal to the intersection of two closed sets. Together with the assumption in this thesis that Ω has a Lipschitz boundary, it follows from Proposition 2.15 that Ω is an extension domain. The sequence $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ is part of the minimizing sequence for any $l \in \{1, \dots, k\}$ and we therefore obtain for any $n \in \mathbb{N}$ that $\lambda \text{Per}(\mathcal{O}_{n,l}, \Omega) \leq \overline{C}$ by (2.27) after possibly increasing λ . This implies

$$\begin{aligned} \|\chi_{\mathcal{O}_{n,l}}\|_{\text{BV}(\Omega)} &= \|\chi_{\mathcal{O}_{n,l}}\|_{L^1(\Omega)} + |\chi_{\mathcal{O}_{n,l}}|_{\text{BV}(\Omega)} \\ &\stackrel{\text{Definition 2.6}}{=} \int_{\Omega} |\chi_{\mathcal{O}_{n,l}}(x)| \, dx + \text{Per}(\mathcal{O}_{n,l}, \Omega) \\ &\leq |\Omega| + \frac{\overline{C}}{\lambda} \\ &< \infty, \end{aligned}$$

due to $|\chi_{\mathcal{O}_{n,l}}(x)| \leq 1$ for any $x \in \Omega$ and the boundedness of Ω , and we get the boundedness of $\|\chi_{\mathcal{O}_{n,l}}\|_{\text{BV}(\Omega)}$. This allows us to apply Theorem 2.17 and it follows the existence of a subsequence $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ converging weakly* with respect to Definition 2.16 in $\text{BV}(\Omega)$ to a function $u_l^* \in \text{BV}(\Omega)$. Moreover, it follows by Definition 2.16 that the chosen subsequence $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ is even convergent in $L^1(\Omega)$ to u_l^* . By Lemma 2.18, we can select a subsequence from this sequence that is point-wise a.e. convergent to u_l^* . Since $\chi_{\mathcal{O}_{n,l}}(x) \in \{0, 1\}$ for every $n \in \mathbb{N}$ and $\chi_{\mathcal{O}_{n,l}}(x) \rightarrow u_l^*(x)$ in $\mathbb{R}$, we obtain that $u_l^*(x) \in \{0, 1\}$ for almost every $x \in \Omega$. We set $\mathcal{O}_l^* := (u_l^*)^{-1}(\{1\})$. This set is measurable because it is the preimage of the measurable set $\{1\}$ under the measurable function u_l^* , which is measurable since it is contained in $\text{BV}(\Omega)$. We get $\chi_{\mathcal{O}_l^*} = u_l^*$ almost everywhere.

Once we have chosen a convergent subsequence of $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ with limit $\mathcal{O}_l^*$ for every $l \in \{1, \dots, k\}$, we have to show that $\mathcal{O}^* = (\mathcal{O}_1^*, \dots, \mathcal{O}_k^*)$ is indeed a partition of Ω to conclude that $\mathcal{O}^* \in \mathcal{A}$. First, we observe that

$$S := \left\{ (a_1, \dots, a_k) \in \mathbb{R}^k \left| \sum_{l=1}^k a_l = 1 \right. \right\}$$

is a closed set. Moreover, we have that the sequence $(\chi_{\mathcal{O}_{n,1}}(x), \dots, \chi_{\mathcal{O}_{n,k}}(x))_{n \in \mathbb{N}}$ is a sequence in S and convergent to $(\chi_{\mathcal{O}_1^*}(x), \dots, \chi_{\mathcal{O}_k^*}(x))$ for almost every $x \in \Omega$ because we have chosen the subsequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ in such a way that for every $l \in \{1, \dots, k\}$ we have point-wise a.e. convergence of $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$. The closedness of S yields $(\chi_{\mathcal{O}_1^*}(x), \dots, \chi_{\mathcal{O}_k^*}(x)) \in S$ almost everywhere, and therefore the partition property of $\mathcal{O}^*$ for Ω follows also almost everywhere. This means that $\mathcal{O}^*$ is not guaranteed to be in $\mathcal{A}$ which is what we need, but we can conclude that it is contained in a superset of $\mathcal{A}$, namely

$$\mathcal{A}' := \left\{ (\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \left| \forall l \in \{1, \dots, k\} : \mathcal{O}_l \subseteq \Omega \text{ measurable,} \right. \right. \\ \left. \left. \text{Per}(\mathcal{O}_l, \Omega) < \infty, \forall m \neq l : |\mathcal{O}_l \cap \mathcal{O}_m| = 0, \left| \Omega \setminus \bigcup_{l=1}^k \mathcal{O}_l \right| = 0 \right\}.$$

This set requires that the partition condition of $\mathcal{A}$ only holds almost everywhere. To achieve that $\mathcal{O}^* \in \mathcal{A}$, we can modify $\mathcal{O}^*$ for every $x \in \Omega$ where we do not have point-wise convergence since these elements are contained in a null set, which means that changing $\mathcal{O}^*$ for these elements does not change the value of J_ϵ . For instance, one can put every $x \in \Omega$ without point-wise convergence into $\mathcal{O}_1$ and remove it from $\mathcal{O}_2, \dots, \mathcal{O}_k$. Therefore, we suppose $\mathcal{O}^* \in \mathcal{A}$. With Proposition 2.19, we obtain for $(\mathcal{O}_n)_{n \in \mathbb{N}}$ and $\mathcal{O}^*$ the inequality

$$\text{Per}(\mathcal{O}_l^*, \Omega) \leq \liminf_{n \rightarrow \infty} \text{Per}(\mathcal{O}_{n,l}, \Omega) \text{ for every } l \in \{1, \dots, k\}. \quad (2.28)$$

We continue with the selection of suitable subsequence of $(\mu_n)_{n \in \mathbb{N}} \subseteq (\mathbb{R}^L)^k$ and $(\Sigma_n)_{n \in \mathbb{N}} \subseteq \mathcal{P}_\epsilon^k$. The sequence $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ and its corresponding sequence of characteristic functions $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ converge to $\mathcal{O}_l^*$ in measure and $\chi_{\mathcal{O}_l^*}$ in $L^1(\Omega)$ and point-wise almost everywhere, respectively, for every $l \in \{1, \dots, k\}$, as chosen before. A problem arises when $\mathcal{O}_l^*$ is a null set. In that case, the integral in J_ϵ (cf. Eq. (2.20)) is 0 in the limit and, consequently, we lose control over $(\mu_{n,l})_{n \in \mathbb{N}}$ and $(\Sigma_{n,l})_{n \in \mathbb{N}}$ when $n \rightarrow \infty$. Therefore, we have to distinguish two cases for each $l \in \{1, \dots, k\}$: the sequence $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ converges to $\mathcal{O}_l^*$ with (1) $|\mathcal{O}_l^*| > 0$, where we keep control over $(\mu_{n,l})_{n \in \mathbb{N}}$ and $(\Sigma_{n,l})_{n \in \mathbb{N}}$, and (2) $|\mathcal{O}_l^*| = 0$, where we cannot control the sequences. Let $l \in \{1, \dots, k\}$.

An inequality that we need in both cases is

$$\begin{aligned} \log \det \Sigma_{n,l} &= \log \prod_{i=1}^L \sigma^2(\Sigma_{n,l})_i \\ &\geq \log \left(\sigma^2(\Sigma_{n,l})_1 [\sigma^2(\Sigma_{n,l})_L]^{L-1} \right) \\ &\geq \log \sigma^2(\Sigma_{n,l})_1 + \log \epsilon^{2(L-1)}, \end{aligned} \quad (2.29)$$

which holds for every $n \in \mathbb{N}$ since $\log$ is monotonically increasing and $\Sigma_{n,l} \in \mathcal{P}_\epsilon$.

Case (1): The boundedness of the minimizing sequence (cf. (2.27)) yields

$$|\mathcal{O}_{n,l}| \log \det \Sigma_{n,l} \leq \overline{C} \text{ for all } n \in \mathbb{N}.$$

We have chosen $(\mathcal{O}_n)_{n \in \mathbb{N}}$ such that $\chi_{\mathcal{O}_{n,l}} \rightarrow \chi_{\mathcal{O}_l^*}$ in $L^1(\Omega)$. By Propositions 2.10 and 2.11, this implies that $|\mathcal{O}_{n,l}| \xrightarrow{n \rightarrow \infty} |\mathcal{O}_l^*| =: c_l$, and $c_l > 0$ by assumption of case (1). Then, there is a fixed $\delta \in (0, c_l)$ and $N_l \in \mathbb{N}$ that imply for $n \geq N_l$

$$0 < c_l - \delta \leq |\mathcal{O}_{n,l}| \leq c_l + \delta. \quad (2.30)$$

We remove the first $N_l - 1$ elements of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ to obtain a sequence for which (2.30) is true for every $n \in \mathbb{N}$. In the case where $\log \det \Sigma_{n,l} \geq 0$, it follows by (2.29)

$$\overline{C} \geq |\mathcal{O}_{n,l}| \log \det \Sigma_{n,l} \geq (c_l - \delta) \log \sigma^2(\Sigma_{n,l})_1 + (c_l - \delta) \log \epsilon^{2(L-1)},$$

with the consequence that

$$\frac{\overline{C}}{c_l - \delta} - \log \epsilon^{2(L-1)} \geq \log \sigma^2(\Sigma_{n,l})_1$$

and therefore

$$\infty > \exp \left(\frac{\overline{C}}{c_l - \delta} - \log \epsilon^{2(L-1)} \right) \geq \sigma^2(\Sigma_{n,l})_1.$$

In the other case, i.e., when $\log \det \Sigma_{n,l} < 0$, it follows analogously from

$$\overline{C} \geq |\mathcal{O}_{n,l}| \log \det \Sigma_{n,l} \geq (c_l + \delta) \log \det \Sigma_{n,l},$$

that

$$\infty > \exp \left(\frac{\overline{C}}{c_l + \delta} - \log \epsilon^{2(L-1)} \right) \geq \sigma^2(\Sigma_{n,l})_1,$$

and we obtain in both cases that $(\Sigma_{n,l})_{n \in \mathbb{N}}$ is bounded with respect to the spectral norm $\|\cdot\|_{2 \rightarrow 2}$ that is equal to the absolute value of the largest eigenvalue for symmetric matrices. Therefore,

$$(\Sigma_{n,l})_{n \in \mathbb{N}} \subseteq \overline{B_r^{\|\cdot\|_{2 \rightarrow 2}}(0)} \cap \mathcal{P}_\epsilon \subseteq \mathbb{R}^{L \times L},$$

for a sufficiently large $r > 0$. The set $\overline{B_r^{\|\cdot\|_{2 \rightarrow 2}}(0)} \cap \mathcal{P}_\epsilon$ is bounded and closed since the closed ball and $\mathcal{P}_\epsilon$ are closed (cf. Lemma 2.13), and so is their intersection. The

Heine-Borel theorem yields its compactness and, consequently, the existence of a convergent subsequence $(\Sigma_{n,l})_{n \in \mathbb{N}} \subseteq \mathcal{P}_\epsilon$ with limit $\Sigma_l^* \in \mathcal{P}_\epsilon$.

We continue in case (1) with the selection of a convergent subsequence of $(\mu_{n,l})_{n \in \mathbb{N}} \subseteq \mathbb{R}^L$. To this end, we observe that the reverse triangle inequality yields for each $x \in \Omega$

$$\|\mu_{n,l} - g(x)\|_2 \geq \left| \|\mu_{n,l}\|_2 - \|g(x)\|_2 \right| \geq \|\mu_{n,l}\|_2 - \|g(x)\|_2.$$

This fact together with (2.27) and $\sigma^2(\Sigma_{n,l}^{-1})_i = 1/\sigma^2(\Sigma_{n,l})_{L-i+1}$ for every $i \in \{1, \dots, L\}$ implies

$$\begin{aligned} \overline{C} &\geq \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx \\ &\geq \int_{\mathcal{O}_{n,l}} \sqrt{(g(x) - \mu_{n,l})^T \Sigma_{n,l}^{-1} (g(x) - \mu_{n,l})} \, dx \\ &\stackrel{\text{Lemma 2.14}}{\geq} \int_{\mathcal{O}_{n,l}} \sqrt{\frac{1}{\sigma^2(\Sigma_{n,l})_1}} \|g(x) - \mu_{n,l}\|_2 \, dx \\ &= \int_{\mathcal{O}_{n,l}} \sqrt{\frac{1}{\sigma^2(\Sigma_{n,l})_1}} \|\mu_{n,l} - g(x)\|_2 \, dx \\ &\geq \sqrt{\frac{1}{\sigma^2(\Sigma_{n,l})_1}} \left(\int_{\mathcal{O}_{n,l}} \|\mu_{n,l}\|_2 \, dx - \int_{\mathcal{O}_{n,l}} \|g(x)\|_2 \, dx \right) \\ &\geq \sqrt{\frac{1}{\sigma^2(\Sigma_{n,l})_1}} \left(|\mathcal{O}_{n,l}| \|\mu_{n,l}\|_2 - |\mathcal{O}_{n,l}| \|g(\cdot)\|_{L^\infty(\mathcal{O}_{n,l})} \right) \\ &\geq \sqrt{\frac{1}{\sigma^2(\Sigma_{n,l})_1}} \left(|\mathcal{O}_{n,l}| \|\mu_{n,l}\|_2 - |\Omega| \|g(\cdot)\|_{L^\infty(\Omega)} \right). \end{aligned}$$

The last inequality used that $|\mathcal{O}_{n,l}| \leq |\Omega|$ since $\mathcal{O}_{n,l} \subseteq \Omega$. From this inequality, we deduce together with (2.30)

$$\infty > \sqrt{\sigma^2(\Sigma_{n,l})_1} \overline{C} + |\Omega| \|g(\cdot)\|_{L^\infty(\Omega)} \geq |\mathcal{O}_{n,l}| \|\mu_{n,l}\|_2 \geq (c_l - \delta) \|\mu_{n,l}\|_2,$$

where the boundedness of the left-hand side follows because $(\Sigma_{n,l})_{n \in \mathbb{N}}$ is chosen such that $\|\Sigma_{n,l}\|_{2 \rightarrow 2} \leq r$ for every $n \in \mathbb{N}$, Ω is bounded and $g \in L^\infty(\Omega, \mathbb{R}^L)$. Division by $(c_l - \delta)$ yields the boundedness of $(\mu_{n,l})_{n \in \mathbb{N}}$ with respect to $\|\cdot\|_2$. Using the Heine-Borel theorem as above, we select a convergent subsequence $(\mu_{n,l})_{n \in \mathbb{N}}$ with limit $\mu_l^* \in \mathbb{R}^L$.

Because two sequences $(a_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ and $(b_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ satisfy

$$\liminf_{n \rightarrow \infty} a_n + \liminf_{n \rightarrow \infty} b_n \leq \liminf_{n \rightarrow \infty} a_n + b_n, \quad (2.31)$$

it is enough to establish an inequality like (2.21) for the summands of the outer sum in (2.20). To this end, we first note that products and sums of point-wise a.e. convergent sequences also converge point-wise a.e. Furthermore, we make use of the continuity of the transpose and the inverse of a matrix as well as the square root, determinant and logarithm. Considering the l -th summand of the outer sum of J_ϵ while having in mind that $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$ converges in measure in Ω to $\mathcal{O}_l^*$ (the characteristic functions point-wise a.e. in $\mathbb{R}$), $(\mu_{n,l})_{n \in \mathbb{N}} \subseteq \mathbb{R}^L$ to $\mu_l^* \in \mathbb{R}^L$ and $(\Sigma_{n,l})_{n \in \mathbb{N}} \subseteq \mathcal{P}_\epsilon$ to $\Sigma_l^* \in \mathcal{P}_\epsilon$, we deduce

$$\begin{aligned}
& \int_{\mathcal{O}_l^*} \|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} + \log \det \Sigma_l^* \, dx + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\
&= \int_{\Omega} \chi_{\mathcal{O}_l^*}(x) \left(\|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} + \log \det \Sigma_l^* \right) \, dx + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\
&= \int_{\Omega} \lim_{n \rightarrow \infty} \chi_{\mathcal{O}_{n,l}}(x) \left(\sqrt{\left(g(x) - \lim_{n \rightarrow \infty} \mu_{n,l} \right)^T \left(\lim_{n \rightarrow \infty} \Sigma_{n,l} \right)^{-1} \left(g(x) - \lim_{n \rightarrow \infty} \mu_{n,l} \right)} + \eta \right. \\
&\quad \left. + \log \det \lim_{n \rightarrow \infty} \Sigma_{n,l} \right) \, dx + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\
&= \int_{\Omega} \lim_{n \rightarrow \infty} \chi_{\mathcal{O}_{n,l}}(x) \lim_{n \rightarrow \infty} \left(\|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} + \log \det \Sigma_{n,l} \right) \, dx + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\
&= \int_{\Omega} \lim_{n \rightarrow \infty} \chi_{\mathcal{O}_{n,l}}(x) \left(\|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}, \eta} + \log \det \Sigma_{n,l} \right) \, dx + \lambda \operatorname{Per} \left(\lim_{n \rightarrow \infty} \mathcal{O}_{n,l}, \Omega \right) \\
&\leq \liminf_{n \rightarrow \infty} \int_{\Omega} \chi_{\mathcal{O}_{n,l}}(x) \left(\|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}, \eta} + \log \det \Sigma_{n,l} \right) \, dx + \liminf_{n \rightarrow \infty} \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) \\
&\leq \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}, \eta} + \log \det \Sigma_{n,l} \, dx + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega). \tag{2.32}
\end{aligned}$$

In the second-last inequality, we used (2.28) and Fatou's lemma (cf. Theorem 2.20) for measurable and possibly negative functions with a lower bound in $L^1(\Omega)$. The lower bound is given with a similar argumentation as in (2.24) and (2.25) as $-|\log \epsilon^{2L}|$, which is integrable since Ω is bounded. The last inequality follows from (2.31). This means that in case (1) where $|\mathcal{O}_l^*| > 0$, we have found a convergent subsequence of the minimizing sequence and a corresponding element such that the inequality in (2.21) can be established.

Case (2): $|\mathcal{O}_l^*| = 0$. It was already mentioned above that since in this case $|\mathcal{O}_l^*| = 0$, we do not have control over $(\Sigma_{n,l})_{n \in \mathbb{N}}$ and $(\mu_{n,l})_{n \in \mathbb{N}}$ in the limit and will therefore only choose a subsequence and elements $\mu_l^* \in \mathbb{R}^L$ and $\Sigma_l^* \in \mathcal{P}_\epsilon$ such that the inequality in (2.21) is fulfilled when the existence of a convergent subsequence cannot be guaranteed in general. First, we set $\mu_{n,l} := 0 \in \mathbb{R}^L$ and $\Sigma_{n,l} := \epsilon^2 I \in \mathcal{P}_\epsilon$ for every $n \in \mathbb{N}$ where $|\mathcal{O}_{n,l}| = 0$. This is possible because it does not change the functional value and the

established properties of the chosen sequence stay intact.

Consider the term $|\mathcal{O}_{n,l}| \log \det \Sigma_{n,l}$ in (2.23). By assumption we have $|\mathcal{O}_{n,l}| \rightarrow 0$ for $n \rightarrow \infty$ and similar to (2.25) we get $\log \det \Sigma_{n,l} \geq 2L \log \epsilon > -\infty$ for any $n \in \mathbb{N}$. The sequence $(\log \det \Sigma_{n,l})_{n \in \mathbb{N}}$ is now either (a) bounded above or (b) has no upper bound. If we are in case (a), then there is a constant $\alpha < \infty$ such that $\log \det \Sigma_{n,l} \leq \alpha$ for any $n \in \mathbb{N}$. We use (2.29) to derive

$$\alpha \geq \log \det \Sigma_{n,l} \geq \log \sigma^2(\Sigma_{n,l})_1 + \log \epsilon^{2(L-1)},$$

and it directly follows

$$\infty > \exp \left[\alpha - \log \epsilon^{2(L-1)} \right] \geq \sigma^2(\Sigma_{n,l})_1.$$

Hence, also here, the boundedness of $(\Sigma_{n,l})_{n \in \mathbb{N}}$ in the spectral norm holds and we can again apply the Heine-Borel theorem that allows for the selection of a subsequence converging to $\Sigma_l^* \in \mathcal{P}_\epsilon$. This implies together with Fatou's lemma (cf. Theorem 2.20) and the integrable lower bound $-\log \epsilon^{2L}$, as before in (2.32),

$$\begin{aligned} |\mathcal{O}_l^*| \log \det \Sigma_l^* &= \int_{\Omega} \chi_{\mathcal{O}_l^*}(x) \log \det \Sigma_l^* \, dx \\ &= \int_{\Omega} \lim_{n \rightarrow \infty} \chi_{\mathcal{O}_{n,l}}(x) \lim_{n \rightarrow \infty} \log \det \Sigma_{n,l} \, dx \\ &= \int_{\Omega} \lim_{n \rightarrow \infty} \chi_{\mathcal{O}_{n,l}}(x) \log \det \Sigma_{n,l} \, dx \\ &\leq \liminf_{n \rightarrow \infty} \int_{\Omega} \chi_{\mathcal{O}_{n,l}}(x) \log \det \Sigma_{n,l} \, dx. \end{aligned}$$

As above, the continuity of $\log$ and $\det$ as well as that products of point-wise a.e. convergent sequences converge point-wise a.e. were used.

In case (b), where the sequence $(\log \det \Sigma_{n,l})_{n \in \mathbb{N}}$ is not bounded from above, we cannot hope to find a convergent subsequence of $(\Sigma_{n,l})_{n \in \mathbb{N}}$ in general. Therefore, we will deviate from the usual way of applying the direct method and only select a subsequence $(\Sigma_{n,l})_{n \in \mathbb{N}}$ that can be non-convergent and an element $\Sigma_l^* \in \mathcal{P}_\epsilon$ which allow us to establish the inequality in (2.21). We select a subsequence $(\Sigma_{n,l})_{n \in \mathbb{N}}$ such that the sequence $(\log \det \Sigma_{n,l})_{n \in \mathbb{N}}$ is non-negative, monotonically increasing and diverging to ∞ . Additionally, we set $\Sigma_l^* := \epsilon^2 I \in \mathcal{P}_\epsilon$ and get

$$\underbrace{|\mathcal{O}_l^*|}_{=0} \log \det \Sigma_l^* = 0 \leq \liminf_{n \rightarrow \infty} \int_{\Omega} \underbrace{\chi_{\mathcal{O}_{n,l}}(x)}_{\geq 0} \underbrace{\log \det \Sigma_{n,l}}_{\geq 0} \, dx.$$

Since the integrand on the right-hand side is non-negative, the limit inferior of the sequence of integrals has to be non-negative, too.

It remains to deal with $(\mu_{n,l})_{n \in \mathbb{N}}$ and μ_l^* . Since μ_l only appears in the integral in (2.23), we concentrate on that term. As before, there is no chance to find a convergent subsequence in general, but since the integral is obviously non-negative for any $n \in \mathbb{N}$ because of its positive integrand (cf. (2.24)), it directly follows with $\mu_l^* := 0 \in \mathbb{R}^L$

$$\int_{\mathcal{O}_l^*} \|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} \, dx = 0 \leq \liminf_{n \rightarrow \infty} \underbrace{\int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx}_{\geq 0}.$$

In particular, this inequality is true, independent of the choice of $(\mu_{n,l})_{n \in \mathbb{N}}$ and $(\Sigma_{n,l})_{n \in \mathbb{N}}$, and we do not need to modify the sequences further.

We can now put together the shown inequalities for the selected subsequences in the case of $|\mathcal{O}_l^*| = 0$ and use (2.28), which yields

$$\begin{aligned} & \int_{\mathcal{O}_l^*} \|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} \, dx + |\mathcal{O}_l^*| \log \det \Sigma_l^* + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\ & \leq \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx + \liminf_{n \rightarrow \infty} \int_{\Omega} \chi_{\mathcal{O}_{n,l}}(x) \log \det \Sigma_{n,l} \, dx \\ & \quad + \liminf_{n \rightarrow \infty} \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) \\ & \leq \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx + \int_{\Omega} \chi_{\mathcal{O}_{n,l}}(x) \log \det \Sigma_{n,l} \, dx + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) \\ & = \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx + \log \det \Sigma_{n,l} \, dx + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega). \end{aligned}$$

The last inequality is a consequence of (2.31).

Summing up, we have selected a subsequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X_\epsilon$ of the minimizing sequence and an element $(\mathcal{O}^*, \mu^*, \Sigma^*) \in X_\epsilon$ such that for every $l \in \{1, \dots, k\}$ the inequality

$$\begin{aligned} & \int_{\mathcal{O}_l^*} \|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} \, dx + \log \det \Sigma_l^* + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \\ & \leq \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} \, dx + \log \det \Sigma_{n,l} + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) \end{aligned}$$

is fulfilled. We can now use (2.31) to show the inequality (2.21):

$$\begin{aligned} & J_\epsilon[\mathcal{O}^*, \mu^*, \Sigma^*] \\ & = \sum_{l=1}^k \left[\int_{\mathcal{O}_l^*} \|g(x) - \mu_l^*\|_{(\Sigma_l^*)^{-1}, \eta} \, dx + \log \det \Sigma_l^* + \lambda \operatorname{Per}(\mathcal{O}_l^*, \Omega) \right] \end{aligned}$$

$$\begin{aligned}
&\leq \liminf_{n \rightarrow \infty} \sum_{l=1}^k \left[\int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} + \log \det \Sigma_{n,l} \, dx + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) \right] \\
&= \liminf_{n \rightarrow \infty} J_\epsilon[\mathcal{O}_n, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n].
\end{aligned}$$

Consequently, the chain of inequalities in (2.22) is fulfilled and therefore, $(\mathcal{O}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ minimizes J_ϵ . $\square$

2.2.3 Γ -convergence

The regularization parameter ϵ for the covariance was introduced to ensure the positive definiteness of the covariance matrix estimate. However, in the previous section, we saw that it also played an important role when proving the existence of a minimizer for J_ϵ . In particular, a central argument was the closedness of $\mathcal{P}_\epsilon$ that is a consequence of the restriction of the eigenvalues by the introduction of ϵ .

Since the parameter is supposed to be chosen small, we will devote this subsection to the question about the consequences if $\epsilon \rightarrow 0$. In particular, because the parameter is data-independent in the sense that there is no rigorous mathematical dependence of ϵ on the data, it is highly difficult to choose in practice. A tool to answer that question and find out about the importance of ϵ for the model is Γ -convergence. It is a framework to analyze the convergence of sequences of functionals that also implies the convergence of minimizers of the functionals, always provided they exist and certain conditions are fulfilled. It is often used to better understand complex or degenerate problems by replacing them with a sequence of approximating problems which Γ -converge to the original problem but can be of a completely different type. Γ -convergence is a quite flexible framework in the sense that it does not require minimizers of the sequence of functionals to be computed and does not use any structure of the domains of the sequence of functionals or the Γ -limit functional. We will show that J_ϵ Γ -converges to a Γ -limit when $\epsilon \rightarrow 0$. In particular, under suitable conditions, Γ -convergence also implies the convergence of a sequence of minimizers to a minimizer of the Γ -limit if these minimizers exist. However, we will observe that in our case, we lose the guarantee of the existence of a minimizer in the Γ -limit by giving a counterexample. As a reference for the information about Γ -convergence given here, we refer to [14].

In the following, we assume that the domain of the considered functionals is a metric space. When we consider convergence of sequences in this space (notation: $x_n \rightarrow x$), we mean convergence in the metric of this space. The Γ -convergence for a sequence of functionals is defined as:

Definition 2.22 (Γ -convergence for sequences of functionals [14, Def. 1.5]).

Let (X, d_X) be a metric space, $f_n: X \rightarrow \overline{\mathbb{R}}$ for $n \in \mathbb{N}$ and $f_\infty: X \rightarrow \overline{\mathbb{R}}$. The sequence of functionals $(f_n)_{n \in \mathbb{N}}$ is said to Γ -converge to f_∞ if we have that for all $x \in X$ the following two conditions are fulfilled:

1. (*lim inf inequality*) for every sequence $(x_n)_{n \in \mathbb{N}} \subseteq X$ with $x_n \rightarrow x$ it holds

$$f_\infty[x] \leq \liminf_{n \rightarrow \infty} f_n[x_n],$$

2. (*lim sup inequality*) there exists a sequence $(x_n)_{n \in \mathbb{N}} \subseteq X$ with $x_n \rightarrow x$ such that

$$f_\infty[x] \geq \limsup_{n \rightarrow \infty} f_n[x_n].$$

In that case, the functional f_∞ is called Γ -limit of $(f_n)_{n \in \mathbb{N}}$. Notation: $f_\infty = \Gamma\text{-}\lim_{n \rightarrow \infty} f_n$.

Since the parameter ϵ is continuous, we are dealing with a family of functionals. In that case, Γ -convergence is defined by

Definition 2.23 (Γ -convergence for families of functionals [14, Def. 1.45]).

Let (X, d_X) be a metric space, $f_\epsilon: X \rightarrow \overline{\mathbb{R}}$ with $\epsilon > 0$ and $f_0: X \rightarrow \overline{\mathbb{R}}$. The family of functionals $(f_\epsilon)_{\epsilon > 0}$ is said to Γ -converge to f_0 if for any sequence $(\epsilon_n)_{n \in \mathbb{N}}$ with $\epsilon_n > 0$ and $\epsilon_n \rightarrow 0$, the sequence $(f_{\epsilon_n})_{n \in \mathbb{N}}$ Γ -converges to f_0 .

The notion of convergence that the metric d_X in X describes takes a crucial role when proving Γ -convergence. A stronger notion of convergence can be advantageous for proving the first condition in Definition 2.22 since there are fewer sequences for which the condition has to be fulfilled. On the other hand, a weaker notion of convergence is beneficial for proving the second condition in Definition 2.22 since a weaker notion of convergence admits more convergent sequences and it can be therefore easier to find a sequence that fulfills the second condition.

We see that the definitions of Γ -convergence require that the functionals are all defined on the same metric space X . Since J_ϵ is defined on X_ϵ , which varies with ϵ , this is not fulfilled for J_ϵ yet. However, defining X in our case as

$$X := \mathcal{A} \times (\mathbb{R}^L)^k \times \mathcal{P}^k,$$

we observe that $X_\epsilon \subseteq X$. The so-called *indicator functional*, that is described below in Remark 2.24, allows us to extend the domains of functionals to a larger set that contains their original domains and incorporate the original domains as constraints into the extended functionals such that the functionals keep their properties on their original domains. We will apply that trick to extend J_ϵ to X and solve the problem of different functional domains.

Remark 2.24 (Indicator functional [15, Ex. 6.23 3.]).

Let Y be a vector space. Define for a set $A \subseteq Y$ the functional

$$\delta_A[y] := \begin{cases} 0 & \text{if } y \in A, \\ \infty & \text{if } y \notin A. \end{cases} \quad (2.33)$$

This functional is often called indicator functional. In order to avoid confusion with the indicator functions of the MS functional, we will use the name only if necessary and in combination with its symbol δ_A .

The functional can be used to incorporate constraints of optimization problems into the objective functional to make the problem unconstrained. Let $J: Y \rightarrow \mathbb{R}$ be a functional and consider the optimization problem

$$\underset{y \in A}{\text{minimize}} \quad J[y]$$

for $A \subseteq Y$. Then, we can equivalently reformulate it to

$$\underset{y \in Y}{\text{minimize}} \quad J[y] + \delta_A[y].$$

With the preceding remark, we can extend J_ϵ using $\delta_{\mathcal{P}_\epsilon}$ to X , i.e., we extend the domain of the functional with respect to Σ . The functional with extended domain will not be stated here but later in (2.35). The reason is the notion of convergence that we choose on X that has to be described by the metric of the metric space. To this end, we have to restrict X with respect to $\mathcal{O}$. This restriction is discussed in the following before J_ϵ with the needed domain is stated, which then allows us to prove its Γ -convergence to a Γ -limit.

The notion of convergence defined by the metric of (X, d_X) is a key factor for the effectiveness of Γ -convergence and should ideally allow for a selection of convergent subsequences for sequences of minimizers. Please note that even though we need a metric for the metric space that induces the wanted notion of convergence, we will define the notion of convergence in X first and then argue that there exists a metric on X that describes the defined notion of convergence. We need compactness of the closed unit ball with respect to the convergence since then, every bounded sequence has a convergent subsequence. The second and third part of X , namely $(\mathbb{R}^L)^k$ and $\mathcal{P}^k$, are finite-dimensional, meaning that convergence in norm, weak and weak* convergence coincide and that the Heine-Borel theorem guarantees the compactness of the closed unit ball with respect to these notions of convergence. Hence, the notion of convergence in $BV(\Omega)$ defined by the chosen metric on that space is crucial. Since we have compactness in $BV(\Omega)$ (cf. Theorem 2.17) with respect to the notion of convergence defined in Definition 2.16, the natural choice for the notion of convergence in X is:

Definition 2.25 (Convergence in X).

Let $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X$ be a sequence. Then, $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ is said to converge weakly* to $(\mathcal{O}, \mu, \Sigma) \in X$ if the three conditions are fulfilled for each $l \in \{1, \dots, k\}$:

1. $\chi_{\mathcal{O}_{n,l}} \xrightarrow{*} \chi_{\mathcal{O}_l}$ in $\text{BV}(\Omega)$ with respect to Definition 2.16,
2. $\mu_{n,l} \rightarrow \mu_l$ in $\mathbb{R}^L$,
3. $\Sigma_{n,l} \rightarrow \Sigma_l$ in $\mathbb{R}^{L \times L}$,

if $n \rightarrow \infty$. We use the notation: $(\mathcal{O}_n, \mu_n, \Sigma_n) \xrightarrow{*} (\mathcal{O}, \mu, \Sigma)$.

Remark 2.26.

If the domain Ω of $\text{BV}(\Omega)$ is sufficiently regular, then weak* convergence in $\text{BV}(\Omega)$ defined in Definition 2.16 describes the usual weak* convergence with respect to a separable predual space of $\text{BV}(\Omega)$ [5, Rem. 3.12], which we do not further explain in this thesis as it is not needed for our purposes. Since in finite-dimensional spaces the notions of strong, weak and weak* convergence are equivalent, we would have in that case that the convergence in X in Definition 2.25 coincides with the usual weak* convergence in X . This motivates the description and notation in Definition 2.25.

Now, to extend the theory about the existence of minimizers of J_ϵ to the case where $\epsilon \rightarrow 0$ using Γ -convergence, we need to apply Γ -convergence with the notion of convergence defined in Definition 2.25. Since we consider a metric space in the context of Γ -convergence, one way to do so is to ensure that we choose a metric on X that describes this notion of convergence because then we have the sequential compactness of the unit ball and, therefore, can choose convergent subsequences of bounded sequences. For the finite-dimensional parts of X such a metric exists since strong, weak and weak* convergence are equivalent in this case and the strong convergence is a convergence in norm, which is equivalent to the notion of convergence described by the metric that is induced by the norm. For the infinite-dimensional part, the existence of such a metric can be guaranteed by Proposition A.16, which requires that we restrict our considerations to the closed ball around the origin of $\text{BV}(\Omega)$ for a constant $C_{\text{BV}} < \infty$, i.e., if we restrict our considerations to

$$\text{BV}(\Omega) \cap B_{C_{\text{BV}}}^{\|\cdot\|_{\text{BV}(\Omega)}}(0) = \left\{ u \in L^1(\Omega) \mid \|u\|_{\text{BV}(\Omega)} = \|u\|_{L^1(\Omega)} + |u|_{\text{BV}(\Omega)} \leq C_{\text{BV}} \right\}.$$

Proposition A.16 is applicable because it can be shown that BV is the dual space of a separable normed space [5, Rem. 3.12]. In particular, this restriction implies a bound on the TV of the functions in that set. Therefore, since we have $\text{Per}(\mathcal{O}, \Omega) = |\chi_{\mathcal{O}}|_{\text{BV}(\Omega)}$ by Definition 2.6 for a measurable set $\mathcal{O} \subseteq \mathbb{R}^d$, we also need to restrict $\mathcal{A}$ to account for that. To this end, we furthermore assume that C_{BV} is chosen sufficiently large

such that we have $C_{\text{BV}} - |\Omega| > 0$. Then, we define the restricted subset of $\mathcal{A}$ that we consider in the following as

$$\mathcal{A}_{C_{\text{BV}}} := \left\{ (\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \left| \begin{array}{l} \forall l \in \{1, \dots, k\} : \mathcal{O}_l \subseteq \Omega \text{ measurable,} \\ \text{Per}(\mathcal{O}_l, \Omega) \leq C_{\text{BV}} - |\Omega|, \forall m \neq l : \mathcal{O}_l \cap \mathcal{O}_m = \emptyset, \bigcup_{l=1}^k \mathcal{O}_l = \Omega \end{array} \right. \right\},$$

which differs from $\mathcal{A}$ (cf. (2.7)) only in the upper bound on $\text{Per}(\mathcal{O}_l, \Omega)$. It holds $\mathcal{A}_{C_{\text{BV}}} \subseteq \mathcal{A}$. Moreover, we find that for any $(\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \in \mathcal{A}_{C_{\text{BV}}}$, we obtain $\chi_{\mathcal{O}_l} \in \text{BV}(\Omega) \cap B_{C_{\text{BV}}}^{\|\cdot\|_{\text{BV}(\Omega)}}(0)$ for every $l \in \{1, \dots, k\}$ because

$$\begin{aligned} \|\chi_{\mathcal{O}_l}\|_{\text{BV}(\Omega)} &= \|\chi_{\mathcal{O}_l}\|_{L^1(\Omega)} + |\chi_{\mathcal{O}_l}|_{\text{BV}(\Omega)} \\ &\stackrel{\text{Definition 2.6}}{=} |\mathcal{O}_l| + \text{Per}(\mathcal{O}_l, \Omega) \\ &\leq |\mathcal{O}_l| + C_{\text{BV}} - |\Omega| \\ &\leq C_{\text{BV}}. \end{aligned}$$

The last inequality uses $|\mathcal{O}_l| \leq |\Omega|$ because $\mathcal{O}_l \subseteq \Omega$.

We define the metric space that we consider to prove Γ -convergence as

$$X_{C_{\text{BV}}} := \mathcal{A}_{C_{\text{BV}}} \times (\mathbb{R}^L)^k \times \mathcal{P}^k, \quad (2.34)$$

equipped with the metric that describes the notion of convergence defined in Definition 2.25 but restricting Definition 2.25 to $\text{BV}(\Omega) \cap B_{C_{\text{BV}}}^{\|\cdot\|_{\text{BV}(\Omega)}}(0)$ (instead of $\text{BV}(\Omega)$) and $X_{C_{\text{BV}}}$ (instead of X). Then, the explanations and construction above allow us to equivalently consider this notion of convergence instead of convergence in the chosen metric. We consider now J_ϵ defined on $X_{C_{\text{BV}}}$:

$$\begin{aligned} J_\epsilon : X_{C_{\text{BV}}} &\rightarrow \overline{\mathbb{R}}, \\ (\mathcal{O}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\mapsto \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \text{Per}(\mathcal{O}_l, \Omega) + \delta_{\mathcal{P}_\epsilon}(\Sigma_l) \right]. \end{aligned} \quad (2.35)$$

The behavior of the functional J_ϵ with respect to $\mathcal{O}$ and $\boldsymbol{\mu}$ remains the same. With respect to $\boldsymbol{\Sigma}$, it acts as before if $\boldsymbol{\Sigma} \in \mathcal{P}_\epsilon^k$, but outside of $\mathcal{P}_\epsilon^k$, i.e., if $\boldsymbol{\Sigma} \notin \mathcal{P}_\epsilon^k$, it is equal to ∞ . It follows that Theorem 2.21 also guarantees the existence of minimizers for this J_ϵ because the restriction to $\mathcal{P}_\epsilon$ from before is now built into the functional. The transfer of Theorem 2.21 to this setting here is possible because we chose $X_{C_{\text{BV}}}$ with an equivalent notion of convergence as we used for the proof of Theorem 2.21.

The restriction to $BV(\Omega) \cap B_{C_{BV}}^{\|\cdot\|_{BV(\Omega)}}(0)$ implies that we only consider sequences that are bounded with respect to $\mathcal{O}_l$ for every $l \in \{1, \dots, k\}$, which means that one can always select a weakly* convergent subsequence by Theorem 2.17. This restriction would simplify the existence proof but is actually not necessary since the reasoning in the proof of Theorem 2.21 did not use it and would still work in the setting of J_ϵ on $X_{C_{BV}}$.

Now, we can state the theorem about the Γ -convergence of J_ϵ :

Theorem 2.27 (Γ -convergence of J_ϵ).

The family of functionals $(J_\epsilon)_{\epsilon>0}$, defined in (2.35), Γ -converges to the Γ -limit

$$J_0: X_{C_{BV}} \rightarrow \overline{\mathbb{R}},$$

$$(\mathcal{O}, \mu, \Sigma) \mapsto \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \right], \quad (2.36)$$

if $\epsilon \rightarrow 0$ in regard to the weak* convergence in $X_{C_{BV}}$ (cf. Definition 2.25).

Proof. Let $(\epsilon_n)_{n \in \mathbb{N}}$ with $\epsilon_n > 0$ for any $n \in \mathbb{N}$ and $\epsilon_n \rightarrow 0$ when $n \rightarrow \infty$, and let $(\mathcal{O}, \mu, \Sigma) \in X_{C_{BV}}$. According to Definitions 2.22 and 2.23, it has to be shown that

- (1) (lim inf inequality): any sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X_{C_{BV}}$ weakly* converging to $(\mathcal{O}, \mu, \Sigma)$ fulfills

$$J_0[\mathcal{O}, \mu, \Sigma] \leq \liminf_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n],$$

- (2) (lim sup inequality): there exists a sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X_{C_{BV}}$ that weakly* converges to $(\mathcal{O}, \mu, \Sigma)$ and fulfills

$$J_0[\mathcal{O}, \mu, \Sigma] \geq \limsup_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n].$$

Please note that as before, when we select subsequences of sequences, we will simplify the notation by denoting these subsequences of a sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ again with $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$.

We establish the lim inf inequality in (1) first: let $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}} \subseteq X_{C_{BV}}$ be a sequence that weakly* converges to $(\mathcal{O}, \mu, \Sigma) \in X_{C_{BV}}$. Without loss of generality, we can make the assumption that the corresponding sequence of evaluations, namely $(J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n])_{n \in \mathbb{N}} \subseteq \overline{\mathbb{R}}$, converges to the limit inferior of this sequence, i.e.,

$$\lim_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n] = \liminf_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n], \quad (2.37)$$

because if this is not fulfilled, we can choose a subsequence of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ for which (2.37) holds.

From the definition of $\delta_{\mathcal{P}_\epsilon}$ in (2.33), it follows that $\delta_{\mathcal{P}_\epsilon}(\Sigma) \geq 0$ for arbitrary $\Sigma \in \mathcal{P}$ and $\epsilon > 0$. Therefore, we obtain for the last term in (2.35)

$$0 \leq \liminf_{n \rightarrow \infty} \delta_{\mathcal{P}_{\epsilon_n}}(\Sigma_{n,l}) \text{ for each } l \in \{1, \dots, k\}. \quad (2.38)$$

Because we assumed $(\mathcal{O}_n, \mu_n, \Sigma_n) \xrightarrow{*} (\mathcal{O}, \mu, \Sigma)$, we have that by definition $(\mu_{n,l})_{n \in \mathbb{N}}$ and $(\Sigma_{n,l})_{n \in \mathbb{N}}$ converge strongly to μ_l and Σ_l , respectively, for any $l \in \{1, \dots, k\}$. Only for $(\mathcal{O}_{n,l})_{n \in \mathbb{N}}$, we have to choose a subsequence to establish the $\liminf$ inequality. To this end, we follow the same strategy as in the proof of Theorem 2.21 and start with considering the first component $(\chi_{\mathcal{O}_{n,1}})_{n \in \mathbb{N}}$, choose a subsequence of $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ that fulfills the requirements, consider the second component $(\chi_{\mathcal{O}_{n,2}})_{n \in \mathbb{N}}$ and select another subsequence of the resulting subsequence such that also that components satisfy our requirements, and so on, until all sequences $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ behave as needed. Let $l \in \{1, \dots, k\}$. For $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ we have that $\chi_{\mathcal{O}_{n,l}} \xrightarrow{*} \chi_{\mathcal{O}_l}$ in $BV(\Omega) \cap B_{CBV}^{\|\cdot\|_{BV(\Omega)}}(0)$. This implies that $(\chi_{\mathcal{O}_{n,l}})_{n \in \mathbb{N}}$ converges to $\chi_{\mathcal{O}_l}$ in $L^1(\Omega)$ by Definition 2.16. Using Lemma 2.18, we can select a subsequence with point-wise a.e. convergence to $\chi_{\mathcal{O}_l}$. An analogous argumentation as in (2.32) yields for any $l \in \{1, \dots, k\}$

$$\begin{aligned} & \int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) \\ & \leq \liminf_{n \rightarrow \infty} \int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} + \log \det \Sigma_{n,l} \, dx + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega). \end{aligned}$$

Together with (2.31) and (2.38), it follows

$$\begin{aligned} J_0[\mathcal{O}, \mu, \Sigma] &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) + 0 \right] \\ &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx + \lambda \operatorname{Per}(\mathcal{O}_l, \Omega) + \delta_{\mathcal{P}}(\Sigma_l) \right] \\ &\leq \liminf_{n \rightarrow \infty} \sum_{l=1}^k \left[\int_{\mathcal{O}_{n,l}} \|g(x) - \mu_{n,l}\|_{\Sigma_{n,l}^{-1}, \eta} + \log \det \Sigma_{n,l} \, dx \right. \\ &\quad \left. + \lambda \operatorname{Per}(\mathcal{O}_{n,l}, \Omega) + \delta_{\mathcal{P}_{\epsilon_n}}(\Sigma_{n,l}) \right] \\ &\leq \liminf_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n]. \end{aligned}$$

Hence, the $\liminf$ inequality is fulfilled.

The second step is to establish (2) the lim sup inequality: we consider again $(\mathcal{O}, \mu, \Sigma) \in X_{CBV}$. The constant sequence $(\mathcal{O}_n, \mu_n, \Sigma_n)_{n \in \mathbb{N}}$ given as $(\mathcal{O}_n, \mu_n, \Sigma_n) = (\mathcal{O}, \mu, \Sigma)$ for every $n \in \mathbb{N}$ satisfies the lim sup inequality. To see this, we first note that this sequence is clearly weakly* convergent with limit $(\mathcal{O}, \mu, \Sigma)$. Let $l \in \{1, \dots, k\}$. Furthermore, we observe that $\mathcal{P}_{\epsilon''} \subseteq \mathcal{P}_{\epsilon'}$ when $\epsilon'' \geq \epsilon' > 0$. Hence, since $\Sigma_l \in \mathcal{P}$ and $\epsilon_n \rightarrow 0$ when $n \rightarrow \infty$, there is an $N \in \mathbb{N}$ such that for $n \geq N$ it follows that ϵ_n is sufficiently small such that we get $\Sigma_{n,l} = \Sigma_l \in \mathcal{P}_{\epsilon_n}$. Therefore, we have $\delta_{\mathcal{P}_{\epsilon_n}}(\Sigma_{n,l}) = \delta_{\mathcal{P}_{\epsilon_n}}(\Sigma_l) = 0$ for all $n \geq N$ and that implies

$$J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n] = J_{\epsilon_n}[\mathcal{O}, \mu, \Sigma] = J_0[\mathcal{O}, \mu, \Sigma].$$

Hence, the sequence $(J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n])_{n \in \mathbb{N}} \subseteq \overline{\mathbb{R}}$ converges to $J_0[\mathcal{O}, \mu, \Sigma]$ in $\overline{\mathbb{R}}$. We conclude

$$J_0[\mathcal{O}, \mu, \Sigma] = \lim_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n] = \limsup_{n \rightarrow \infty} J_{\epsilon_n}[\mathcal{O}_n, \mu_n, \Sigma_n],$$

and the lim sup inequality is established. $\square$

2.2.4 Importance of ϵ -regularization

The Γ -convergence of a family of functionals to a Γ -limit also implies the convergence of minimizers of that set of functionals to a minimizer of the Γ -limit if a certain coercivity condition holds and provided that the minimizers of the functionals exist. Coercivity is a property of a functional that describes a fast growth at the boundaries of the functional domain. By Theorem 2.21, the minimizers of the family of functionals exist in our case. However, as we will see in the following, the coercivity condition is violated in general by the functionals as the functionals do not grow at all domain boundaries. The regularization with ϵ compensates for that lack of coercivity by making the functional domains closed at these boundaries. But since the Γ -limit is not regularized anymore, minimizers do not exist in general. In particular, we will give a concrete counterexample where this is the case.

Coercivity can be defined as follows:

Definition 2.28 (Coercivity [14, Def. 1.19]).

Let (X, d_X) be a metric space. We call a function $f: X \rightarrow \overline{\mathbb{R}}$ mildly coercive if there exists a non-empty and compact set $K \subseteq X$ that fulfills

$$\inf_{x \in X} f[x] = \inf_{x \in K} f[x].$$

We call a function $f: X \rightarrow \overline{\mathbb{R}}$ coercive if for any $t \in \mathbb{R}$ the sublevel set

$$\{f \leq t\} := \{x \in X \mid f[x] \leq t\}$$

is precompact, i.e., its closure is compact. Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functionals with $f_n: X \rightarrow \overline{\mathbb{R}}$. This sequence is called *equi-mildly coercive* if it exists a non-empty and compact set $K \subseteq X$ that fulfills

$$\inf_{x \in X} f_n[x] = \inf_{x \in K} f_n[x] \quad \text{for every } n \in \mathbb{N}.$$

Please note that a coercive function is obviously also mildly coercive. The reverse is not true in general. An example of a mildly coercive but not coercive function is a periodic function [14, Rem. 1.20]. It has to be emphasized that this definition only requires X to be a metric space. The closure of a subset can only be built with respect to a larger set. As a consequence, the closure of a precompact subset still has to be a subset of X and cannot contain points that are not elements of X . This means in our case that if we have a sublevel set of J_ϵ or J_0 that is precompact in $X_{C_{BV}}$ (cf. Eq. (2.34)), its closure with respect to the part in $\mathcal{P}$ cannot contain any positive semidefinite matrix because these matrices are not contained in the considered metric space that only encompasses positive definite matrices. Hence, in the sense of the inherited notions of non-closedness and boundedness in $\mathcal{P}$ from $\mathbb{R}^{L \times L}$, a compact subset K of $X_{C_{BV}}$ has to be a proper subset of $X_{C_{BV}}$, i.e., $K \subsetneq X_{C_{BV}}$, and therefore, the coercivity requires a functional to show a fast increase when approaching the extremes of the functional domain. The mild coercivity weakens this concept but still requires that when leaving the compact subset K , the functional has to show sufficient growth such that the infimum lies inside of K .

Theorem 2.29 (Convergence of minima [14, Th. 1.21]).

Consider a metric space (X, d_X) and a sequence $(f_n)_{n \in \mathbb{N}}$ of equi-mildly coercive functionals on X . Let $f_\infty := \Gamma\text{-}\lim_{n \rightarrow \infty} f_n$ be the Γ -limit of $(f_n)_{n \in \mathbb{N}}$. Then, f_∞ attains its minimum on X and it holds

$$\min_{x \in X} f_\infty[x] = \lim_{n \rightarrow \infty} \inf_{x \in X} f_n[x].$$

We see that for a sequence of equi-mildly coercive functionals that converge to a Γ -limit, the Γ -limit attains its minimum. However, the family of functionals $(J_\epsilon)_{\epsilon > 0}$ is not equi-mildly coercive in general because the compact set $K \subseteq X_{C_{BV}}$ for the mild coercivity of J_ϵ can depend on ϵ . If that is the case for any $\epsilon > 0$, then it is impossible to find a non-empty and compact set for the equi-mild coercivity and consequently, Theorem 2.29 is not applicable. Of course, this does not mean that the Γ -limit does not have a minimum in general; however, we will now consider a counterexample where exactly this happens. In particular, we will see that the problem is that for certain images g , terms in the data term can vanish, which makes the functional non-coercive. This problem is circumvented by the ϵ -regularization that bounds the eigenvalues of the covariance matrices below and makes $\mathcal{P}_\epsilon$ closed. When $\epsilon \rightarrow 0$, i.e.,

in the case of J_0 , however, the functional domain X_{CBV} is not closed anymore with respect to Σ_l because $\mathcal{P}$ is not closed, with the consequence that we lose the existence guarantee of minimizers.

The culprits for the described behavior are the covariances in the functional. We can see in the proof of Theorem 2.21 that J_ϵ is bounded below with respect to $\mathcal{O} \in \mathcal{A}$ and $\boldsymbol{\mu} \in (\mathbb{R}^L)^k$ for fixed $\boldsymbol{\Sigma} \in (\mathcal{P})^k$, independent of ϵ . The same is true for J_0 . The term $\log \det \Sigma_l$ is the one that can cause trouble since it is not bounded below in general and where we used ϵ to ensure the boundedness from below. It is possible that this term and $\|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta}$ balance each other in certain situations such that the sum of these terms is bounded below without ϵ -regularization. We will look at that in Section 2.2.5. However, it can happen that some terms in $\log \det \Sigma_l$ are not present in $\|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta}$. Then, the former is able to push the functional value to $-\infty$. Let $l \in \{1, \dots, k\}$. For $(\mathcal{O}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \in X_{CBV}$, we consider the first two terms of the l -th summand in (2.36). The $\text{Per}(\mathcal{O}_l, \Omega)$ term is omitted as it does not play a role in the following considerations. In particular, we consider the eigendecomposition of $\Sigma_l = V_l D_l^2 V_l^T$ where $D_l^2 \in \mathbb{R}^{L \times L}$ is diagonal with the eigenvalues of Σ_l on its diagonal and $V_l \in O(L)$. Furthermore, denote the columns of V_l with $v_{l,1}, \dots, v_{l,L}$. As already discussed in Section 2.2.1, the eigendecomposition of Σ_l^{-1} is obtained as $\Sigma_l^{-1} = V_l (D_l^2)^{-1} V_l^T$. Then, we have

$$\begin{aligned}
& \int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx \\
&= \int_{\mathcal{O}_l} \sqrt{(g(x) - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g(x) - \mu_l) + \eta} \, dx + |\mathcal{O}_l| \log \left(\prod_{i=1}^L \sigma^2(\Sigma_l)_i \right) \\
&= \int_{\mathcal{O}_l} \sqrt{[V_l^T (g(x) - \mu_l)]^T (D_l^2)^{-1} [V_l^T (g(x) - \mu_l)] + \eta} \, dx + |\mathcal{O}_l| \log \left(\prod_{i=1}^L \sigma^2(\Sigma_l)_i \right) \\
&= \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \frac{[v_{l,i}^T (g(x) - \mu_l)]^2}{\sigma^2(\Sigma_l)_i}} + \eta \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i. \tag{2.39}
\end{aligned}$$

During minimization, the square root term aims to make $\sigma^2(\Sigma_l)_i$ for $i \in \{1, \dots, L\}$ large since this minimizes the fraction inside the square root, while $\log \det \Sigma_l$ aims to push $\sigma^2(\Sigma_l)_i$ towards 0 since then $\log \det \Sigma_l$ tends to $-\infty$. We will see in Section 2.2.5 that the two terms balance each other under certain conditions such that $\sigma^2(\Sigma_l)_i$ can neither become 0 nor ∞ where the minimum would be attained.

We will now consider an example where these two terms do not balance each other and therefore J_0 does not have a minimizer. We consider an image $g \in L^\infty(\Omega, \mathbb{R}^L)$ where $g(\Omega)$ is $(L - m)$ -dimensional for a fixed $m \in \{1, \dots, L - 1\}$. From the boundedness of Ω , it follows that $g \in L^2(\Omega, \mathbb{R}^L)$. We fix a segment $\mathcal{O}_l \subseteq \Omega$ and a mean

spectrum $\mu_l \in \mathbb{R}^L$ and choose an orthonormal basis for $\text{span}(g(\Omega))$. The operator span describes the span of a set of vectors. This basis can be completed to an orthonormal basis of $\mathbb{R}^L$, and we arrange the basis vectors as columns in a matrix $V_l \in \mathbb{R}^{L \times L}$. The basis vectors $v_{l,1}, \dots, v_{l,L-m} \in \mathbb{R}^L$ then form a basis of $\text{span}(g(\Omega))$ and basis vectors $v_{l,L-m+1}, \dots, v_{l,L} \in \mathbb{R}^L$ form a basis of $\text{span}(g(\Omega))^\perp$ where $A^\perp$ for a set $A \subseteq \mathbb{R}^L$ is the orthogonal complement of A . Furthermore, we select positive numbers $\sigma_{l,1}^2, \dots, \sigma_{l,L}^2 > 0$ and write them as the diagonal entries into a diagonal matrix $D_l^2 \in \mathbb{R}^{L \times L}$. Then, we can define the matrix $\Sigma_l := V_l D_l^2 V_l^T$, which is symmetric and positive definite and therefore contained in $\mathcal{P}$. It was already mentioned in Remark 2.8 that we can use this matrix to obtain an inner product $\langle \cdot, \cdot \rangle_{\Sigma_l^{-1}}$ on $\mathbb{R}^L$. For a vector $z \in \text{span}(g(\Omega))$, it holds

$$z^T V_l = \left(z^T v_{l,1}, \dots, z^T v_{l,L-m}, \underbrace{0, \dots, 0}_{m\text{-times}} \right)$$

since $v_{l,L-m+1}, \dots, v_{l,L} \in \mathbb{R}^L$ are elements of $\text{span}(g(\Omega))^\perp$. Moreover, we also have for every $i \in \{1, \dots, m\}$ that

$$V_l^T v_{l,L-m+i} = e_{L-m+i},$$

where the vector $e_{L-m+i} \in \mathbb{R}^L$ is the unit vector. We now use the orthogonal projection $\mathbb{P}: \mathbb{R}^L \rightarrow \mathbb{R}^L$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\Sigma_l^{-1}}$ onto the subspace spanned by $v_{l,L-m+1}, \dots, v_{l,L}$. Let $\mathcal{I}$ describe the identity operator. It follows together with the linearity of the inner product

$$\begin{aligned} & \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta}^2 \\ &= \|g(x) - (\mathcal{I} - \mathbb{P})\mu_l - \mathbb{P}\mu_l\|_{\Sigma_l^{-1}, \eta}^2 \\ &= \langle g(x) - (\mathcal{I} - \mathbb{P})\mu_l - \mathbb{P}\mu_l, g(x) - (\mathcal{I} - \mathbb{P})\mu_l - \mathbb{P}\mu_l \rangle_{\Sigma_l^{-1}} + \eta \\ &= \langle g(x) - (\mathcal{I} - \mathbb{P})\mu_l, g(x) - (\mathcal{I} - \mathbb{P})\mu_l \rangle_{\Sigma_l^{-1}} - \underbrace{2 \langle \mathbb{P}\mu_l, g(x) - (\mathcal{I} - \mathbb{P})\mu_l \rangle_{\Sigma_l^{-1}}}_{=0} \\ &\quad + \langle \mathbb{P}\mu_l, \mathbb{P}\mu_l \rangle_{\Sigma_l^{-1}} + \eta \\ &= \langle g(x) - (\mathcal{I} - \mathbb{P})\mu_l, g(x) - (\mathcal{I} - \mathbb{P})\mu_l \rangle_{\Sigma_l^{-1}} + \langle \mathbb{P}\mu_l, \mathbb{P}\mu_l \rangle_{\Sigma_l^{-1}} + \eta \\ &= \|g(x) - (\mathcal{I} - \mathbb{P})\mu_l\|_{\Sigma_l^{-1}, \eta}^2 + \langle \mathbb{P}\mu_l, \mathbb{P}\mu_l \rangle_{\Sigma_l^{-1}} \\ &\geq \|g(x) - (\mathcal{I} - \mathbb{P})\mu_l\|_{\Sigma_l^{-1}, \eta}^2. \end{aligned}$$

The mixed term in the fourth row is 0 because the left argument is in $\text{span}(g(\Omega))^\perp$ and the right argument in $\text{span}(g(\Omega))$. The inequality is a consequence of the positive

definiteness of an inner product that implies $\langle \mathbb{P}\mu_l, \mathbb{P}\mu_l \rangle_{\Sigma_l^{-1}} \geq 0$. Hence, a projection of μ_l onto the subspace including $g(\Omega)$ can only reduce the functional value of J_0 . We can, therefore, assume that $\mu_l \in \text{span}(g(\Omega))$ when we minimize J_0 . With (2.39), we obtain

$$\begin{aligned} & \int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx \\ &= \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \frac{[v_{l,i}^T(g(x) - \mu_l)]^2}{\sigma_{l,i}^2}} + \eta \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma_{l,i}^2 \\ &= \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^{L-m} \frac{[v_{l,i}^T(g(x) - \mu_l)]^2}{\sigma_{l,i}^2}} + \eta \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma_{l,i}^2. \end{aligned}$$

The last equality is a consequence of $\mu_l \in \text{span}(g(\Omega))$ and $v_{l,L-m+1}, \dots, v_{l,L} \in \text{span}(g(\Omega))^\perp$. What we can observe is that the eigenvalues $\sigma_{l,L-m+1}^2, \dots, \sigma_{l,L}^2$ vanished from the square root term. During minimization, we are now allowed to let these values approach 0, which lets the $\log \det \Sigma_l$ term decrease without increasing the square root term, and therefore makes the functional value tend to $-\infty$. Geometrically speaking, this makes sense since these values quantify the variation of the data in $\text{span}(g(\Omega))^\perp$, which is 0 by assumption. However, it causes problems. One problem is that the functional J_0 is not bounded from below because $\lim_{x \rightarrow 0^+} \log x = -\infty$. This can be fixed by, e.g., substituting $\log$ by $\log(\cdot + 1)$ because this function is bounded below by 0. However, the term would still benefit from letting the eigenvalues of Σ_l approach 0 and, therefore, make the eigenvalues leave the admissible set in the limit, resulting in Σ_l becoming only positive semidefinite. A minimizer would not exist in general. This is circumvented in the case of J_ϵ defined on X_ϵ (cf. (2.20)) by a restriction of the admissible set to the closed set $\mathcal{P}_\epsilon$. The scenario explained in the counterexample would then lead to eigenvalues

$$\sigma_{l,L-m+1}^2 = \dots = \sigma_{l,L}^2 = \epsilon^2$$

in the case of a minimization, and the closedness of $\mathcal{P}_\epsilon$ allows for finding a minimizer. In terms of the mild coercivity, the compact set K in Definition 2.28 with respect to Σ_l is a subset of $\mathcal{P}_\epsilon$ that is allowed to cover the boundary case where $\sigma_{l,i}^2 = \epsilon^2$ for some $i \in \{1, \dots, L\}$ because $\mathcal{P}_\epsilon$ is a closed subset of $\mathbb{R}^{L \times L}$. We obtain mild coercivity with respect to Σ_l in this case. The extension of J_ϵ with respect to Σ to $X_{C_{BV}}$ (cf. (2.35)) shifts the argument for the mild coercivity from the closedness of $\mathcal{P}_\epsilon$ to the addition of $\delta \mathcal{P}_\epsilon$, which lets the functional value explode when we try to leave $\mathcal{P}_\epsilon$ with Σ_l . In the case of the Γ -limit when $\epsilon \rightarrow 0$, however, neither of these two arguments

for the mild coercivity hold for J_0 , and it was shown in the example above that J_0 does not have a minimizer in general. This consideration also rules out coercivity since it would imply mild coercivity.

We have seen that the ϵ -regularization of the eigenvalues has several important consequences for the feasibility of the model: by bounding the eigenvalues from below, we obtain the invertibility of Σ_l , a lower bound for the functional, the closedness of the admissible set $\mathcal{P}_\epsilon$ or the coercivity of the functional when extended with respect to Σ to X_{CBV} and therefore a minimizer. In particular, this lower bound for the eigenvalues restricts the set of admissible covariance matrices and, hence, the search space for minimizers. Consequently, it potentially excludes minima of the functional. However, the choice of ϵ is difficult because it is a data-independent parameter that has to be selected before applying the model. Therefore, it would be of interest to work with a model that does not need such a parameter.

Two follow-up questions to this counterexample are: (1) can we find conditions on the image data g that ensure a data-dependent lower bound for the eigenvalues and, therefore, the feasibility of the model and the existence of minimizers, and (2) can we modify the model to make it coercive without an extra hyperparameter like ϵ such that the model is feasible and always admits a minimizer, regardless of any conditions on the data? The results of the next subsection can provide a starting point to answer these questions.

2.2.5 Lower bound for eigenvalues of covariance matrix

We have seen in Sections 2.2.2 to 2.2.4 that it is the eigenvalues of the covariance matrix that can cause problems in terms of the existence of minimizers and the feasibility of the model. We will now derive a positive lower bound for the eigenvalues when $\|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta}$ and $\log \det \Sigma_l$ balance each other, i.e., when $|\mathcal{O}_l| > 0$ and every eigenvalue is present in both terms, which requires $g(\Omega)$ to be L -dimensional. This lower bound can serve as a starting point to formulate conditions on the data and to find a modification of J_0 that is coercive in the eigenvalues of the covariance matrices.

We can observe that the summands of the outer sum in the definition of J_0 in (2.36) are uncoupled with respect to $\mathcal{O}_l$, μ_l and Σ_l and that we can therefore consider only the l -th summand with $l \in \{1, \dots, k\}$. Additionally, the eigenvalues of Σ_l only appear inside the integral and we can restrict our analysis to the integral of the l -th summand. A fact that we need is that a concave function

$$f: D \subseteq \mathbb{R} \rightarrow \mathbb{R}$$

fulfills for coefficients $a_1, \dots, a_n \geq 0$ with $\sum_{i=1}^n a_i = 1$ and points $x_1, \dots, x_n \in D$ the

inequality

$$f\left(\sum_{i=1}^n a_i x_i\right) \geq \sum_{i=1}^n a_i f(x_i). \quad (2.40)$$

For the matrix $\Sigma_l \in \mathcal{P}$, we have the eigendecomposition $\Sigma_l = V_l D_l^2 V_l^T$ with an orthogonal matrix $V \in O(L)$ and a diagonal matrix $D_l^2 \in \mathbb{R}^{L \times L}$. The columns of V_l are denoted by $v_{l,i}$ for $i \in \{1, \dots, L\}$ and the diagonal elements of D_l^2 are the eigenvalues of Σ_l . Starting with a reformulation as in (2.39), we derive

$$\begin{aligned} & \int_{\mathcal{O}_l} \|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \, dx \\ (2.39) \quad &= \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \frac{[v_{l,i}^T (g(x) - \mu_l)]^2}{\sigma^2(\Sigma_l)_i}} + \eta \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i \\ &\geq \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \frac{[v_{l,i}^T (g(x) - \mu_l)]^2}{\sigma^2(\Sigma_l)_i}} \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i \\ &= \int_{\mathcal{O}_l} \sqrt{L} \sqrt{\sum_{i=1}^L \frac{1}{L} [v_{l,i}^T (g(x) - \mu_l)]^2 \frac{1}{\sigma^2(\Sigma_l)_i}} \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i \\ (2.40) \quad &\geq \int_{\mathcal{O}_l} \frac{\sqrt{L}}{L} \sum_{i=1}^L \sqrt{[v_{l,i}^T (g(x) - \mu_l)]^2 \frac{1}{\sigma^2(\Sigma_l)_i}} \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i \\ &= \int_{\mathcal{O}_l} \frac{1}{\sqrt{L}} \sum_{i=1}^L |v_{l,i}^T (g(x) - \mu_l)| \frac{1}{\sqrt{\sigma^2(\Sigma_l)_i}} \, dx + |\mathcal{O}_l| \sum_{i=1}^L \log \sigma^2(\Sigma_l)_i \\ &= \sum_{i=1}^L \left(\frac{1}{\sqrt{L}} \int_{\mathcal{O}_l} |v_{l,i}^T (g(x) - \mu_l)| \, dx \frac{1}{\sqrt{\sigma^2(\Sigma_l)_i}} + |\mathcal{O}_l| \log \sigma^2(\Sigma_l)_i \right). \quad (2.41) \end{aligned}$$

We see that for each $i \in \{1, \dots, L\}$ the i -th eigenvalue $\sigma^2(\Sigma_l)_i$ only appears in the i -th summand of the outer sum. We can, therefore, analyze the summands of that sum independently. When interpreting the term in the i -th summand as a function in $\sigma^2(\Sigma_l)_i$, it is formulated as

$$h(x) := \frac{a}{\sqrt{x}} + b \log x, \quad (2.42)$$

with two real numbers $a \geq 0$ and $b > 0$. The function domain of h is $(0, \infty)$. If $a = 0$, h is not bounded below and hence, does not have a minimizer. We assume $a > 0$.

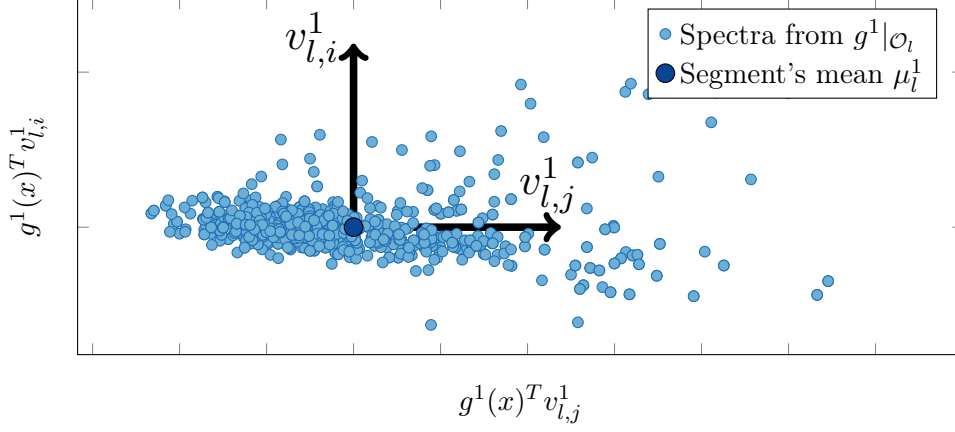
Then, the first and second derivative h' and h'' are given by

$$\begin{aligned} h'(x) &= -\frac{a}{2}x^{-\frac{3}{2}} + \frac{b}{x}, \\ h''(x) &= \frac{3a}{4}x^{-\frac{5}{2}} - \frac{b}{x^2}. \end{aligned}$$

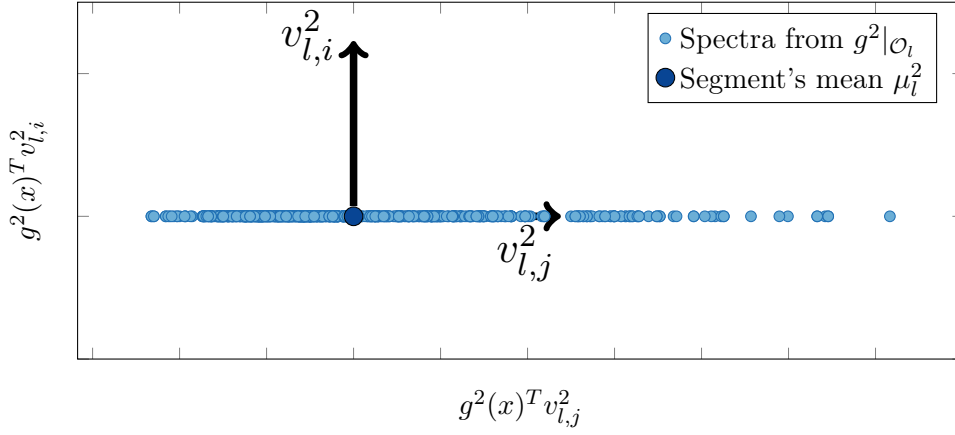
One can directly verify that $x^* = (\frac{a}{2b})^2$ is a local minimum since $h'(x^*) = 0$ and $h''(x^*) > 0$. This minimum is, in fact, global because $h'(x) < 0$ for $x < x^*$, i.e., the function is strictly monotonically decreasing for arguments smaller than x^* , and $h'(x) > 0$ for $x > x^*$, i.e., the function is strictly monotonically increasing for arguments larger than x^* . Replacing a and b by the corresponding expressions from (2.41), we find that the minimizing eigenvalue of the i -th summand is

$$\sigma^2(\Sigma_l)_i^* = \left(\frac{\int_{\mathcal{O}_l} |v_{l,i}^T (g(x) - \mu_l)| \, dx}{2\sqrt{L} |\mathcal{O}_l|} \right)^2. \quad (2.43)$$

We assumed that $|\mathcal{O}_l| > 0$; hence, the denominator is positive. The numerator, however, is not guaranteed to be positive. If $a = 0$, then also the numerator is 0, and $\sigma^2(\Sigma_l)_i^*$ does not exist because in that case the function in (2.42) does not have a minimizer. This happens when the vector $g(x) - \mu_l$ is orthogonal to $v_{l,i}$ for almost every $x \in \mathcal{O}_l$. The scenario corresponds to the case in the counterexample in Section 2.2.4 if we assume that $v_{l,i} \in \text{span}(g(\Omega))^\perp$. Hence, we must require that the integral in the numerator is always positive. Then, the eigenvalue is bounded away from 0. If we take a geometrical perspective, this condition means that we require sufficient deviation of the data $g(x)$ from a mean spectrum μ_l in the direction of $v_{l,i}$ for $x \in \mathcal{O}'_l \subseteq \mathcal{O}_l$ with $|\mathcal{O}'_l| > 0$. If we think of spectra of the pixels in segment $\mathcal{O}_l$, we ask the data to show sufficient variation of the spectra in the direction of $v_{l,i}$ for almost every $x \in \mathcal{O}_l$. This is what has caused the problem in the counterexample before: $g(\Omega)$ does not extend in every dimension of $\mathbb{R}^L$ and, therefore, there have to be directions $v_{l,i}$ in which there is no variation and, thus, the integral in the numerator in (2.43) is equal to 0 for these $i \in \{1, \dots, L\}$. Fig. 2.3 visualizes these findings. The spectra from the spectral distribution $g^1|_{\mathcal{O}_l}$ in Fig. 2.3a show variation along both directions $v_{l,i}^1$ and $v_{l,j}^1$. Consequently, for both directions the integral in the numerator in Eq. (2.43) is greater 0 and the corresponding eigenvalues of a minimizing Σ_l^1 are kept away from 0. Assuming that we also have variation in the directions of the remaining eigenvectors $v_{l,t}^1$, $t \in \{1, \dots, L\} \setminus \{i, j\}$, of Σ_l^1 , it follows $\Sigma_l^1 \in \mathcal{P}$. In the situation in Fig. 2.3b, we have variation only along the direction of eigenvector $v_{l,j}^2$. Therefore, the estimate for the eigenvalue describing the variation along $v_{l,i}^2$ equals 0. That implies $\Sigma_l^2 \notin \mathcal{P}$, i.e., a Σ_l^2 that would be optimal in terms of described data



(a) Sampled spectra from spectral distribution $g^1|_{\mathcal{O}_l}$, showing variation in both depicted directions.



(b) Sampled spectra from spectral distribution $g^2|_{\mathcal{O}_l}$, showing variation in only one direction.

Figure 2.3: Sampled spectra from two spectral distributions $g^m|_{\mathcal{O}_l}$ of a segment $\mathcal{O}_l \subseteq \Omega$ corresponding to images $g^m \in L^\infty(\Omega, \mathbb{R}^L)$, $m \in \{1, 2\}$, projected onto the i -th and j -th eigenvector $v_{l,i}^m \in \mathbb{R}^L$ and $v_{l,j}^m \in \mathbb{R}^L$ of the covariance matrix $\Sigma_l^m \in \mathbb{R}^{L \times L}$.

variation is not admissible. In other words, this means that $a = 0$ in (2.42), making the function unbounded from below. Therefore, no minimizer exists.

In practice, measured data can be located only in a low-dimensional subspace of the measured space. In that case, the model would be infeasible by this intrinsic property of the data. As a simple heuristic to circumvent this problem, we propose to use the PCA [88] as a preprocessing step and keep the principal components that account for at least 99.9% of the variance in the data. Then, there is a good chance

that the data shows variation in every direction such that we obtain $\sigma^2(\Sigma_l)_i^* > 0$ for every $i \in \{1, \dots, L\}$. An application of the MNF can have the same effect. However, there is no clear criterion, as in the case of the PCA telling us how many components have to be kept to describe almost the full variation of the data and avoid directions without any variation of the data. In the discrete case, it is enough to have for each segment and each direction in the spectral space one spectrum that deviates from the mean to ensure the boundedness of the eigenvalues.

2.3 The ϕ -regularized anisotropic Mumford–Shah model

We will now propose a modification of J_0 (cf. (2.36)) where we replace η under the square root in $\|g(x) - \mu_l\|_{\Sigma_l^{-1}, \eta}$ by a function that lets the functional value grow rapidly if an eigenvalue without variation in the data in the corresponding direction approaches 0. The idea is to check whether there is variation in the data along a certain direction. If this is the case, the new term adds only a small quantity close to 0 to the term under the square root. If there is no variation, the new extra term adds a certain quantity multiplied by the reciprocal of the corresponding eigenvalue such that the functional value diverges to ∞ when the eigenvalue approaches 0. This modification was originally meant as a direction- and variation-dependent regularization to make J_0 coercive with respect to the eigenvalues of the covariance matrices and without a data-independent parameter like ϵ that leads to a data-independent lower bound for the eigenvalues. However, as we will see, also this model leads to a data-independent lower bound for the eigenvalues that is, in this case, dependent on the choice of a function that is involved in the new regularization. Nevertheless, it is a step in the direction of a functional without fixed lower bound for the eigenvalues but feasibility independent of the data: It shows how extra terms under the square root in J_0 affect the lower bound for the eigenvalues (cf. (2.43)) and that it is possible to estimate the main directions of variation (the eigenvectors of the covariance matrices) and corresponding variances (eigenvalues of the covariance matrices) directly instead of optimizing the covariance matrices. This allows for more flexibility in the functional design since we can consider the eigenvectors and eigenvalues also independent of each other in the functional. Moreover, this model does not need the parameter η to ensure the differentiability with respect to the parameters appearing inside the square root because the positivity of the term under the square root is now guaranteed by the new regularizer, as we will see in formulas below. We now introduce the modification of J_0 in Section 2.3.1 and then show the existence of minimizers also for this model in Section 2.3.2.

2.3.1 Model description and lower bound of eigenvalues

We have seen in Section 2.2.5 that an eigenvalue of Σ_l can be pushed arbitrarily close to 0 by the functional when $|\mathcal{O}_l| > 0$ but the spectra showing variation in the direction of the corresponding eigenvector belong to pixels that are contained in a subset of $\mathcal{O}_l$ that is a null set. In that case, the integral in (2.41) is 0 and the eigenvalue is not bounded away from 0. In the case of ϵ AMS, the ϵ -regularization of the eigenvalues by clipping eigenvalues smaller than $\epsilon^2 > 0$ to ϵ^2 has forbidden such a behavior and defined a data-independent lower bound for the eigenvalues. We will go a different way now to bound the eigenvalues away from 0. In order to define the modified version of J_0 , which we call J_ϕ , let $\alpha, \beta > 0$ be two parameters and

$$\phi: [0, \infty) \rightarrow [0, \infty) \quad (2.44)$$

be a continuous and monotonically decreasing function such that

$$\phi(0) = 1 \text{ and } \lim_{t \rightarrow \infty} \phi(t) = 0. \quad (2.45)$$

The variable Σ_l of J_0 was defined over the symmetric positive definite matrices $\mathcal{P}$. However, we will have terms in J_ϕ that use explicitly the eigenvalues or eigenvectors of Σ_l . We will, therefore, equivalently represent the elements of $\mathcal{P}$ by their eigen-decompositions and define J_ϕ over the orthogonal matrices $O(L)$ and positive real numbers $\mathbb{R}_{>0}^L$. For ease of notation, we will use the following symbols:

$$\mathbf{V} := (V_1, \dots, V_k) \in (O(L))^k, \quad \boldsymbol{\sigma}^2 := (\sigma_1^2, \dots, \sigma_k^2) \in (\mathbb{R}_{>0}^L)^k.$$

Please note that in the previous sections, σ^2 was the mapping returning the eigenvalues of a matrix. Here, it changes slightly its meaning in the sense that $\sigma_l^2 \in \mathbb{R}_{>0}^L$, $l \in \{1, \dots, k\}$, is now a vector of values which are meant as the eigenvalues of the covariance matrix Σ_l of segment l , which was previously denoted by $\sigma^2(\Sigma_l)$. Let $g \in L^\infty(\Omega, \mathbb{R}^L)$ and define

$$X_\phi := \mathcal{A} \times (\mathbb{R}^L)^k \times (O(L))^k \times (\mathbb{R}_{>0}^L)^k.$$

Building on (2.39), we propose the following modified version of J_0 :

$$\begin{aligned} J_\phi: X_\phi &\rightarrow \mathbb{R}, \\ (\mathcal{O}, \boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\sigma}^2) &\mapsto J_\phi[\mathcal{O}, \boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\sigma}^2], \end{aligned} \quad (2.46)$$

and we compute $J_\phi [\mathcal{O}, \mu, V, \sigma^2]$ as

$$\begin{aligned} J_\phi [\mathcal{O}, \mu, V, \sigma^2] &:= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \left[[V_l^T (g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T (g(x) - \mu_l)]_i^2 \right) \right] \frac{1}{\sigma_{l,i}^2}} \right. \\ &\quad \left. + \beta \sum_{i=1}^L \log \sigma_{l,i}^2 \, dx + \lambda \operatorname{Per} (\mathcal{O}_l, \Omega) \right]. \end{aligned} \quad (2.47)$$

Analog to the naming of ϵ AMS, we call this model ϕ -regularized anisotropic Mumford–Shah model (short: ϕ AMS). Hence, the indicator function of segment $l \in \{1, \dots, k\}$ is in this case

$$\begin{aligned} f_{\phi,l} (x; \mu_l, V_l, \sigma_l^2) &:= \\ &\sqrt{\sum_{i=1}^L \left[[V_l^T (g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T (g(x) - \mu_l)]_i^2 \right) \right] \frac{1}{\sigma_{l,i}^2}} + \beta \sum_{i=1}^L \log \sigma_{l,i}^2 \end{aligned} \quad (2.48)$$

for each $x \in \Omega$. With $\Sigma_l := V_l D_l^2 V_l^T$ where $D_l^2 \in \mathbb{R}^{L \times L}$ is a diagonal matrix having σ_l^2 as diagonal entries, we can reformulate $J_\phi [\mathcal{O}, \mu, V, \sigma^2]$ as

$$\begin{aligned} J_\phi [\mathcal{O}, \mu, V, \sigma^2] &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L [V_l^T (g(x) - \mu_l)]_i^2 \frac{1}{\sigma_{l,i}^2} + \alpha \sum_{i=1}^L \phi \left([V_l^T (g(x) - \mu_l)]_i^2 \right) \frac{1}{\sigma_{l,i}^2}} \right. \\ &\quad \left. + \beta \log \det \Sigma_l \, dx + \lambda \operatorname{Per} (\mathcal{O}_l, \Omega) \right] \end{aligned} \quad (2.49)$$

$$\begin{aligned} &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \sqrt{(g(x) - \mu_l)^T \Sigma_l^{-1} (g(x) - \mu_l) + \alpha \sum_{i=1}^L \phi \left([V_l^T (g(x) - \mu_l)]_i^2 \right) \frac{1}{\sigma_{l,i}^2}} \right. \\ &\quad \left. + \beta \log \det \Sigma_l \, dx + \lambda \operatorname{Per} (\mathcal{O}_l, \Omega) \right]. \end{aligned} \quad (2.50)$$

We see from (2.50) that, in principle, J_ϕ coincides with J_0 (cf. (2.36)) when $\beta = 1$ and η in J_0 is replaced by the second summand involving ϕ in the square root. To understand what the regularization with ϕ does, we consider the square root term

in (2.49). The first summand equals the anisotropic distance of the spectrum $g(x)$ to the segment's mean μ_l , only that the vector-matrix-vector product is written in scalar form using the eigendecomposition of Σ_l . We used the same reformulation, for instance, in (2.39) and (2.41) to understand the behavior of the eigenvalues of Σ_l . Let us consider the i -th variance or eigenvalue $\sigma_{l,i}^2$ for $i \in \{1, \dots, L\}$ of segment $l \in \{1, \dots, k\}$. We have seen in Sections 2.2.4 and 2.2.5 that we need variation in the direction of the i -th eigenvector. This variation is quantified by $[V_l^T (g(x) - \mu_l)]_i$. In particular, if either $g(x) = \mu_l$ or $g(x) - \mu_l$ is orthogonal to the i -th eigenvector (i -th column of V_l), then $[V_l^T (g(x) - \mu_l)]_i = 0$. In this case, as before, only the $\log \det \Sigma_l$ term would see $\sigma_{l,i}^2$ if we did not have the second term under the square root in (2.49), which would result in the unboundedness of the functional. Now, the quantification $[V_l^T (g(x) - \mu_l)]_i$ of the variation of the data in the direction of the i -th eigenvector is input in ϕ . If there is no variation in that direction, the argument of ϕ equals 0. However, by construction of ϕ the function returns 1 in that case. Through the multiplication with $1/\sigma_{l,i}^2$, the eigenvalue remains visible for the term under the square root that benefits from making $\sigma_{l,i}^2$ large. Hence, the $\log \det \Sigma_l$ term is not allowed to push $\sigma_{l,i}^2$ to 0 without increasing the square root term that would tend to ∞ in that case. Consequently, the new term (the second summand) under the square root ensures a rapid growth of the functional value at the lower bound of the functional domain with respect to $\sigma_{l,i}^2$ when $|\mathcal{O}_l| > 0$, similar to coercivity, which is needed for the lower bound of the functional and the existence of minimizers since we do not have the closedness of the functional domain here. If there is variation along the i -th eigenvector, then $[V_l^T (g(x) - \mu_l)]_i \neq 0$ and hence, the square root term sees $\sigma_{l,i}^2$, preventing the $\log \det \Sigma_l$ term from pushing $\sigma_{l,i}^2$ to 0. Since ϕ is monotonically decreasing and converging to 0 when its argument tends to ∞ by construction, it contributes less the more variation we have along the considered direction. Ideally, ϕ decays rapidly such that J_ϕ differs significantly from J_0 only when the variation in the data along a certain direction is extremely small. This leads to the idea to use a discontinuous function ϕ such that $\phi(0) = 1$ and $\phi(t) = 0$ for $t \in \mathbb{R}_{>0}$. However, then we could choose $\mathcal{O}_l$ in such a way that it has positive measure but the spectra show arbitrary small positive variation in a certain direction. In such a scenario, ϕ would never contribute but the influence of the anisotropic distance could be made arbitrarily small, giving more and more impact to the $\log \det \Sigma_l$ term such that it can make the functional tend to $-\infty$. Hence, ϕ has to contribute when its argument comes very close to 0.

We can observe that at least one of the terms under the square root is positive. Both are bounded below by 0. Hence, the argument of the square root is now always positive, which makes the regularization for differentiability by η dispensable.

With the same argumentation as in Section 2.2.5, we can derive a lower bound for the entries of σ^2 in J_ϕ when $|\mathcal{O}_l| > 0$. We obtain for the l -th summand of the outer

sum in J_ϕ

$$\begin{aligned} & \int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \left[[V_l^T(g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T(g(x) - \mu_l)]_i^2 \right) \right] \frac{1}{\sigma_{l,i}^2}} + \beta \sum_{i=1}^L \log \sigma_{l,i}^2 \, dx \\ & \geq \sum_{i=1}^L \left[\frac{1}{\sqrt{L}} \int_{\mathcal{O}_l} \sqrt{[V_l^T(g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T(g(x) - \mu_l)]_i^2 \right)} \, dx \frac{1}{\sqrt{\sigma_{l,i}^2}} \right. \\ & \quad \left. + \beta |\mathcal{O}_l| \log \sigma_{l,i}^2 \right]. \end{aligned}$$

Again, we have for each $i \in \{1, \dots, L\}$ a function of the form (2.42), which has a global minimum at

$$(\sigma_{l,i}^2)^* := \left(\frac{\int_{\mathcal{O}_l} \sqrt{[V_l^T(g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T(g(x) - \mu_l)]_i^2 \right)} \, dx}{2\sqrt{L} |\mathcal{O}_l| \beta} \right)^2.$$

Note that the expression does not scale with $|\mathcal{O}_l|$. We can interpret the integrand in the numerator as a function

$$\begin{aligned} f &: [0, \infty) \rightarrow [0, \infty), \\ t &\mapsto \sqrt{t + \alpha \phi(t)}, \end{aligned}$$

evaluated at $[V_l^T(g(x) - \mu_l)]_i^2$. Because of the continuity of ϕ and $\phi(0) = 1$ as well as $\lim_{t \rightarrow \infty} \phi(t) = 0$, we obtain that f is continuous and

$$f(0) = \sqrt{\alpha}, \quad \lim_{t \rightarrow \infty} f(t) = \infty.$$

In particular, the non-negativity and continuity of ϕ as well as $\phi(0) = 1$ imply that $c_\phi := \inf_{t \in [0, \infty)} f(t) > 0$. Hence, we find for $(\sigma_{l,i}^2)^*$

$$(\sigma_{l,i}^2)^* = \left(\frac{\int_{\mathcal{O}_l} f \left([V_l^T(g(x) - \mu_l)]_i^2 \right) \, dx}{2\sqrt{L} |\mathcal{O}_l| \beta} \right)^2 \geq \left(\frac{\int_{\mathcal{O}_l} c_\phi \, dx}{2\sqrt{L} |\mathcal{O}_l| \beta} \right)^2 = \left(\frac{c_\phi}{2\sqrt{L} \beta} \right)^2 > 0, \quad (2.51)$$

and define that as

$$\underline{c}_\phi := \left(\frac{c_\phi}{2\sqrt{L} \beta} \right)^2. \quad (2.52)$$

Hence, also for ϕ AMS the eigenvalues are bounded away from 0 by a fixed lower bound, which depends on the choice of ϕ but not on the data. Although this does not seem to be a step forward compared to ϵ AMS, this functional works without the regularization with η for differentiability of the square root. Whether the function ϕ is easier to choose in advance than the parameter ϵ in ϵ AMS is unclear. However, the choice for ϕ that we use for the numerical experiments will turn out to work for all datasets we test on while different choices for ϵ are needed to obtain good results on the different datasets. Moreover, we also observe that the faster $\phi(t) \rightarrow 0$ when $t \rightarrow \infty$, the closer c_ϕ to 0. Hence, also the introduction of a data-dependent hyperparameter controlling the strength of the decay of ϕ would be possible.

We need to optimize J_ϕ with respect to parameters that appear inside ϕ . If the optimization is carried out gradient-based, ϕ also has to be differentiable which implies the required continuity. A possible choice for the function ϕ is

$$\phi(t) = \frac{1}{\exp t}, \quad (2.53)$$

which we will use for the numerical experiments in Chapter 4.

2.3.2 Existence of minimizers

The proof for the existence result works essentially as the proof in the case of ϵ AMS. A result that we will need is the compactness of $O(L)$:

Lemma 2.30 (Compactness of $O(L)$).

The set of orthogonal matrices $O(L) \subseteq \mathbb{R}^{L \times L}$ is compact.

Proof. We define the function

$$\begin{aligned} f: \mathbb{R}^{L \times L} &\rightarrow \mathbb{R}^{L \times L}, \\ M &\mapsto M^T M, \end{aligned}$$

which is continuous. Then, $O(L) = f^{-1}(\{I\})$, and by the closedness of $\{I\}$ and the continuity of f , we obtain the closedness of $O(L)$. The boundedness follows because the operator norm is

$$\|M\|_2 = \sup_{\substack{z \in \mathbb{R}^L, \\ \|z\|_2 \leq 1}} \|Mz\|_2 = \sup_{\substack{z \in \mathbb{R}^L, \\ \|z\|_2 \leq 1}} \sqrt{z^T M^T M z} = \sup_{\substack{z \in \mathbb{R}^L, \\ \|z\|_2 \leq 1}} \sqrt{z^T z} = \sup_{\substack{z \in \mathbb{R}^L, \\ \|z\|_2 \leq 1}} \|z\|_2 = 1$$

for every $M \in O(L)$. Since $\mathbb{R}^{L \times L}$ is finite-dimensional, the Heine-Borel theorem yields the compactness of $O(L)$. $\square$

Theorem 2.31 (Existence of minimizers of ϕ AMS).

For the functional J_ϕ , defined in (2.46) and (2.47), there exists a minimizer in X_ϕ .

Proof. We will proceed analogously to the proof of Theorem 2.21 and aim to apply the direct method in the calculus of variations as sketched in Section 2.2.2. We will essentially only point out and describe in more detail where the two proofs differ. Please keep in mind that subsequences of the minimizing sequence are chosen consecutively for each component of the minimizing sequence and that each subsequence is denoted as its parent sequence to avoid too many subindices.

The functional J_ϕ is bounded from below because, as already argued in Section 2.3.1, the term under the square root in (2.47) is always positive, the perimeter is non-negative and $\sigma_{l,i}^2$ is bounded below by (2.51). We define

$$\underline{J}_\phi := \inf_{(\mathcal{O}, \mu, V, \sigma^2) \in X_\phi} J_\phi [\mathcal{O}, \mu, V, \sigma^2] > -\infty.$$

It follows the existence of a minimizing sequence

$$(\mathcal{O}_n, \mu_n, V_n, \sigma_n^2)_{n \in \mathbb{N}} \subseteq X_\phi$$

such that

$$J_\phi [\mathcal{O}_n, \mu_n, V_n, \sigma_n^2] \rightarrow \underline{J}_\phi \text{ when } n \rightarrow \infty.$$

As before, the sequence of evaluations $(J_\phi [\mathcal{O}_n, \mu_n, V_n, \sigma_n^2])_{n \in \mathbb{N}}$ is bounded since it is convergent in $\mathbb{R}$. Hence, we have

$$J_\phi [\mathcal{O}_n, \mu_n, V_n, \sigma_n^2] \leq \bar{C}$$

for any $n \in \mathbb{N}$ and a sufficiently large $\bar{C} < \infty$.

We find the convergent subsequences and the distinguished element fulfilling the inequality in (2.21) as follows: the convergent subsequence of $(\mathcal{O}_n)_{n \in \mathbb{N}}$ with limit $\mathcal{O}^*$ is obtained exactly as in the proof of Theorem 2.21. Furthermore, by Lemma 2.30 we can directly choose a subsequence of $(V_n)_{n \in \mathbb{N}}$ converging to V^* . For the following considerations, we rewrite the functional J_ϕ (cf. (2.47)) by applying the linearity of the integral and computing the resulting integrals of the two integrands separately:

$$\begin{aligned} J_\phi [\mathcal{O}, \mu, V, \sigma^2] &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \sqrt{\sum_{i=1}^L \left[[V_l^T (g(x) - \mu_l)]_i^2 + \alpha \phi \left([V_l^T (g(x) - \mu_l)]_i^2 \right) \right]} \frac{1}{\sigma_{l,i}^2} dx \right. \\ &\quad \left. + \sum_{i=1}^L |\mathcal{O}_l| \beta \log \sigma_{l,i}^2 + \lambda \text{Per} (\mathcal{O}_l, \Omega) \right]. \end{aligned}$$

For $(\mu_n)_{n \in \mathbb{N}}$ and $(\sigma_n^2)_{n \in \mathbb{N}}$, we distinguish for any $l \in \{1, \dots, k\}$ again the two cases (1) $|\mathcal{O}_l^*| > 0$ and (2) $|\mathcal{O}_l^*| = 0$. In case (1) we find for each $n \in \mathbb{N}$ and each $i \in \{1, \dots, L\}$ using (2.30) that if $\log \sigma_{n,l,i}^2 < 0$, we obtain

$$\begin{aligned} \bar{C} &\geq |\mathcal{O}_{n,l}| \beta \log \sigma_{n,l,i}^2 \geq (c_l + \delta) \beta \log \sigma_{n,l,i}^2 \\ &\Rightarrow \infty > \exp \left(\frac{\bar{C}}{(c_l + \delta) \beta} \right) \geq \sigma_{n,l,i}^2, \end{aligned}$$

and if $\log \sigma_{n,l,i}^2 \geq 0$, it follows

$$\begin{aligned} \bar{C} &\geq |\mathcal{O}_{n,l}| \beta \log \sigma_{n,l,i}^2 \geq (c_l - \delta) \beta \log \sigma_{n,l,i}^2 \\ &\Rightarrow \infty > \exp \left(\frac{\bar{C}}{(c_l - \delta) \beta} \right) \geq \sigma_{n,l,i}^2. \end{aligned}$$

Consequently, $\sigma_{n,l,i}^2$ is bounded above for any $n \in \mathbb{N}$ and also bounded below with a positive lower bound c_ϕ (cf. (2.51) and (2.52)). Therefore, $\sigma_{n,l,i}^2$ is included in a bounded and closed interval with positive lower bound, which is a compact set. Hence, a convergent subsequence can be chosen that converges to a $(\sigma_{l,i}^2)^* \in \mathbb{R}_{>0}$. The boundedness of $(\mu_{n,l})_{n \in \mathbb{N}}$ follows in case (1) from

$$\begin{aligned} \bar{C} &\geq \int_{\mathcal{O}_{n,l}} \sqrt{\sum_{i=1}^L \left[\left[V_{n,l}^T (g(x) - \mu_{n,l}) \right]_i^2 + \underbrace{\alpha \phi \left(\left[V_{n,l}^T (g(x) - \mu_{n,l}) \right]_i^2 \right)}_{\geq 0}} \right] \frac{1}{\sigma_{n,l,i}^2} \, dx} \\ &\geq \int_{\mathcal{O}_{n,l}} \sqrt{\sum_{i=1}^L \left[V_{n,l}^T (g(x) - \mu_{n,l}) \right]_i^2 \frac{1}{\sigma_{n,l,i}^2}} \, dx \\ &\geq \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \int_{\mathcal{O}_{n,l}} \sqrt{(g(x) - \mu_{n,l})^T V_{n,l} V_{n,l}^T (g(x) - \mu_{n,l})} \, dx \\ &= \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \int_{\mathcal{O}_{n,l}} \|\mu_{n,l} - g(x)\|_2 \, dx \\ &\geq \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \int_{\mathcal{O}_{n,l}} \left| \|\mu_{n,l}\|_2 - \|g(x)\|_2 \right| \, dx \\ &\geq \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \int_{\mathcal{O}_{n,l}} \|\mu_{n,l}\|_2 - \|g(x)\|_2 \, dx \\ &\geq \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \left(|\mathcal{O}_{n,l}| \|\mu_{n,l}\|_2 - \int_{\mathcal{O}_{n,l}} \|g(x)\|_2 \, dx \right) \end{aligned}$$

$$\geq \sqrt{\frac{1}{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2}} \left((c_l - \delta) \|\mu_{n,l}\|_2 - |\Omega| \|g(\cdot)\|_2 \|L^\infty(\Omega)\| \right),$$

which implies

$$\infty > \frac{\sqrt{\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2} \overline{C} + |\Omega| \|g(\cdot)\|_2 \|L^\infty(\Omega)\|}{c_l - \delta} \geq \|\mu_{n,l}\|_2$$

since $\max_{j \in \{1, \dots, L\}} \sigma_{n,l,j}^2$ is bounded, as shown before. Hence, by the boundedness, we can choose a convergent subsequence of $(\mu_{n,l})_{n \in \mathbb{N}}$.

The inequality in (2.21) for the l -th summand of the outer sum is then established analogously to the proof of Theorem 2.21 using the continuity of ϕ .

In case (2), as in the proof for Theorem 2.21, there is no control about $(\sigma_{n,l,i}^2)_{n \in \mathbb{N}}$ in general when $n \rightarrow \infty$ for $l \in \{1, \dots, k\}$ and $i \in \{1, \dots, L\}$. However, if $(\log \sigma_{n,l,i}^2)_{n \in \mathbb{N}}$ is bounded above, it follows the boundedness of $(\sigma_{n,l,i}^2)_{n \in \mathbb{N}}$ and we can again choose a convergent subsequence with limit $(\sigma_{l,i}^2)^* \in \mathbb{R}_{>0}$. If it is not bounded, we can choose a diverging, monotonically increasing subsequence of $(\sigma_{n,l,i}^2)_{n \in \mathbb{N}}$ such that $\log \sigma_{n,l,i}^2 \geq 0$ for any $n \in \mathbb{N}$ and an element $(\sigma_{l,i}^2)^* := 1 \in \mathbb{R}_{>0}$. We obtain

$$\underbrace{|\mathcal{O}_l^*|}_{=0} \beta \log(\sigma_{l,i}^2)^* = 0 \leq \liminf_{n \rightarrow \infty} \underbrace{|\mathcal{O}_{n,l}|}_{\geq 0} \underbrace{\beta}_{>0} \underbrace{\log \sigma_{n,l,i}^2}_{\geq 0}.$$

The remaining parts of the proof, i.e., the selection of a subsequence of $(\mu_{n,l})_{n \in \mathbb{N}}$ and an element μ_l^* as well as showing the inequality in (2.21), work analogously to the steps in the proof of Theorem 2.21, using the positivity of the square root term and the continuity of ϕ . Hence, $(\mathcal{O}^*, \mu^*, V^*, (\sigma^2)^*) \in X_\phi$ minimizes J_ϕ . $\square$

2.4 The kernelized Mumford–Shah model

In Section 2.1, we have introduced the piecewise constant MS model as the basis of ϵ AMS and ϕ AMS and explained our modifications to generalize the classical indicator function (cf. (2.6)) to make the model more flexible and powerful. However, the indicator functions of ϵ AMS and ϕ AMS process the spectra in their original space or in the space after a potential preprocessing and hence, have to separate the spectral clusters as they are. An alternative way to enhance the model is to use a (potentially non-linear) mapping as part of the model that maps the data into a space where the spectral clusters are ideally well separated, making it easier to detect these clusters by the model. Such a mapping can also additionally get spatial information to combine it with the spectral information and enable a disentanglement of overlapping clusters in

the original spectral space. We will choose mappings that lift the spectra into an inner product space. Hence, an inner product is available and we maintain the possibility to measure angles and lengths. The reference for the theoretical and background information given in this section is [99].

2.4.1 Model description

We use essentially the same indicator function as the traditional MS model in (2.6), namely, the Euclidean distance between spectrum and mean, but compute this distance in the embedding inner product space and with respect to the norm induced by the inner product in that space. That is, let

$$\Phi: \mathbb{R}^L \rightarrow \mathcal{H}$$

be the mapping that embeds the spectra that live in $\mathbb{R}^L$ into an inner product space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The map Φ is called the *feature map* and the space $\mathcal{H}$ the *feature space*. Then, we define the indicator function for the l -th segment of an image $g \in L^\infty(\Omega, \mathbb{R}^L)$ with $l \in \{1, \dots, k\}$ at position $x \in \Omega$ as

$$f_{\Phi,l}(x; \mu_l^\Phi) := \|\Phi(g(x)) - \mu_l^\Phi\|_{\mathcal{H}}^2, \quad (2.54)$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm on $\mathcal{H}$ induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\mu_l^\Phi \in \mathcal{H}$ is the mean feature of the embedded spectra in segment $\mathcal{O}_l$ with respect to $\|\cdot\|_{\mathcal{H}}$. The resulting functional is (cf. (2.9))

$$J_\Phi: \mathcal{A} \times \mathcal{H}^k \rightarrow \mathbb{R},$$

$$(\mathcal{O}_1, \dots, \mathcal{O}_k, \mu_1^\Phi, \dots, \mu_k^\Phi) \mapsto \sum_{l=1}^k \left[\int_{\mathcal{O}_l} \|\Phi(g(x)) - \mu_l^\Phi\|_{\mathcal{H}}^2 dx + \lambda \text{Per}(\mathcal{O}_l, \Omega) \right]. \quad (2.55)$$

We can either compute J_Φ by choosing an explicit Φ or we can shift the choice of the feature map and its evaluation to the choice and evaluation of a so-called *kernel* since there is a connection between feature maps and kernels. We will use this connection for our purpose and introduce now the notion of a kernel.

Let X be a non-empty set.

Definition 2.32 (Positive semidefinite kernel [99, Definitions 2.3, 2.4 and 2.5]).

A function $K: X \times X \rightarrow \mathbb{R}$ is called a symmetric, positive semidefinite kernel if for any $x, x' \in X$ we have

$$K(x, x') = K(x', x),$$

and if for each $m \in \mathbb{N}$ and any $x_1, \dots, x_m \in X$, the associated Gram matrix (or kernel matrix) $G \in \mathbb{R}^{m \times m}$ with respect to these elements, given as the $m \times m$ matrix with entries

$$G_{i,j} := K(x_i, x_j)$$

for $i, j \in \{1, \dots, m\}$, is positive semidefinite, i.e.,

$$\sum_{i,j=1}^m z_i z_j G_{i,j} = \sum_{i,j=1}^m z_i z_j K(x_i, x_j) \geq 0$$

for any $z_1, \dots, z_m \in \mathbb{R}$.

The terminology of the kind of kernels described in the definition is not consistent in the literature. Kernels with these properties are also sometimes called *positive definite*, which can be confusing when compared to the definition of a positive definite matrix. In this thesis, we aim to be consistent with the usual terminology in matrix theory to avoid confusion with the positive definiteness of the covariance matrix that we required and discussed in the previous sections.

A kernel can be understood as measuring the similarity of two elements in X . An inner product also measures the similarity of two vectors. Therefore, given an inner product space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as feature space and a feature map $\Phi: X \rightarrow \mathcal{H}$, one can obtain a kernel on X by defining

$$K(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}. \quad (2.56)$$

This kernel is symmetric by the symmetry of the inner product. It is furthermore positive semidefinite because for any $m \in \mathbb{N}$, $x_1, \dots, x_m \in X$ and $z_1, \dots, z_m \in \mathbb{R}$, we find with the linearity of the inner product

$$\sum_{i,j=1}^m z_i z_j K(x_i, x_j) = \left\langle \sum_{i=1}^m z_i \Phi(x_i), \sum_{j=1}^m z_j \Phi(x_j) \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^m z_i \Phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.$$

The connection between feature map and kernel also works the other way around:

Proposition 2.33 ([87, Th. 1.2]).

If X is a non-empty set and $K: X \times X \rightarrow \mathbb{R}$ a symmetric, positive semidefinite kernel, it follows the existence of a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a map $\Phi: X \rightarrow \mathcal{H}$ such that $\mathcal{H} = \text{span}\{\Phi(x) \mid x \in X\}$ and

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} \quad (2.57)$$

for any $x, x' \in X$.

Hence, we can choose a kernel such that its evaluations correspond to inner product evaluations in some feature space.

Remark 2.34 (Reproducing kernel Hilbert space).

The proposition only states the existence of a feature map and a feature space and does not explain how to get explicit representations of them. For the purpose of this thesis, this is enough as we want to stick to evaluations of the kernel that allow us to do computations in the feature space without knowing the feature map. However, it is possible to explicitly construct the feature map and feature space from a given kernel. One such construction leads to the so-called reproducing kernel Hilbert space. The reproducing kernel Hilbert space with respect to a kernel uniquely determines this kernel in the sense that if two kernels result in the same reproducing kernel Hilbert space, one can show that these kernels function equally when evaluated at any pair of elements from X . We refer to [99] for an explanation of the construction of the reproducing kernel Hilbert space.

Evaluating a kernel instead of mapping the data points into the feature space and computing the inner product there has the advantage that one avoids computations in a possibly very high-dimensional or even infinite-dimensional space, which can be computationally expensive compared to a kernel evaluation in the original data space. Another advantage is the so-called *kernel trick*:

Remark 2.35 (Kernel trick [99, Rem. 2.8]).

We are given two different kernels K and $\tilde{K}$ and an algorithm that is defined in terms of kernel evaluations of K . By replacing K with the alternative kernel $\tilde{K}$ in the algorithm, we can apply the same algorithm in a different inner product space without any additional effort. This works because by Proposition 2.33 the two kernels have associated feature spaces $\mathcal{H}$ and $\tilde{\mathcal{H}}$ and associated feature maps Φ and $\tilde{\Phi}$, respectively, such that (2.57) holds. In particular, this means that the algorithm based on K computes inner products of elements $\Phi(x)$ in $\mathcal{H}$, while the replacement of K with $\tilde{K}$ means that the same algorithm now computes inner products of elements $\tilde{\Phi}(x)$ in $\tilde{\mathcal{H}}$.

Hence, by simply replacing one kernel with another kernel, we can run an algorithm in a different Hilbert space without even knowing how to map the data into that Hilbert space. This makes kernels a powerful and flexible tool to enhance algorithms. It was already commented above that also an inner product is a similarity measure. If X is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_X$, we can choose Φ as the identity, i.e., $\mathcal{H} = X$, and obtain with the definition in (2.56) a kernel that equals the inner product. Applying the kernel trick in this setting then means that any algorithm working in X and being based on evaluations of $\langle \cdot, \cdot \rangle_X$ can be easily carried out in another Hilbert space by making use of the kernel trick and replacing the inner product evaluations with evaluations of another kernel. In general, we can apply the

kernel trick to any algorithm that can be expressed in terms of inner products and introduce non-linearity to it.

The kernel trick can also be applied in our setting, namely to J_Φ . To this end, we switch to the discrete setting, i.e., g now maps from $\mathcal{N}$ to $W \subseteq \mathbb{R}^L$ and we denote one element from $\mathcal{N}$ with $\mathbf{i} = (i_1, \dots, i_d)$. It is enough to focus only on the indicator function since only there did we introduce the feature map and therefore, it is the only place where the kernel trick shows its effect. In our case, an indicator function sees only the spectra, i.e., $g(\mathbf{i})$ for $\mathbf{i} \in \mathcal{N}$, and hence $X := \{g(\mathbf{i}) \mid \mathbf{i} \in \mathcal{N}\} \subseteq \mathbb{R}^L$ in our case. Let $K: X \times X \rightarrow \mathbb{R}$ be a symmetric, positive semidefinite kernel and $\Phi: X \rightarrow \mathcal{H}$ be a feature map from X to a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that (2.57) holds (guaranteed to exist by Proposition 2.33). For every $l \in \{1, \dots, k\}$, we denote the cardinality of the (now discrete and therefore finite set) $\mathcal{O}_l$ with $n_l \in \mathbb{N}_0$ and choose $\mu_l^\Phi \in \mathcal{H}$ as the arithmetic mean of the embedded spectra of segment $\mathcal{O}_l \subseteq \mathcal{N}$, i.e.,

$$\mu_l^\Phi := \frac{1}{n_l} \sum_{\mathbf{i} \in \mathcal{O}_l} \Phi(g(\mathbf{i})). \quad (2.58)$$

This choice will be justified in Section 3.3.3. Then, we can reformulate the discrete indicator function of J_Φ evaluated at $\mathbf{i} \in \mathcal{N}$ using the linearity of the inner product as

$$\begin{aligned} f_{\Phi,l}(\mathbf{i}; \mu_l^\Phi) &= \|\Phi(g(\mathbf{i})) - \mu_l^\Phi\|_{\mathcal{H}}^2 \\ &= \langle \Phi(g(\mathbf{i})) - \mu_l^\Phi, \Phi(g(\mathbf{i})) - \mu_l^\Phi \rangle_{\mathcal{H}} \\ &= \langle \Phi(g(\mathbf{i})), \Phi(g(\mathbf{i})) \rangle_{\mathcal{H}} - 2 \langle \Phi(g(\mathbf{i})), \mu_l^\Phi \rangle_{\mathcal{H}} + \langle \mu_l^\Phi, \mu_l^\Phi \rangle_{\mathcal{H}} \\ &= \langle \Phi(g(\mathbf{i})), \Phi(g(\mathbf{i})) \rangle_{\mathcal{H}} - 2 \left\langle \Phi(g(\mathbf{i})), \frac{1}{n_l} \sum_{\mathbf{j} \in \mathcal{O}_l} \Phi(g(\mathbf{j})) \right\rangle_{\mathcal{H}} \\ &\quad + \left\langle \frac{1}{n_l} \sum_{\mathbf{j} \in \mathcal{O}_l} \Phi(g(\mathbf{j})), \frac{1}{n_l} \sum_{\mathbf{m} \in \mathcal{O}_l} \Phi(g(\mathbf{m})) \right\rangle_{\mathcal{H}} \\ &= \langle \Phi(g(\mathbf{i})), \Phi(g(\mathbf{i})) \rangle_{\mathcal{H}} - \frac{2}{n_l} \sum_{\mathbf{j} \in \mathcal{O}_l} \langle \Phi(g(\mathbf{i})), \Phi(g(\mathbf{j})) \rangle_{\mathcal{H}} \\ &\quad + \frac{1}{n_l^2} \sum_{\mathbf{j} \in \mathcal{O}_l} \sum_{\mathbf{m} \in \mathcal{O}_l} \langle \Phi(g(\mathbf{j})), \Phi(g(\mathbf{m})) \rangle_{\mathcal{H}}. \end{aligned}$$

The computation of the indicator function is now based on inner products of the spectra in the feature space. With (2.57), we can write $f_{\Phi,l}$ in terms of kernel

evaluations. That yields

$$f_{\Phi,l}(\mathbf{i}; \mu_l^\Phi) = K(g(\mathbf{i}), g(\mathbf{i})) - \frac{2}{n_l} \sum_{j \in \mathcal{O}_l} K(g(\mathbf{i}), g(\mathbf{j})) + \frac{1}{n_l^2} \sum_{j \in \mathcal{O}_l} \sum_{\mathbf{m} \in \mathcal{O}_l} K(g(\mathbf{j}), g(\mathbf{m})). \quad (2.59)$$

In particular, we can now apply the kernel trick and replace K with any other symmetric and positive semidefinite kernel $\tilde{K}$ to carry out the computations in another feature space. This motivates the name *kernelized Mumford–Shah model* (kMS). Moreover, since $\mathcal{N}$ is finite, we have only finitely many spectra $g(\mathbf{i})$ for $\mathbf{i} \in \mathcal{N}$. Since the computation of $f_{\Phi,l}(\mathbf{i}; \mu_l^\Phi)$ is only based on kernel evaluations of pairs of spectra, which are exactly the entries of the Gram matrix, and neither the kernel nor the spectra change during minimization, we can precompute the Gram matrix to speed up computations. The computation of $f_{\Phi,l}(\mathbf{i}; \mu_l^\Phi)$ then boils down to extracting the corresponding entries from the Gram matrix and inserting them into the formula. A drawback of precomputing the kernel matrix in practice is that it has the size *number of pixels* $\times$ *number of pixels*. Hence, even for very small images of size, e.g., 145×145 pixels, which is the size of a benchmark dataset recorded with AVIRIS, it has a size of 21025×21025 entries. If each entry is stored in single precision, such a matrix already needs about 1.77 gigabytes of memory. Unfortunately, there are kernels like, e.g., the so-called Gaussian radial basis function (RBF) kernel (cf. (2.62)), which are of great interest in practice but for which one can show that the Gram matrix has full rank in general. Nevertheless, there are techniques to tackle that problem and approximate the kernel matrix with low-rank approximations tailored for kernel matrices, see for example [99].

2.4.2 Kernels for hyperspectral imaging

It remains to answer the question of how to choose suitable kernels for hyperspectral imaging. We will explain three different kernels that are designed for hyperspectral image classification and are proposed in [19]. These kernels will also be used in Chapter 4 when testing kMS.

The idea is to use spectral and spatial information as well as their combination and introduce this information in the kernel evaluation. To this end, we use the following facts about symmetric, positive semidefinite kernels that allow designing new kernels from given ones.

Proposition 2.36 ([30, Prop. 3.12]).

Let $X \subseteq \mathbb{R}^L$, $\alpha > 0$ and let $M \in \mathbb{R}^{L \times L}$ be a symmetric and positive semidefinite matrix. Furthermore, let K_1 and K_2 be two symmetric and positive semidefinite kernels such that

$$K_i: X \times X \rightarrow \mathbb{R},$$

for $i \in \{1, 2\}$. Then, the functions

1. $K(x, x') := K_1(x, x') + K_2(x, x')$,
2. $K(x, x') := \alpha K_1(x, x')$,
3. $K(x, x') := x^T M x'$,

are also symmetric, positive semidefinite kernels.

A second fact we need is that given two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$, respectively, the direct sum of $\mathcal{H}_1$ and $\mathcal{H}_2$ is again a Hilbert space $\mathcal{H} := \mathcal{H}_1 \oplus \mathcal{H}_2$ with inner product

$$\langle (x_1, x_2), (x'_1, x'_2) \rangle_{\mathcal{H}} := \langle x_1, x'_1 \rangle_{\mathcal{H}_1} + \langle x_2, x'_2 \rangle_{\mathcal{H}_2} \quad (2.60)$$

for two elements $(x_1, x_2), (x'_1, x'_2) \in \mathcal{H}$ [29]. We can use this for our purposes in the following way: For each pixel at $\mathbf{i} \in \mathcal{N}$, we are given its spectral information, encoded with a subscript ω , as the recorded spectrum $g_\omega^{\mathbf{i}} := g(\mathbf{i})$ and consider the *contextual information*, encoded with a subscript s , denoted with $g_s^{\mathbf{i}} \in \mathbb{R}^{N_s}$ for some $N_s \in \mathbb{N}$, as some extracted features from a certain area around the pixel at position $\mathbf{i}$. This could be, for instance, the mean spectrum in this area or the channel-wise standard deviation. Let $X_\omega := \{g_\omega^{\mathbf{i}} \mid \mathbf{i} \in \mathcal{N}\} \subseteq \mathbb{R}^L$ the set containing the spectral information and $X_s := \{g_s^{\mathbf{i}} \mid \mathbf{i} \in \mathcal{N}\} \subseteq \mathbb{R}^{N_s}$ the set containing the contextual information. We choose symmetric and positive semidefinite kernels $K_\omega: X_\omega \times X_\omega \rightarrow \mathbb{R}$ and $K_s: X_s \times X_s \rightarrow \mathbb{R}$ for the spectral and contextual information, respectively. As explained in Section 2.4.1, this gives us a feature map

$$\Phi_\omega: X_\omega \rightarrow \mathcal{H}_\omega$$

into a Hilbert space $\mathcal{H}_\omega$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\omega}$ for the spectral features and a feature map

$$\Phi_s: X_s \rightarrow \mathcal{H}_s$$

into a Hilbert space $\mathcal{H}_s$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_s}$ for the contextual features. Camps-Valls et al. [19] build on this and propose as composite kernels the *direct summation kernel (DSK)*, the *weighted summation kernel (WSK)* and the *cross-information kernel (CIK)*.

Direct summation kernel

The DSK combines spectral with contextual information by directly adding the kernel values of K_ω and K_s . To derive an associated feature map from Φ_ω and Φ_s , we define a feature map

$$\Phi: X \rightarrow \mathcal{H}$$

with $X := X_\omega \times X_s$ and $\mathcal{H} := \mathcal{H}_\omega \oplus \mathcal{H}_s$ given as

$$\Phi \left((g_\omega^i, g_s^i) \right) := \left(\Phi_\omega(g_\omega^i), \Phi_s(g_s^i) \right).$$

Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ describe the inner product on $\mathcal{H}$. Using the linearity of inner products, this leads to

$$\begin{aligned} K \left((g_\omega^i, g_s^i), (g_\omega^j, g_s^j) \right) &:= K_\omega(g_\omega^i, g_\omega^j) + K_s(g_s^i, g_s^j) \\ &= \left\langle \Phi_\omega(g_\omega^i), \Phi_\omega(g_\omega^j) \right\rangle_{\mathcal{H}_\omega} + \left\langle \Phi_s(g_s^i), \Phi_s(g_s^j) \right\rangle_{\mathcal{H}_s} \\ (2.60) \quad &= \left\langle \left(\Phi_\omega(g_\omega^i), \Phi_s(g_s^i) \right), \left(\Phi_\omega(g_\omega^j), \Phi_s(g_s^j) \right) \right\rangle_{\mathcal{H}} \\ &= \left\langle \Phi \left((g_\omega^i, g_s^i) \right), \Phi \left((g_\omega^j, g_s^j) \right) \right\rangle_{\mathcal{H}}. \end{aligned} \tag{2.61}$$

We see that this feature map Φ and feature space $\mathcal{H}$ correspond to the DSK that acts on $X \times X$. Depending on the specific choices for K_ω and K_s , the feature maps Φ_ω and Φ_s as well as the Hilbert spaces $\mathcal{H}_\omega$ and $\mathcal{H}_s$ change, resulting in a different Φ and $\mathcal{H}$.

Weighted summation kernel

The WSK is a generalization of the DSK. The idea is to use a convex combination of K_ω and K_s . Let $\alpha \in (0, 1)$. Then, we define the WSK as

$$K \left((g_\omega^i, g_s^i), (g_\omega^j, g_s^j) \right) := \alpha K_\omega(g_\omega^i, g_\omega^j) + (1 - \alpha) K_s(g_s^i, g_s^j),$$

which is a symmetric and positive semidefinite kernel by Proposition 2.36. Hence, we obtain the WSK by replacing the kernels in the DSK by their weighted counterparts. An advantage of this approach is that we can insert the prior knowledge into the model if it should rely more on the spectral or contextual information. In this thesis, we choose the same α for every segment, but one could also consider choosing α for each segment separately.

Cross-information kernel

As the name of the kernel suggests, the idea is to compare also spectral with contextual information in a kernel. To this end, both need the same dimension, i.e., we need to construct the contextual information such that $N_s = L$ holds. In the two examples of possible constructions of the contextual information given above, namely, as the mean spectrum or as the channel-wise standard deviation in a local area around the pixel, this condition is fulfilled. In addition to the two chosen kernels for the spectral

and contextual information $K_\omega: X_\omega \times X_\omega \rightarrow \mathbb{R}$ and $K_s: X_s \times X_s \rightarrow \mathbb{R}$ from above and the corresponding feature maps Φ_ω and Φ_s as well as feature spaces $\mathcal{H}_\omega$ and $\mathcal{H}_s$, we define $X_{\omega s} := X_\omega \cup X_s$ and choose a symmetric, positive semidefinite kernel $K_{\omega s}: X_{\omega s} \times X_{\omega s} \rightarrow \mathbb{R}$ with an associated feature map

$$\Phi_{\omega s}: X_{\omega s} \rightarrow \mathcal{H}_{\omega s},$$

where the Hilbert space $\mathcal{H}_{\omega s}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\omega s}}$ is the feature space. Then, we can set $X := X_\omega \times X_s$ and $\mathcal{H} := \mathcal{H}_\omega \oplus \mathcal{H}_s \oplus \mathcal{H}_{\omega s}$ which is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and define a composite feature map via

$$\begin{aligned} \Phi: X &\rightarrow \mathcal{H}, \\ (g_\omega^i, g_s^i) &\mapsto \left(\Phi_\omega(g_\omega^i), \Phi_s(g_s^i), \Phi_{\omega s}(g_\omega^i) + \Phi_{\omega s}(g_s^i) \right). \end{aligned}$$

We define the CIK as follows and find with a similar computation as in (2.61):

$$\begin{aligned} &K \left((g_\omega^i, g_s^i), (g_\omega^j, g_s^j) \right) \\ &:= K_\omega(g_\omega^i, g_\omega^j) + K_s(g_s^i, g_s^j) + K_{\omega s}(g_\omega^i, g_\omega^j) + K_{\omega s}(g_\omega^i, g_s^j) + K_{\omega s}(g_s^i, g_\omega^j) + K_{\omega s}(g_s^i, g_s^j) \\ &= \left\langle \Phi_\omega(g_\omega^i), \Phi_\omega(g_\omega^j) \right\rangle_{\mathcal{H}_\omega} + \left\langle \Phi_s(g_s^i), \Phi_s(g_s^j) \right\rangle_{\mathcal{H}_s} + \left\langle \Phi_{\omega s}(g_\omega^i), \Phi_{\omega s}(g_\omega^j) \right\rangle_{\mathcal{H}_{\omega s}} \\ &\quad + \left\langle \Phi_{\omega s}(g_\omega^i), \Phi_{\omega s}(g_s^j) \right\rangle_{\mathcal{H}_{\omega s}} + \left\langle \Phi_{\omega s}(g_s^i), \Phi_{\omega s}(g_\omega^j) \right\rangle_{\mathcal{H}_{\omega s}} + \left\langle \Phi_{\omega s}(g_s^i), \Phi_{\omega s}(g_s^j) \right\rangle_{\mathcal{H}_{\omega s}} \\ &= \left\langle \Phi_\omega(g_\omega^i), \Phi_\omega(g_\omega^j) \right\rangle_{\mathcal{H}_\omega} + \left\langle \Phi_s(g_s^i), \Phi_s(g_s^j) \right\rangle_{\mathcal{H}_s} \\ &\quad + \left\langle \Phi_{\omega s}(g_\omega^i) + \Phi_{\omega s}(g_s^i), \Phi_{\omega s}(g_\omega^j) + \Phi_{\omega s}(g_s^j) \right\rangle_{\mathcal{H}_{\omega s}} \\ &= \left\langle \left(\Phi_\omega(g_\omega^i), \Phi_s(g_s^i), \Phi_{\omega s}(g_\omega^i) + \Phi_{\omega s}(g_s^i) \right), \left(\Phi_\omega(g_\omega^j), \Phi_s(g_s^j), \Phi_{\omega s}(g_\omega^j) + \Phi_{\omega s}(g_s^j) \right) \right\rangle_{\mathcal{H}} \\ &= \left\langle \Phi \left((g_\omega^i, g_s^i) \right), \Phi \left((g_\omega^j, g_s^j) \right) \right\rangle_{\mathcal{H}}. \end{aligned}$$

We see that the feature map Φ and feature space $\mathcal{H}$ from above correspond to the CIK that is defined on $X \times X$. When considering the term $K_{\omega s}(g_\omega^i, g_\omega^j)$, we observe that $K_{\omega s}$ can be restricted to $X_\omega \times X_\omega$ in that case without changing the expression above. Hence, by Proposition 2.36, we can combine $K_\omega(g_\omega^i, g_\omega^j)$ and the restricted version of $K_{\omega s}(g_\omega^i, g_\omega^j)$ to obtain a symmetric and positive semidefinite kernel on $X_\omega \times X_\omega$ in practice that we again denote by K_ω . The same can be done with $K_s(g_s^i, g_s^j)$ and $K_{\omega s}(g_s^i, g_s^j)$ and we denote the resulting kernel by K_s . This yields the following formulation of K which is the final CIK

$$K \left((g_\omega^i, g_s^i), (g_\omega^j, g_s^j) \right) = K_\omega(g_\omega^i, g_\omega^j) + K_s(g_s^i, g_s^j) + K_{\omega s}(g_\omega^i, g_s^j) + K_{\omega s}(g_s^i, g_\omega^j).$$

Choices for spectral and contextual kernels

The spectral and contextual kernels have been abstract functions so far. Two popular concrete choices for kernels that Camps-Valls et al. [19] used for hyperspectral image classification in their experiments with their proposed composite kernels that we described above are the *polynomial kernel* and the *Gaussian RBF kernel*. Let $X \subseteq \mathbb{R}^L$ and $x, x' \in X$. The Gaussian RBF kernel is defined as

$$K(x, x') := \exp \left(-\gamma \|x - x'\|_2^2 \right), \quad (2.62)$$

with a scaling factor $\gamma \in (0, \infty)$. The polynomial kernel is given by

$$K(x, x') := (\delta x^T x' + c)^p, \quad (2.63)$$

with a scaling factor $\delta \in (0, \infty)$, a shift $c \geq 0$ and a *kernel degree* $p \in \mathbb{N}$. In the numerical experiments in Chapter 4, we follow Camps-Valls et al. [19] and use the polynomial kernel as the spectral kernel K_ω and the cross-information kernel $K_{\omega s}$ and the Gaussian RBF kernel as the contextual kernel K_s .

Chapter 3

Optimization

In the previous chapter, we have defined three new models for hyperspectral image segmentation, namely, ϵ AMS, ϕ AMS and kMS, and shown some theoretical properties of them. This chapter is devoted to the question of how to find minimizers of the functionals in practice to obtain segmentations of images. We start with a general high-level explanation of how a functional of the form (2.9) with a generic indicator function can be optimized with respect to the segmentation and potential degrees of freedom of the indicator functions, explain the initialization of the degrees of freedom and propose a stopping criterion for our optimization procedure in Section 3.1. Then, we describe the subsequent steps of our segmentation framework in more detail. We start with how to optimize over the admissible segmentations in Section 3.2. It follows an explanation of the optimization over the parameters of the indicator functions that are model-specific in Section 3.3. We conclude the chapter with comments on the computational complexity of ϵ AMS in Section 3.4. The description of how to optimize ϵ AMS, the computational complexity of ϵ AMS and the stopping criterion were already published in [26].

We restrict our considerations in this chapter to the case of a two-dimensional image domain, i.e., $\Omega \subseteq \mathbb{R}^d$ and $\mathcal{N} \subseteq \mathbb{R}^d$ with $d = 2$ are subsets of the two-dimensional real space, because the images used for the numerical experiments have only two-dimensional image domains and a more general description after discretization makes the formulas more complicated without giving more insights.

3.1 Optimization framework

We will now describe how we find minimizers of the MS functional J_{MS} as given in (2.9) with generic indicator functions $f_l(x; \theta)$ that depend on degrees of freedom θ which reside in a suitable space. The symbol θ is meant here as a placeholder for the model-specific degrees of freedom later. This procedure is used for all models

proposed in this thesis; only the subsequent optimization with respect to θ differs across the models and is therefore handled separately in Section 3.3. We use an alternating optimization approach, i.e., we optimize with respect to the partition $(\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k) \in \mathcal{A}$ of Ω while θ is fixed and fix the partition afterward to optimize θ . We alternate between these two steps several times.

However, in order to find good minimizers, we do not optimize the functional J_{MS} directly but relax the optimization problem by making it convex with respect to the partition, at the cost of enlarging the search space for solutions, having to project the found solution onto the original admissible set and losing the guarantee for the existence of a minimizer if $k > 2$.

3.1.1 Convexification

As a relaxation, we use the convexification proposed by Zach et al. [110]. Starting with J_{MS} (cf. (2.9)) and applying Definition 2.6, we reformulate it as follows (now also under consideration of possible parameters of the indicator functions):

$$\begin{aligned} J_{\text{MS}}[\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k, \theta] &= \sum_{l=1}^k \left[\int_{\mathcal{O}_l} f_l(x; \theta) \, dx + \lambda \text{Per}(\mathcal{O}_l, \Omega) \right] \\ &= \sum_{l=1}^k \left[\int_{\Omega} \chi_{\mathcal{O}_l}(x) f_l(x; \theta) \, dx + \lambda |\chi_{\mathcal{O}_l}|_{\text{BV}(\Omega)} \right]. \end{aligned}$$

The characteristic functions $\chi_{\mathcal{O}_l}: \Omega \rightarrow \{0, 1\}$ for each $l \in \{1, \dots, k\}$ are then replaced by functions $u_l: \Omega \rightarrow [0, 1]$ that allow function values in the interval $[0, 1]$ instead of restricting the output to the values $\{0, 1\}$, but we still require $\sum_{l=1}^k u_l(x) = 1$ for any $x \in \Omega$. We describe the k -dimensional unit simplex as

$$\mathbb{K}_k := \left\{ (s_1, s_2, \dots, s_k) \in \mathbb{R}^k \left| \forall l \in \{1, \dots, k\} : s_l \geq 0, \sum_{l=1}^k s_l = 1 \right. \right\} \quad (3.1)$$

and compose the so-called *labeling function* of $u_1, \dots, u_k$ as

$$u := (u_1, u_2, \dots, u_k): \Omega \rightarrow \mathbb{K}_k.$$

Then, the relaxed formulation of J_{MS} is

$$J_{\text{Zach}}[u, \theta] := \sum_{l=1}^k \left[\int_{\Omega} u_l(x) f_l(x; \theta) \, dx + \lambda |u_l|_{\text{BV}(\Omega)} \right]. \quad (3.2)$$

Since the integral term is linear in u and therefore convex, it follows with Lemma 2.2 the convexity of J_{Zach} with respect to u because sums of convex functions are still

convex. From a geometrical point of view, the convexification has the following effect: In the original formulation of J_{MS} (cf. (2.9)), a partition of Ω given as $(\chi_{\mathcal{O}_1}, \dots, \chi_{\mathcal{O}_k})$ evaluated at $x \in \Omega$ always yields a point in $\{0, 1\}^k \cap \mathbb{K}_k$, i.e., on one of the vertices of the unit simplex $\mathbb{K}_k$, by the partition property of $(\mathcal{O}_1, \dots, \mathcal{O}_k)$, while in the case of the relaxed version a labeling function u evaluated at $x \in \Omega$ yields a point in $\mathbb{K}_k$, i.e., also points inside the unit simplex are allowed. Hence, given a labeling function u , we obtain a unique segment label $c(x)$ for position $x \in \Omega$ by performing a thresholding

$$c(x) := \min \left\{ \arg \max_{l \in \{1, \dots, k\}} \{u_l(x)\} \right\} \in \{1, \dots, k\}. \quad (3.3)$$

The outer minimum is applied to handle the case where the $\arg \max$ operator might not return a singleton. This label can then be equivalently described as a one-hot encoding

$$(\delta_{c(x)1}, \delta_{c(x)2}, \dots, \delta_{c(x)k})^T \in \{0, 1\}^k \cap \mathbb{K}_k, \quad (3.4)$$

with δ_{rs} for $r, s \in \{1, \dots, k\}$ being the Kronecker delta. This one-hot encoding corresponds to a vertex of the unit simplex and therefore, each vertex of the unit simplex represents one segment. By performing the thresholding, which is essentially a projection onto a vertex of the unit simplex for any $x \in \Omega$, we obtain a partition of Ω in $\mathcal{A}$.

Due to the thresholding, the relaxation only plays a role when optimizing with respect to the partition $(\chi_{\mathcal{O}_1}, \dots, \chi_{\mathcal{O}_k})$ of Ω . Once we have found the minimizing labeling function u of J_{Zach} for fixed indicator function parameters θ , we perform the thresholding in (3.3) and the transformation into the one-hot encoding in (3.4) afterward. In particular, we reset the labeling function to (3.4), i.e., for any $x \in \Omega$, we set

$$u(x) := (\delta_{c(x)1}, \delta_{c(x)2}, \dots, \delta_{c(x)k})^T \in \{0, 1\}^k \cap \mathbb{K}_k. \quad (3.5)$$

This way, we perform a *hard assignment*, i.e., each pixel obtains exactly one label and is not considered as a mixture of different classes (called *soft assignment*).

3.1.2 Discretization

We use the approach to *discretize, then optimize*, i.e., the objective functional has to be discretized in practical applications first. Then, the resulting finite-dimensional approximation can be optimized. If the given image g has a continuous domain Ω , one can obtain a discretized version from it, e.g., by dividing Ω into regular grid cells and sampling g on these cells with a sampling function to bring it in the form

of a digital image (cf. Definition 1.2). The spectra we obtain when evaluating g on $\mathcal{N} = \{1, \dots, N_1\} \times \{1, \dots, N_2\}$ are denoted in the following by

$$g^{i,j} := g(i, j) \in W \subseteq \mathbb{R}^L$$

for $(i, j) \in \mathcal{N}$. The indicator functions and labeling function are discretized analogously.

To discretize the TV term, we assume a certain smoothness for u_l for every $l \in \{1, \dots, k\}$, namely $u_l \in H^{1,1}(\Omega)$. Then, we can use the identity in (2.10) in Remark 2.4 and obtain

$$\begin{aligned} |u_l|_{\text{BV}(\Omega)} &= \sup \left\{ \int_{\Omega} u_l(x) \operatorname{div} \varphi(x) \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^2), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\} \\ &= \sup \left\{ \int_{\Omega} -\nabla u_l(x)^T \varphi(x) \, dx \mid \varphi \in C_c^\infty(\Omega, \mathbb{R}^2), \forall x \in \Omega: \|\varphi(x)\|_2 \leq 1 \right\}. \end{aligned} \quad (3.6)$$

We use the notation

$$\begin{aligned} u_l^{i,j} &:= u_l(i, j), \\ f_l^{i,j}(\theta) &:= f_l(i, j; \theta), \end{aligned}$$

for $(i, j) \in \mathcal{N}$ and $l \in \{1, \dots, k\}$ and also

$$\begin{aligned} f_l(\theta) &:= \left(f_l^{i,j}(\theta) \right)_{\substack{i=1, \dots, N_1, \\ j=1, \dots, N_2}} \in \mathbb{R}^{N_1 \times N_2}, \\ u^{i,j} &:= (u_1^{i,j}, \dots, u_k^{i,j})^T \in \mathbb{K}_k, \\ u_l &:= (u_l^{i,j})_{\substack{i=1, \dots, N_1, \\ j=1, \dots, N_2}} \in [0, 1]^{N_1 \times N_2}, \\ u &:= (u^{i,j})_{\substack{i=1, \dots, N_1, \\ j=1, \dots, N_2}} = (u_1, \dots, u_k) \in (\mathbb{K}_k)^{N_1 \times N_2}. \end{aligned}$$

Hence, the introduced notation has the same structure as the notation in the continuous setting and is only the discrete counterpart. The gradient operator in (3.6) is approximated with its discrete counterpart [21], denoted by

$$B: \mathbb{R}^{N_1 \times N_2} \rightarrow (\mathbb{R}^{N_1 \times N_2})^2 \quad (3.7)$$

and based on forward differences, i.e., for $(i, j) \in \mathcal{N}$ and $l \in \{1, \dots, k\}$, we compute for a grid spacing $h := (h_1, h_2)$ of the underlying regular grid $\mathcal{N}$ with $h_1, h_2 > 0$ the discrete gradient of u_l at $(i, j) \in \mathcal{N}$ via

$$(Bu_l)_1^{i,j} := \begin{cases} \frac{u_l^{i+1,j} - u_l^{i,j}}{h_1} & \text{if } i < N_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$(Bu_l)_2^{i,j} := \begin{cases} \frac{u_l^{i,j+1} - u_l^{i,j}}{h_2} & \text{if } j < N_2, \\ 0 & \text{otherwise.} \end{cases}$$

With (3.6) and the set

$$P_h := \left\{ \varphi \in (\mathbb{R}^{N_1 \times N_2})^2 \mid \forall (i, j) \in \mathcal{N}: \|\varphi^{i,j}\|_2 \leq 1 \right\},$$

which is the admissible set of test vectors for the TV in the discretized setting, we define the discretized TV semi-norm as

$$|u_l|_{\text{BV}(\mathcal{N})_h} := \sup_{\varphi \in P_h} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [(Bu_l)^{i,j}]^T \varphi^{i,j}. \quad (3.8)$$

In the case of hyperspectral images, one can usually assume the same grid spacing in both directions, i.e., $h_1 = h_2$.

The resulting discrete objective functional with respect to the generic indicator functions $f_l(i, j; \theta)$, $l \in \{1, \dots, k\}$, and based on Zach's convexification in (3.2) is

$$E(u, \theta) := \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_l^{i,j}(\theta) + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}. \quad (3.9)$$

3.1.3 Optimization algorithm

We use an alternating optimization approach to minimize E with respect to u and θ . After an application of the MNF (cf. Section 1.5) for dimensionality and noise reduction, we initialize u and θ (cf. Section 3.1.4). Then, we alternate between the following two steps until the stopping criterion (cf. Section 3.1.5) is met or a maximum number of iterations $m^{\max} \in \mathbb{N}$ is reached: (1) we update θ (including the regularization in the case of ϵ AMS, cf. Section 3.3) and use the new values to compute the indicator values $f_l^{i,j}(\theta)$ for every $(i, j) \in \mathcal{N}$ and $l \in \{1, \dots, k\}$. (2) Then, the indicator values are used to update u (cf. Section 3.2) and the thresholding as in Eqs. (3.3) and (3.4) is applied, giving a partition of $\mathcal{N}$ and therefore a segmentation of g . These two steps together form one *outer* iteration in the optimization procedure. The iterations of the algorithms to optimize the labeling function u and the indicator function parameters θ are *inner* iterations. The overall segmentation procedure including our optimization framework is summarized in Algorithm 3.1.

3.1.4 Initialization

To initialize the labeling function u and the parameters of the indicator functions θ for the optimization procedure, an initial segmentation is needed. We obtain it

Algorithm 3.1: Alternating optimization of objective functional (3.9)

Input: Hyperspectral image $\tilde{g}: \mathcal{N} \rightarrow \tilde{W} \subseteq \mathbb{R}^{\tilde{L}}$, $\lambda > 0$ and hyperparameters of indicator functions;
Result: Segmentation as a one-hot encoding in u ;
Dimensionality reduction: Apply MNF to $\tilde{g}$ and obtain $g: \mathcal{N} \rightarrow W \subseteq \mathbb{R}^L$ (cf. Section 1.5);
Initialization: Initialize $u \in (\{0, 1\}^k \cap \mathbb{K}_k)^{N_1 \times N_2}$ and θ (cf. Section 3.1.4);
while *maximum number of outer iterations not reached* **and** *stopping criterion (cf. Section 3.1.5) not met* **do**
 Update θ (including regularization if applicable) and compute indicator values (cf. Section 3.3);
 Update u (cf. Section 3.2);
 Threshold u (cf. Eqs. (3.3) and (3.4));
end

from one run of the k -means algorithm with random initialization to the spectra $(g^{i,j})_{(i,j) \in \mathcal{N}} \in W^{N_1 \times N_2}$. This yields a labeling of the spectra and, therefore, also the pixels, i.e., we have for each pixel position $(i, j) \in \mathcal{N}$ a label $c(i, j) \in \{1, \dots, k\}$. This can then be turned into a one-hot encoding by the operation described in (3.4), which is used to initialize $u^{i,j}$ for every $(i, j) \in \mathcal{N}$ by (3.5). The initialization of θ is also based on the initial segmentation since it gives a rough estimate, which pixels can be considered similar. To initialize θ , we go through the segments and estimate the parameters of the respective segment based on the pixels corresponding to that segment in the initial segmentation. Detailed descriptions for the indicator function parameters of ϵ AMS, ϕ AMS and kMS are given in Sections 3.3.1 to 3.3.3, respectively.

3.1.5 Stopping criterion

A natural idea is to stop iterating once the segmentation, which contains the discrete class labels of the pixels, does not change anymore. This, however, is not a suitable criterion in our case. The problem is caused by the relaxation of u and the thresholding in (3.3). In a late outer iteration after an update of u , it can occur that for some $(i, j) \in \mathcal{N}$ there are indices $l, n \in \{1, \dots, k\}$ of segments such that $u_l^{i,j} \approx u_n^{i,j}$ and $u_l^{i,j} > u_r^{i,j}$ and $u_n^{i,j} > u_r^{i,j}$ for any $r \in \{1, \dots, k\} \setminus \{l, n\}$, i.e., at pixel (i, j) the two segments with indices l, n obtain the highest and approximately the same values by the labeling function. If we now apply the thresholding, we will get a label, say l , for the pixel at (i, j) . Since we consider a late outer iteration, we are quite close to a local minimum. Thus, the model parameters will be adjusted only slightly and after the next update of u we can be in almost the same situation, only that now

the other segment n becomes the label of (i, j) . This can result in an alternation between two segmentations where the label of (i, j) jumps between l and n , causing the method to converge extremely slowly or even preventing it from converging. Therefore, it is better to rely on an aggregating statistic that is not discrete. We base our stopping criterion on the change in the mean spectra of the segments. However, since in segments with only a few pixels a change in some pixels can have a greater impact on the mean spectrum of that segment than in segments with many pixels, we weight the differences of last and current mean estimate with the number of pixels in that segment. In this way, we can reduce this effect.

Let $m \in \mathbb{N}$ describe the current outer iteration. Please note that we can assume that $u^{(m)} \in (\{0, 1\}^k \cap \mathbb{K}_k)^{N_1 \times N_2}$, i.e., $(u^{(m)})^{i,j}$ lies on one vertex of $\mathbb{K}_k$ and $(u^{(m)})_l^{i,j} \in \{0, 1\}$ for any $(i, j) \in \mathcal{N}$ and $l \in \{1, \dots, k\}$. The reason is that the check of the stopping criterion is done after the thresholding and the reset of u to one of the vertices of the unit simplex (cf. Eqs. (3.3) and (3.4)). To measure the change in the segmentation, we use the expression

$$\sum_{l=1}^k \frac{n_l^{(m)}}{N_1 N_2} \left\| \bar{\mu}_l^{(m)} - \bar{\mu}_l^{(m-1)} \right\|_{\infty},$$

where

$$n_l^{(m)} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left(u^{(m)} \right)_l^{i,j}$$

is the number of pixels in the l -th segment and

$$\bar{\mu}_l^{(m)} := \frac{1}{n_l^{(m)}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left(u^{(m)} \right)_l^{i,j} g^{i,j}$$

the empirical mean (and not the estimated mean from the update of the parameters of the indicator functions such that we stop after an update of the segmentation) of the spectra in the l -th segment according to the segmentation after the m -th outer iteration.

3.2 Optimization with respect to the labeling function

As commented in Section 3.1.1, Zach's convexification makes the objective functional in (3.9) convex in u . Methods for optimization like gradient flows only require differentiability and are therefore applicable in a broad range of problems; however, they are unable to exploit structure like convexity and can thus converge slowly. Special

algorithms designed for convex optimization can accelerate the numerical computation of minimizers significantly and sometimes do not even require differentiability. This is of special interest in our case because the TV is convex but non-differentiable. We will now explain the *primal-dual hybrid gradient* (PDHG) method, proposed by Chambolle and Pock [23], that we will use to optimize with respect to the labeling function u . It is a primal-dual method that is designed for convex and non-smooth optimization problems and based on evaluations of the so-called *proximal operator*. In Section 3.2.1, the PDHG method is explained in detail before the formulas for our concrete case are derived in Section 3.2.2.

3.2.1 The PDHG method

The introduction of the PDHG method requires some preparation since it can only be applied to a certain class of functions, namely, the class of proper, closed and convex functions. We start with introducing these functions, which then allows for an explanation of the proximal operator that needs functions of this class to be well-defined. In the following, we consider functions $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ on a finite-dimensional real vector space X with inner product $\langle \cdot, \cdot \rangle_X$ and induced norm $\|\cdot\|_X = \sqrt{\langle \cdot, \cdot \rangle_X}$.

Remark 3.1.

Chambolle and Pock introduce the PDHG method in their article only for finite-dimensional real spaces but point out that most of the theory is also correct in general real Hilbert spaces. We will also restrict our considerations to finite-dimensional real spaces. In particular, several of the following definitions and results also hold in general spaces, but we formulate only special cases of them according to our setting. A consequence of the finite-dimensional setting is that the considered space X is a Hilbert space because every finite-dimensional normed real space is complete. The Riesz representation theorem, therefore, yields that X is isomorphic to its dual space and for each functional from the dual space exists a unique element in X such that we can represent the functional from the dual space with its representative in X and the inner product on X . Consequently, we will identify each element of the dual space with an element in X and formulate definitions and results in this section in terms of elements from X , the dual space will be used only implicitly. That is, different from the definitions given in this thesis, the subgradient is generally an element from the dual space (cf. Definition A.2) and also the Fenchel conjugate (cf. Definition 3.10) maps from the dual space into the real numbers in general.

Definition 3.2 (Proper, closed, convex functional [8, Def. 2.2]).

We define the effective domain and the epigraph of f as

$$\text{dom}(f) := \{x \in X \mid f(x) < \infty\}$$

and

$$\text{epi}(f) := \{(x, t) \in X \times \mathbb{R} \mid f(x) \leq t\},$$

respectively. We call f a *proper functional* if there exists an $x \in X$ such that $f(x) < \infty$ which means that $\text{dom}(f) \neq \emptyset$. The functional f is said to be *closed* if $\text{epi}(f)$ is a closed set. We describe with $\Gamma_0(X)$ the set of closed, proper and convex functionals on X .

For proper functions f , one can equivalently define convexity (cf. Definition A.5) by requiring that $\text{epi}(f)$ is a convex set [8].

Proposition 3.3 ([8, Th. 2.6]).

A function $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ is closed if and only if it is lower semi-continuous.

Hence, we can replace the closedness requirement in the definition of $\Gamma_0(X)$ with the requirement that a function in $\Gamma_0(X)$ is lower semi-continuous. Furthermore, the closedness of functions is preserved under summation and positive scaling.

Proposition 3.4 (Summation preserves closedness [8, Th. 2.7 (b)]).

Let $a_1, \dots, a_m \in \mathbb{R}_{>0}$ and consider closed functions $f_1, \dots, f_m: X \rightarrow \mathbb{R} \cup \{\infty\}$. Then, we have that the function $f = \sum_{i=1}^m a_i f_i$ is also closed.

Definition 3.5 (Proximal operator [8, Def. 6.1]).

We define the proximal operator of a function $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ evaluated at $x \in X$ as

$$\begin{aligned} \text{prox}_f: X &\rightarrow P(X), \\ x &\mapsto \arg \min_{u \in X} \left(f(u) + \frac{1}{2} \|u - x\|_X^2 \right), \end{aligned}$$

where $P(X)$ describes the power set of X .

At the moment, this operator returns a set that can be empty, a singleton or contain several elements. However, if f is proper, closed and convex, prox_f always returns a singleton.

Proposition 3.6 ([8, Th. 6.3, first prox theorem]).

If $f \in \Gamma_0(X)$, then the proximal operator prox_f evaluated at any $x \in X$ returns a singleton.

Hence, for functions from $\Gamma_0(X)$, we can consider prox_f as a function from X to X , i.e.,

$$\begin{aligned} \text{prox}_f: X &\rightarrow X, \\ x &\mapsto \arg \min_{u \in X} \left(f(u) + \frac{1}{2} \|u - x\|_X^2 \right). \end{aligned}$$

Although we have to solve an optimization problem to evaluate the proximal operator, there are functions that are simple in the sense that their proximal operators can be computed in a closed form. This is what we will exploit later in Section 3.2.2 when applying the proximal operator in the PDHG algorithm. If we weight f with a positive scalar $\tau > 0$ and consider $\text{prox}_{\tau f}$, we find

$$\text{prox}_{\tau f}(x) = \arg \min_{u \in X} \left(\tau f(u) + \frac{1}{2} \|u - x\|_X^2 \right) = \arg \min_{u \in X} \left(f(u) + \frac{1}{2\tau} \|u - x\|_X^2 \right).$$

We can observe that τ introduces a weighting of the summands in the proximal operator and controls how far away from $x \in X$ we are allowed to look for a minimizing $u \in X$. In this sense, such a scaling factor can be understood as a step size in algorithms using the proximal operator. The proximal operator can be related to the subdifferential of f (cf. Definition A.2). We recall that the subdifferential yields an optimality condition for convex functions:

Proposition 3.7 ([8, Th. 3.63]).

If $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ is a proper and convex function, then we have that $x^* \in X$ minimizes f if and only if $0 \in \partial f(x^*)$.

This implies the following equivalence:

Proposition 3.8 ([8, Th. 6.39, second prox theorem]).

If $f \in \Gamma_0(X)$, then we have for any $x, u \in X$ that $u = \text{prox}_f(x)$ if and only if $x - u \in \partial f(u)$.

A direct consequence of the two preceding results is that $x \in X$ minimizes $f \in \Gamma_0(X)$ if and only if $x = \text{prox}_f(x)$. We also recall that the subdifferential obeys the following calculus rule:

Proposition 3.9 ([8, Th. 3.35]).

If $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ is a proper function and $\alpha > 0$, then the equation

$$\partial(\alpha f)(x) = \alpha \partial f(x)$$

is true for any $x \in \text{dom}(f)$.

We use the *Fenchel duality* to derive later a primal-dual formulation for the optimization problem that is tackled by PDHG method:

Definition 3.10 (Fenchel conjugate [8, Def. 4.1]).

For the function $f: X \rightarrow \mathbb{R} \cup \{\infty\}$, we define the function

$$f^*: X \rightarrow \mathbb{R} \cup \{\infty\},$$

$$y \mapsto \sup_{x \in X} (\langle y, x \rangle_X - f(x)),$$

which we call the Fenchel conjugate of f .

We need to apply Proposition 3.7 to the Fenchel conjugate of $f \in \Gamma_0(X)$ which requires that f^* is proper and convex. This is guaranteed by the following proposition:

Proposition 3.11 ([8, Th. 4.3]).

The Fenchel conjugate f^* of a function $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ is closed and convex. If f is additionally proper and convex, f^* is proper.

We will later apply the Fenchel conjugate to f^* . For functions in $\Gamma_0(X)$, it holds:

Proposition 3.12 ([8, Th. 4.8]).

Let $f \in \Gamma_0(X)$. Then, we have $f^{**} = f$.

We will now derive the PDHG algorithm. To this end, we consider two finite-dimensional spaces X and Y with inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$ as well as induced norms $\|\cdot\|_X = \sqrt{\langle \cdot, \cdot \rangle_X}$ and $\|\cdot\|_Y = \sqrt{\langle \cdot, \cdot \rangle_Y}$. Moreover, let $G \in \Gamma_0(X)$ and $H \in \Gamma_0(Y)$ and $B: X \rightarrow Y$ a linear, continuous operator with the induced norm

$$\|B\|_{X \rightarrow Y} = \max_{\substack{x \in X, \\ \|x\|_X \leq 1}} \|Bx\|_Y. \quad (3.10)$$

The problem type we want to solve is

$$\underset{x \in X}{\text{minimize}} \quad G(x) + H(Bx). \quad (3.11)$$

This kind of problem often appears in image processing and, in particular, also our optimization problem, namely the optimization of E with respect to u , can be cast into this form. Since $H \in \Gamma_0(Y)$, we obtain by Proposition 3.12 that $H^{**} = H$. Using the linearity and symmetry of the inner product, this allows us to reformulate the optimization problem:

$$\inf_{x \in X} G(x) + H(Bx)$$

$$\stackrel{\text{Proposition 3.12}}{=} \inf_{x \in X} G(x) + H^{**}(Bx)$$

$$\begin{aligned}
\text{Definition 3.10} &= \inf_{x \in X} G(x) + \sup_{y \in Y} (\langle Bx, y \rangle_Y - H^*(y)) \\
&= \inf_{x \in X} \sup_{y \in Y} (\langle Bx, y \rangle_Y + G(x) - H^*(y)). \tag{3.12}
\end{aligned}$$

In (3.12), we have a saddle-point problem, which is the primal-dual formulation of the original problem in (3.11). It is possible to derive necessary and sufficient conditions for a solution of the primal-dual problem. Let $B^\dagger$ denote the adjoint operator of B .

Proposition 3.13 ([95, Th. 31.3]).

Let $(x, y) \in X \times Y$. The element (x, y) is a solution of the primal-dual problem in (3.12) if (x, y) fulfills the following necessary and sufficient conditions

$$Bx \in \partial H^*(y), \quad -B^\dagger y \in \partial G(x).$$

Hence, assuming that $(x, y) \in X \times Y$ solves the primal-dual problem and having in mind that scaling a proper, closed and convex function with a positive number is still a proper, closed and convex function, we can reformulate the conditions equivalently for $\tau, \kappa > 0$:

$$\begin{aligned}
& Bx \in \partial H^*(y) \\
& \Leftrightarrow \kappa Bx \in \kappa \partial H^*(y) \\
\text{Proposition 3.9} & \Leftrightarrow \kappa Bx \in \partial (\kappa H^*)(y) \\
& \Leftrightarrow y + \kappa Bx - y \in \partial (\kappa H^*)(y) \\
\text{Proposition 3.8} & \Leftrightarrow y = \text{prox}_{\kappa H^*}(y + \kappa Bx)
\end{aligned}$$

and

$$\begin{aligned}
& -B^\dagger y \in \partial G(x) \\
& \Leftrightarrow -\tau B^\dagger y \in \tau \partial G(x) \\
\text{Proposition 3.9} & \Leftrightarrow -\tau B^\dagger y \in \partial (\tau G)(x) \\
& \Leftrightarrow x - \tau B^\dagger y - x \in \partial (\tau G)(x) \\
\text{Proposition 3.8} & \Leftrightarrow x = \text{prox}_{\tau G}(x - \tau B^\dagger y).
\end{aligned}$$

Propositions 3.8 and 3.9 are applicable in the first chain of equivalences because $H^* \in \Gamma_0(Y)$ by Proposition 3.11. The conditions written in terms of the proximal operators motivate the PDHG algorithm [23] with the following update in the m -th iteration, $m \in \mathbb{N}_0$, including an extrapolation step

$$\begin{aligned}
y^{(m+1)} &= \text{prox}_{\kappa H^*} \left(y^{(m)} + \kappa B\bar{x}^{(m)} \right), \\
x^{(m+1)} &= \text{prox}_{\tau G} \left(x^{(m)} - \tau B^\dagger y^{(m+1)} \right), \\
\bar{x}^{(m+1)} &= x^{(m+1)} + \vartheta \left(x^{(m+1)} - x^{(m)} \right), \tag{3.13}
\end{aligned}$$

where $\vartheta \in [0, 1]$ describes a scaling factor for the extrapolation step in the last line and $\bar{x}^{(0)} = x^{(0)} \in X$ and $y^{(0)} \in Y$ are initial guesses. One can observe that the decomposition of the second summand in the objective function (cf. (3.11)) into H and B has the effect that one only needs to compute the proximal operator of H^* but not of B . This algorithm has a guaranteed convergence for certain choices of τ , κ and ϑ if a solution of the primal-dual problem exists.

Proposition 3.14 (Convergence of the PDHG algorithm [23, Th. 1 (c)]).

Let the primal-dual problem in (3.12) have a solution. Set $L_0 := \|B\|_{X \rightarrow Y}$ and $\vartheta = 1$ and choose $\tau, \kappa > 0$ such that $\tau\kappa L_0^2 < 1$. Then, there exists a saddle-point $(x^*, y^*) \in X \times Y$ of (3.12) such that the sequence of points $(x^{(m)}, y^{(m)})_{m \in \mathbb{N}} \subseteq X \times Y$ generated by the updates in (3.13) satisfies

$$(x^{(m)}, y^{(m)}) \rightarrow (x^*, y^*) \text{ when } m \rightarrow \infty.$$

In our numerical experiments, we stop applying the update rules in (3.13) if a maximum number of iterations $m_{\text{PDHG}}^{\max} \in \mathbb{N}$ is reached or the inequality

$$\|x^{(m+1)} - x^{(m)}\|_{\infty} < t_{\text{PDHG}}$$

for a threshold $t_{\text{PDHG}} > 0$ is fulfilled. It is enough to consider iterates of x in the stopping criterion since there is a mutual dependence between the iterates of x and y in the PDHG update rules (cf. (3.13)).

3.2.2 Application of PDHG method to MS model

The PDHG method is a popular tool in image processing. In particular, it is suitable to minimize Zach's convexification of the MS model with respect to u . To this end, we will formulate our minimization problem in the form of (3.11), show that the functions G and H^* are proper, closed and convex in our case such that the proximal operators are well-defined and compute the corresponding proximal operators according to (3.13).

Let the parameters θ of the indicator functions $f_1(\theta), \dots, f_k(\theta)$ be arbitrary but fixed. Then, also the resulting indicator values are fixed. We, therefore, omit the argument in the following and only write $f_1, \dots, f_k \in \mathbb{R}^{N_1 \times N_2}$ and $E(u)$ instead of $E(u, \theta)$ for the discrete objective functional (cf. (3.9)), which is then stated as

$$E(u) = \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_l^{i,j} + \lambda |u|_{\text{BV}(\mathcal{N})_h}.$$

The optimization problem we want to apply the PDHG method to is

$$\begin{aligned} & \underset{u \in (\mathbb{K}_k)^{N_1 \times N_2}}{\text{minimize}} && E(u). \end{aligned} \tag{3.14}$$

A reformulation is needed to bring it into the form considered in the PDHG algorithm. We define

$$f := (f_1, \dots, f_k) \in \left(\mathbb{R}^k\right)^{N_1 \times N_2},$$

consider

$$P_h^k = \left\{ (\varphi_1, \dots, \varphi_k) \in \left[\left(\mathbb{R}^k\right)^{N_1 \times N_2} \right]^2 \mid \forall l \in \{1, \dots, k\} \forall (i, j) \in \mathcal{N}: \|\varphi_l^{i,j}\|_2 \leq 1 \right\},$$

and the discrete gradient operator from (3.7) for a grid spacing $h := (h_1, h_2)$ with $h_1, h_2 > 0$ for $(i, j) \in \mathcal{N}$ and $l \in \{1, \dots, k\}$ such that it operates on the full u instead of u_l for segment l , i.e.,

$$B: \left(\mathbb{R}^k\right)^{N_1 \times N_2} \rightarrow \left[\left(\mathbb{R}^k\right)^{N_1 \times N_2} \right]^2$$

and

$$(Bu)_{l,1}^{i,j} := \begin{cases} \frac{u_l^{i+1,j} - u_l^{i,j}}{h_1} & \text{if } i < N_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$(Bu)_{l,2}^{i,j} := \begin{cases} \frac{u_l^{i,j+1} - u_l^{i,j}}{h_2} & \text{if } j < N_2, \\ 0 & \text{otherwise.} \end{cases}$$

The domain of B as well as the space containing the image of B are written with several exponents to emphasize that we virtually keep the structure of u , i.e., B is applied to u that has the same spatial structure as the considered image g since it assigns to each pixel a point in the unit simplex, and B returns an object that also has the spatial structure as g and at each pixel position we now have k two-dimensional vectors representing the approximated spatial gradients of each u_l for $l \in \{1, \dots, k\}$. However, it would also be possible to roll out u to a long vector and let B operate on that vector and return a vector of twice the size since for each pixel and each segment, we have one gradient of dimension two. Therefore, the considered finite-dimensional real vector spaces X and Y in Section 3.2.1 are in our setting

$$X := \left(\mathbb{R}^k\right)^{N_1 \times N_2}, \quad Y := \left[\left(\mathbb{R}^k\right)^{N_1 \times N_2} \right]^2.$$

The discrete TV semi-norm is given in this setting as

$$|u|_{\text{BV}(\mathcal{N})_h} := \sup_{\varphi \in P_h^k} \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left[(Bu)_l^{i,j} \right]^T \varphi_l^{i,j}.$$

The expression in the supremum, including the triple sum, is essentially the Euclidean inner product of Bu and φ when rolling out both tensors to vectors. Therefore, we equip Y with the inner product

$$\langle y, w \rangle_Y := \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left[y_l^{i,j} \right]^T w_l^{i,j}, \quad (3.15)$$

resulting in

$$|u|_{\text{BV}(\mathcal{N})_h} = \sup_{\varphi \in P_h^k} \langle Bu, \varphi \rangle_Y.$$

For X , we choose the inner product

$$\langle x, v \rangle_X := \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} x_l^{i,j} v_l^{i,j}, \quad (3.16)$$

which also coincides with the usual Euclidean inner product when rolling out the arguments to vectors. These preparations together with Remark 2.24 allow us to rewrite the optimization problem in (3.14).

$$\begin{aligned} & \arg \min_{u \in (\mathbb{K}_k)^{N_1 \times N_2}} E(u) \\ &= \arg \min_{u \in (\mathbb{K}_k)^{N_1 \times N_2}} \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_l^{i,j} + \lambda |u|_{\text{BV}(\mathcal{N})_h} \\ &= \arg \min_{u \in (\mathbb{K}_k)^{N_1 \times N_2}} \langle u, f \rangle_X + \lambda \sup_{\varphi \in P_h^k} \langle Bu, \varphi \rangle_Y \\ &= \arg \min_{u \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{\lambda} \langle u, f \rangle_X + \sup_{\varphi \in P_h^k} \langle Bu, \varphi \rangle_Y \\ \text{Remark 2.24} &= \arg \min_{u \in X} \frac{1}{\lambda} \langle u, f \rangle_X + \delta_{(\mathbb{K}_k)^{N_1 \times N_2}}(u) + \sup_{\varphi \in Y} \langle Bu, \varphi \rangle_Y - \delta_{P_h^k}(\varphi). \end{aligned}$$

If we now define the functions

$$G(u) := \frac{1}{\lambda} \langle u, f \rangle_X + \delta_{(\mathbb{K}_k)^{N_1 \times N_2}}(u), \quad (3.17)$$

$$H(q) := \sup_{\varphi \in Y} \langle q, \varphi \rangle_Y - \delta_{P_h^k}(\varphi), \quad (3.18)$$

we obtain

$$\arg \min_{u \in (\mathbb{K}_k)^{N_1 \times N_2}} E(u) = \arg \min_{u \in X} G(u) + H(Bu),$$

which is exactly the form we need to apply the PDHG method.

A requirement on G and H to apply the PDHG method is that both functions are proper, closed and convex. We now compute H^* which is needed for the application of the PDHG method and argue that $G \in \Gamma_0(X)$ and $H \in \Gamma_0(Y)$. To this end, we need two properties of the indicator functional δ_A and the closedness and convexity of the sets $(\mathbb{K}_k)^{N_1 \times N_2}$ and P_h^k .

Lemma 3.15 (Closedness of indicator functional [8, Prop. 2.3]).

Let V be a vector space. The indicator functional δ_A (cf. Remark 2.24) of a set $A \subseteq V$ is a closed function if and only if the set A is closed.

Lemma 3.16 (Convexity of indicator functional [8]).

Let V be a vector space. The indicator functional δ_A (cf. Remark 2.24) of a set $A \subseteq V$ is a convex function if and only if the set A is convex.

Lemma 3.17 (Closedness and convexity of the unit simplex).

The set $(\mathbb{K}_k)^{N_1 \times N_2}$ is closed and convex.

Proof. We obtain the closedness of $\mathbb{K}_k$ because the continuous equalities and continuous non-strict inequalities in the definition of $\mathbb{K}_k$ (cf. (3.1)) are preserved in the limit. Then, also the Cartesian product $(\mathbb{K}_k)^{N_1 \times N_2}$ is closed.

The set $\mathbb{K}_k$ is convex because linear conditions are preserved under convex combinations. Consequently, also $(\mathbb{K}_k)^{N_1 \times N_2}$ as the Cartesian product of convex sets is convex. $\square$

Lemma 3.18 (Closedness and convexity of P_h^k).

The set P_h^k is closed and convex.

Proof. We note that by Lemma A.6 norms are convex and continuous. The convexity of P_h^k follows from the convexity of norms because for $v, w \in P_h^k$ and $\alpha \in [0, 1]$, we have that $\|v_l^{i,j}\|_2 \leq 1$ and $\|w_l^{i,j}\|_2 \leq 1$, and therefore

$$\|(\alpha v + (1 - \alpha)w)_l^{i,j}\|_2 \leq \alpha \|v_l^{i,j}\|_2 + (1 - \alpha) \|w_l^{i,j}\|_2 \leq \alpha + (1 - \alpha) = 1$$

for every $l \in \{1, \dots, k\}$ and $(i, j) \in \mathcal{N}$. Therefore, $\alpha v + (1 - \alpha)w \in P_h^k$.

The closedness of P_h^k follows from the continuity of the norm and because the non-strict inequality in the definition of P_h^k is preserved in the limit as it depends continuously on the considered element of P_h^k . $\square$

We start with showing that $G \in \Gamma_0(X)$. By Lemmas 3.15 to 3.17, we obtain that $\delta_{(\mathbb{K}_k)^{N_1 \times N_2}}$ is closed and convex. It is furthermore proper because $(\mathbb{K}_k)^{N_1 \times N_2}$ is non-empty. The function $\frac{1}{\lambda} \langle \cdot, f \rangle_X$ is linear and therefore convex. It is also continuous which can be shown with the Cauchy-Schwarz inequality. In particular, the continuity implies the lower semi-continuity, and we obtain by Proposition 3.3 that it is closed. Since inner products attain only finite values, the function is also proper and its effective domain is $(\mathbb{R}^k)^{N_1 \times N_2}$. Therefore, both terms of G are proper, closed and convex. By Proposition 3.4 and because sums of convex functions are convex, we have that G is closed and convex. It is also proper since $\text{dom}(G)$ is only determined by $\delta_{(\mathbb{K}_k)^{N_1 \times N_2}}$ and, thus, equals $(\mathbb{K}_k)^{N_1 \times N_2}$. It follows $G \in \Gamma_0(X)$.

To compute the Fenchel conjugate of H , we observe that P_h^k is obviously non-empty and therefore $\delta_{P_h^k}$ is proper. By Lemmas 3.15, 3.16 and 3.18, we obtain that $\delta_{P_h^k}$ is also closed and convex, and hence $\delta_{P_h^k} \in \Gamma_0(Y)$. This implies together with Proposition 3.11 that $\delta_{P_h^k}^* \in \Gamma_0(Y)$. Since H is defined (cf. (3.18)) as the Fenchel conjugate of $\delta_{P_h^k}$ (cf. Definition 3.10), it follows $H = \delta_{P_h^k}^* \in \Gamma_0(Y)$. Another consequence of $\delta_{P_h^k} \in \Gamma_0(Y)$ is that Proposition 3.12 yields $\delta_{P_h^k}^{**} = \delta_{P_h^k}$. This observation results in

$$H^* = \delta_{P_h^k}^{**} = \delta_{P_h^k}.$$

Since G and H are proper, closed and convex functions, the PDHG method is applicable in our setting. To compute the necessary proximal operators, we need to determine the proximal operator of an indicator functional δ_A for a set A , which, in fact, equals the orthogonal projection onto A . This is summarized in the following remark.

Remark 3.19 (Proximal operator of an indicator functional [8, Th. 6.24, Th. 6.25]). *Let $\mathbb{P}_A$ denote the orthogonal projection onto a non-empty, convex and closed subset A of a finite-dimensional Hilbert space V with induced norm $\|\cdot\|_V$. Then, the proximal operator prox_{δ_A} of the indicator functional δ_A evaluated at $x \in V$ is*

$$\text{prox}_{\delta_A}(x) = \arg \min_{y \in V} \left(\delta_A(y) + \frac{1}{2} \|y - x\|_V^2 \right) = \arg \min_{y \in A} \frac{1}{2} \|y - x\|_V^2.$$

This coincides with the usual definition of the orthogonal projection onto convex and closed sets. In particular, this remains true if we consider $\text{prox}_{\alpha\delta_A}$ for $\alpha > 0$ because we have $\alpha\delta_A = \delta_A$. Since A is assumed to be non-empty, closed and convex, δ_A is a proper, closed and convex functional by Lemmas 3.15 and 3.16, yielding the well-definedness of prox_{δ_A} by Proposition 3.6. Therefore, we have $\text{prox}_{\delta_A} = \mathbb{P}_A$.

With this remark we can directly compute $\text{prox}_{\kappa H^*}$ for $\kappa > 0$ and get

$$\text{prox}_{\kappa H^*}(q) = \mathbb{P}_{P_h^k}(q)$$

for any $q \in Y$. The computation of $\text{prox}_{\tau G}$ with $\tau > 0$ requires more work. We will start with the orthogonal projection onto the unit simplex evaluated at a specific point and then show that this coincides with $\text{prox}_{\tau G}(u)$. We derive with the linearity and symmetry of the inner product

$$\begin{aligned}
& \mathbb{P}_{(\mathbb{K}_k)^{N_1 \times N_2}} \left(u - \frac{\tau}{\lambda} f \right) \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left\| u - \frac{\tau}{\lambda} f - v \right\|_X^2 \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left\langle u - \frac{\tau}{\lambda} f - v, u - \frac{\tau}{\lambda} f - v \right\rangle_X \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left(\left\langle u - v, u - \frac{\tau}{\lambda} f - v \right\rangle_X - \left\langle \frac{\tau}{\lambda} f, u - \frac{\tau}{\lambda} f - v \right\rangle_X \right) \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left(\langle u - v, u - v \rangle_X - \left\langle u - v, \frac{\tau}{\lambda} f \right\rangle_X - \left\langle \frac{\tau}{\lambda} f, u - \frac{\tau}{\lambda} f \right\rangle_X + \left\langle \frac{\tau}{\lambda} f, v \right\rangle_X \right) \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left(\|u - v\|_X^2 - \left\langle u, \frac{\tau}{\lambda} f \right\rangle_X + 2 \left\langle v, \frac{\tau}{\lambda} f \right\rangle_X - \left\langle \frac{\tau}{\lambda} f, u - \frac{\tau}{\lambda} f \right\rangle_X \right).
\end{aligned}$$

The terms that are independent of v do not change the minimizing v and can therefore be omitted. If we additionally incorporate the constraint on v into the optimization problem by adding $\delta_{(\mathbb{K}_k)^{N_1 \times N_2}}$ (cf. Remark 2.24), we arrive at

$$\begin{aligned}
& \mathbb{P}_{(\mathbb{K}_k)^{N_1 \times N_2}} \left(u - \frac{\tau}{\lambda} f \right) \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \left(\|u - v\|_X^2 + 2 \left\langle v, \frac{\tau}{\lambda} f \right\rangle_X \right) \\
&= \arg \min_{v \in (\mathbb{K}_k)^{N_1 \times N_2}} \frac{1}{2} \|u - v\|_X^2 + \frac{\tau}{\lambda} \langle v, f \rangle_X + \delta_{(\mathbb{K}_k)^{N_1 \times N_2}}(v) \\
(3.17) \quad &= \arg \min_{v \in (\mathbb{R}^k)^{N_1 \times N_2}} \frac{1}{2} \|u - v\|_X^2 + \tau G(v)
\end{aligned}$$

Definition 3.5 $= \text{prox}_{\tau G}(u)$.

Hence, we have computed the proximal operators for G and H^* . The orthogonal projection $\mathbb{P}_{P_h^k}(q)$ can be easily computed for any $l \in \{1, \dots, k\}$ and $(i, j) \in \mathcal{N}$ by

$$\left(\mathbb{P}_{P_h^k}(q) \right)_l^{i,j} = \frac{1}{\max \left\{ \left\| q_l^{i,j} \right\|_2, 1 \right\}} q_l^{i,j}.$$

The computation of the orthogonal projection $\mathbb{P}_{(\mathbb{K}_k)^{N_1 \times N_2}}$ onto the unit simplex requires more work. Nevertheless, there exist efficient algorithms to solve that problem. One example is the method proposed in [106] with a computational complexity of $\mathcal{O}(k \log k)$ that we use in our numerical experiments.

Since we are in a finite-dimensional real setting, the adjoint operator of the discrete gradient B that we apply in the update rules of the PDHG method in (3.13) can be computed as the transpose of the discrete gradient operator. It turns out to coincide with the negative discrete divergence operator and is described in [21] for $k = 1$. By applying this for each segment $l \in \{1, \dots, k\}$ separately, the negative discrete divergence operator is given in our case as

$$B^\dagger: \left[(\mathbb{R}^k)^{N_1 \times N_2} \right]^2 \rightarrow (\mathbb{R}^k)^{N_1 \times N_2}$$

and

$$(B^\dagger y)_l^{i,j} := -\frac{1}{h_1} \begin{cases} y_{l,1}^{i,j} & \text{if } i = 1, \\ y_{l,1}^{i,j} - y_{l,1}^{i-1,j} & \text{if } 1 < i < N_1, \\ -y_{l,1}^{i-1,j} & \text{if } i = N_1, \end{cases} \\ -\frac{1}{h_2} \begin{cases} y_{l,2}^{i,j} & \text{if } j = 1, \\ y_{l,2}^{i,j} - y_{l,2}^{i,j-1} & \text{if } 1 < j < N_2, \\ -y_{l,2}^{i,j-1} & \text{if } j = N_2, \end{cases}$$

for any $l \in \{1, \dots, k\}$ and $(i, j) \in \mathcal{N}$. In particular, this is in accordance with the continuous case in which one can show that $\nabla^\dagger = -\text{div}$, see for example [15, Th. 6.88].

Remark 3.20 (Operator norm of discrete gradient operator).

In order to have guaranteed convergence by Proposition 3.14, it is necessary to know $L_0 = \|B\|_{X \rightarrow Y}$ such that we can choose step sizes τ and κ that fulfill the requirements. One can show that with an analogue computation as in [21], i.e., using that $\|B\|_{X \rightarrow Y}^2 = \|B^\dagger\|_{Y \rightarrow X}^2$ [4] and the analogue inequality of (2.40) for convex functions, but with the norms on X and Y induced by (3.16) and (3.15), respectively, it holds $L_0^2 = \|B\|_{X \rightarrow Y}^2 \leq 8/\min\{h_1, h_2\}^2$. Hence, the product of τ and κ needs to be smaller than $\min\{h_1, h_2\}^2/8$ for a convergence guarantee.

3.3 Optimization with respect to parameters of indicator functions

From Section 3.2, we now know how to optimize with respect to the labeling function u . It remains to explain how to optimize with respect to the parameters of the indicator

functions, which is the second step in the alternating optimization procedure, that is described in Section 3.1.3, to optimize the objective functional. In this section, detailed descriptions of the optimization procedures for the indicator functions of the models ϵ AMS (cf. Section 3.3.1), ϕ AMS (cf. Section 3.3.2) and kMS (cf. Section 3.3.3) are given.

As in Section 3.1.5, we will always assume that $u \in (\{0, 1\}^k \cap \mathbb{K}_k)^{N_1 \times N_2}$ throughout this section which is possible because the update of u in the outer iteration is directly followed by the thresholding which projects u onto a vertex of the unit simplex. Furthermore, for every $l \in \{1, \dots, k\}$ we set

$$n_l := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}$$

for the remainder of the chapter, which equals the number of pixels in the l -th segment.

3.3.1 Parameters of ϵ AMS

Based on the notation of Section 2.2, the discretized and relaxed version of J_ϵ (cf. (2.20)) is

$$\begin{aligned} & E_\epsilon(u, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k) \\ &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_{\eta,l}^{i,j}(\mu_l, \Sigma_l) + \lambda |u_l|_{\text{BV}(\mathcal{N})_h} \\ &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left[\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \right] + \lambda |u_l|_{\text{BV}(\mathcal{N})_h} \\ &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left[\sqrt{(g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) + \eta} + \log \det \Sigma_l \right] + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}. \end{aligned} \tag{3.19}$$

For a given u , this function has to be optimized with respect to the segments' means $\mu_1, \dots, \mu_k$ and covariance matrices $\Sigma_1, \dots, \Sigma_k$. The square root in the indicator function makes it difficult to derive a closed-form solution for the parameters and to the best of our knowledge, such a solution for this specific functional has not been found yet. We will, therefore, derive a fixed point iteration scheme based on the gradient to optimize these parameters. One advantage of the fixed point iteration that we introduce below is that it is independent of the scaling of the gradient and, hence, does not need a step size.

For an arbitrary but fixed $l \in \{1, \dots, k\}$, we need the derivatives of E_ϵ with respect to μ_l and Σ_l . We observe that the summands of the outer sum are uncoupled with respect to these parameters and therefore μ_l and Σ_l only appear in the l -th summand and the others can be omitted in the computation as their derivative with respect to these parameters is 0. In addition, the entire TV term only depends on $u_1, \dots, u_k$ and is independent of μ_l and Σ_l for any $l \in \{1, \dots, k\}$. Hence, we can and will omit also this term in the following computations of the derivatives.

Fixed point equation of μ_l

We first note that $(\Sigma_l^{-1})^T = \Sigma_l^{-1}$ because the inverse of a symmetric matrix is again symmetric. When computing the gradient of E_ϵ with respect to μ_l using the linearity of the derivative, the chain rule, Lemma A.1 (b) and the symmetry of Σ_l^{-1} , we obtain

$$\begin{aligned}
& \partial_{\mu_l} E_\epsilon(u, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\mu_l} \left[\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \right] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\mu_l} \sqrt{(g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) + \eta} \\
&\stackrel{\text{Lemma A.1 (b)}}{=} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \frac{1}{2} \frac{1}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} \left[\Sigma_l^{-1} (\mu_l - g^{i,j}) + (\Sigma_l^{-1})^T (\mu_l - g^{i,j}) \right] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \frac{1}{2} \frac{1}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} 2 \Sigma_l^{-1} (\mu_l - g^{i,j}) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} \Sigma_l^{-1} (\mu_l - g^{i,j}).
\end{aligned}$$

Setting the derivative to 0 and isolating one μ_l yields

$$\begin{aligned}
0 &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} \mu_l - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} g^{i,j} \\
\Rightarrow \mu_l &= \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} g^{i,j}}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}}}. \tag{3.20}
\end{aligned}$$

The numerator is vector-valued while the denominator is a scalar, and the division of a vector by a scalar is meant component-wise here. This formula can be interpreted

geometrically: since $u_l^{i,j} = 0$ for all pixels not in segment l in the current outer iteration, we sum in the numerator over all spectra that belong to the segment but weight this sum by the reciprocals of the distance of the spectrum to the mean, measured in the anisotropic distance with respect to the covariance of the segment. Hence, spectra that are far away from μ_l get less weight, while close spectra are weighted stronger. The denominator sums over all these weights and takes care that the weights in the numerator sum up to 1. If the anisotropic distance was equal to 1 for every pixel, the formula would compute the arithmetic mean of the spectra in segment l .

Fixed point equation of Σ_l

Again with the help of the linearity of the derivative, the chain rule, the symmetry of Σ_l^{-1} and additionally Lemma A.1 (c) and (d), we find

$$\begin{aligned}
& \partial_{\Sigma_l} E_\epsilon(u, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\Sigma_l} \left[\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta} + \log \det \Sigma_l \right] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\Sigma_l} \sqrt{(g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) + \eta} \\
&\quad + \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\Sigma_l} \log \det \Sigma_l \\
&\stackrel{\text{Lemma A.1 (c), (d)}}{=} - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{1}{2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} (\Sigma_l^{-1})^T (g^{i,j} - \mu_l)(g^{i,j} - \mu_l)^T (\Sigma_l^{-1})^T \\
&\quad + \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} (\Sigma_l^{-1})^T \\
&= - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{1}{2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} \Sigma_l^{-1} (g^{i,j} - \mu_l)(g^{i,j} - \mu_l)^T \Sigma_l^{-1} \\
&\quad + \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \Sigma_l^{-1}.
\end{aligned}$$

Setting this to 0 and multiplying from left and right with Σ_l , we obtain

$$\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \Sigma_l = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{1}{2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} (g^{i,j} - \mu_l)(g^{i,j} - \mu_l)^T$$

$$\Rightarrow \Sigma_l = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{1}{2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l\|_{\Sigma_l^{-1}, \eta}} (g^{i,j} - \mu_l)(g^{i,j} - \mu_l)^T}{n_l}. \quad (3.21)$$

Again, the division of a matrix by a scalar is meant element-wise. The equation has a similar structure as the one derived for μ_l . We see also here that the factor $u_l^{i,j}$ sorts out the spectra that are assumed to belong to other segments than l in the current outer iteration and that the outer product in the numerator is then weighted by the reciprocal of the anisotropic distance of the spectrum to the mean. In particular, one can observe that the idea of making the indicator function robust against outliers by not squaring it (cf. Section 2.2.1) also shows its effect in the derived formula for the mean and covariance because we give less weight to spectra that deviate stronger from the mean.

Fixed point iteration update rules

We see that in the derived necessary conditions for an optimal μ_l and Σ_l in Eqs. (3.20) and (3.21), respectively, the variables appear on both sides of the respective equation. However, the conditions have a fixed point form and we can, therefore, directly turn these equations into a fixed point iteration scheme. Let $\mathbb{P}_{\mathcal{P}_\epsilon}$ describe the projection of a symmetric, positive semidefinite matrix onto $\mathcal{P}_\epsilon$ by applying the ϵ -regularization described in Section 2.2.1. We define the update rules for $m \in \mathbb{N}_0$ as

$$\begin{aligned} \mu_l^{(m+1)} &= \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l^{(m)}\|_{(\Sigma_{\epsilon,l}^{(m)})^{-1}, \eta}} g^{i,j}}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\|g^{i,j} - \mu_l^{(m)}\|_{(\Sigma_{\epsilon,l}^{(m)})^{-1}, \eta}}}, \\ \Sigma_{\epsilon,l}^{(m+1)} &= \mathbb{P}_{\mathcal{P}_\epsilon} \left(\frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2 \|g^{i,j} - \mu_l^{(m)}\|_{(\Sigma_{\epsilon,l}^{(m)})^{-1}, \eta}} (g^{i,j} - \mu_l^{(m+1)}) (g^{i,j} - \mu_l^{(m+1)})^T}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}} \right). \end{aligned} \quad (3.22)$$

The reason that the formula for $\Sigma_{\epsilon,l}^{(m+1)}$ is partly based on the old and partly on the updated value of μ_l is the attempt to reduce the necessary floating point operations and memory usage in the computation of the new iterates. The update rule is defined in such a way that already computed intermediate values can be reused when computing the updates. We see that the argument of $\mathbb{P}_{\mathcal{P}_\epsilon}$ is symmetric. Moreover, it is positive

semidefinite because for $v \in \mathbb{R}^L$ the matrix $vv^T \in \mathbb{R}^{L \times L}$ is positive semidefinite since for any $w \in \mathbb{R}^L$ we obtain

$$w^T(vv^T)w = (w^T v)v^T w = (w^T v)^2 \geq 0.$$

A multiplication of a positive semidefinite matrix with a non-negative factor as well as the summation of such matrices preserves the positive semidefiniteness. Hence, $\mathbb{P}_{\mathcal{P}_\epsilon}$ can be applied and we obtain $\Sigma_{\epsilon,l}^{(m+1)} \in \mathcal{P}_\epsilon$, i.e., the new iterate is admissible.

In the first outer iteration, i.e., in the first iteration of the alternating optimization procedure (cf. Section 3.1.3), we propose to initialize mean and covariance iterates $\mu_l^{(0)}$ and $\Sigma_{\epsilon,l}^{(0)}$ as the empirical mean and empirical covariance of the pixels in segment l , given by the initial segmentation from the k -means algorithm. That is, in the first outer iteration, we set

$$\begin{aligned} \mu_l^{(0)} &:= \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} g^{i,j}}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}}, \\ \Sigma_{\epsilon,l}^{(0)} &:= \mathbb{P}_{\mathcal{P}_\epsilon} \left(\frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} (g^{i,j} - \mu_l^{(0)}) (g^{i,j} - \mu_l^{(0)})^T}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} - 1} \right). \end{aligned}$$

The projection $\mathbb{P}_{\mathcal{P}_\epsilon}$ ensures the admissibility of the initial guess. These estimates are improved in the first application of the fixed point iteration by the update rules above. For the following outer iterations, we always reuse the outputs of the last outer iteration as initialization for the current ones.

The fixed point iteration scheme is stopped when a maximum number of iterations $m_{\mu,\Sigma}^{\max} \in \mathbb{N}$ is reached or when the change in the last and current iterates falls below a certain threshold. We measure the difference between the last and current estimates via

$$\left\| \mu_l^{(m+1)} - \mu_l^{(m)} \right\|_2 + \left\| (D_{\epsilon,l}^2)^{(m+1)} - (D_{\epsilon,l}^2)^{(m)} \right\|_F + \left\| V_l^{(m+1)} - V_l^{(m)} \right\|_F,$$

where $(D_{\epsilon,l}^2)^{(m)}$ and $V_l^{(m)}$ are obtained from the eigendecomposition of $\Sigma_{\epsilon,l}^{(m)}$, i.e.,

$$\Sigma_{\epsilon,l}^{(m)} = V_l^{(m)} (D_{\epsilon,l}^2)^{(m)} (V_l^{(m)})^T.$$

The norm $\|\cdot\|_F$ describes the Frobenius norm.

3.3.2 Parameters of ϕ AMS

Building on the notation of Section 2.3, the discretized objective functional of ϕ AMS is obtained by inserting $f_{\phi,l}$, stated in (2.48), into (3.9), resulting in

$$\begin{aligned}
& E_{\phi}(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\
&= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_{\phi,l}^{i,j}(\mu_l, V_l, \sigma_l^2) + \lambda |u_l|_{\text{BV}(\mathcal{N})_h} \\
&= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left[\sqrt{\sum_{n=1}^L \left[[V_l^T (g^{i,j} - \mu_l)]_n^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \right] \frac{1}{\sigma_{l,n}^2}} \right. \\
&\quad \left. + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}. \tag{3.23}
\end{aligned}$$

Since we optimize alternately between u and the parameters of the indicator function, we have to optimize over $\mu_1, \dots, \mu_k$, $V_1, \dots, V_k$ and $\sigma_1^2, \dots, \sigma_k^2$ while u is fixed. Special care has to be taken in the case of V_l and σ_l^2 because we have to ensure that the iterates always fulfill the constraints that $V_l \in O(L)$ and $\sigma_l^2 \in \mathbb{R}_{>0}^L$ for every $l \in \{1, \dots, k\}$. The way we go is to derive fixed point iteration schemes for the optimization of μ_l and σ_l^2 and we will see that the constraint for σ_l^2 is automatically fulfilled by the update rule. Unfortunately, a similar update rule for V_l that directly ensures the orthogonality is particularly difficult to be derived. Hence, we go a different way in this case and apply an update rule proposed by [107] that is an adaption of the gradient descent and able to preserve the orthogonality constraint. Since we base our update rules on the gradient, we have to make the requirements on ϕ slightly stronger than in Section 2.3 and assume ϕ to be continuously differentiable, which in turn implies the continuity of ϕ .

Let $l \in \{1, \dots, k\}$ be arbitrary but fixed. First, we observe that also for this indicator function, μ_l , V_l and σ_l^2 only appear in the l -th summand of the outer sum and, since u is fixed, the TV term is only a constant shift. Hence, the other summands and the TV term can be directly dropped in the computation of the derivative. We first derive the fixed point iteration schemes for μ_l and σ_l^2 , before we explain the derivation of the update rule for V_l . Define for every $(i, j) \in \mathcal{N}$ the function

$$h^{i,j}(\mu_l, V_l, \sigma_l^2) := \sqrt{\sum_{n=1}^L \left[[V_l^T (g^{i,j} - \mu_l)]_n^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \right] \frac{1}{\sigma_{l,n}^2}}, \tag{3.24}$$

which is the square root term in $f_{\phi,l}^{i,j}$.

Fixed point equation of μ_l

To compute the derivative of E_ϕ with respect to μ_l , we define $\Sigma_l := V_l D_l^2 V_l^T$ with $D_l^2 \in \mathbb{R}^{L \times L}$ being a diagonal matrix with σ_l^2 as the diagonal and reformulate $h^{i,j}$ and E_ϕ as in (2.50) to

$$h^{i,j}(\mu_l, V_l, \sigma_l^2) = \left[(g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right]^{\frac{1}{2}}$$

and

$$\begin{aligned} & E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\ &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left[\left((g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) \right. \right. \\ & \quad \left. \left. + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}. \end{aligned}$$

Let $v_{l,n} \in \mathbb{R}^L$ denote for $n \in \{1, \dots, L\}$ the n -th column of V_l . We derive from this formulation of E_ϕ together with the linearity of the derivative, the chain rule, the symmetry of Σ_l^{-1} and Lemma A.1 (a), (b)

$$\begin{aligned} & \partial_{\mu_l} E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\mu_l} \left[\left((g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) \right. \right. \\ & \quad \left. \left. + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\mu_l} \left((g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) \right. \\ & \quad \left. + \alpha \sum_{n=1}^L \phi \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(\partial_{\mu_l} (g^{i,j} - \mu_l)^T \Sigma_l^{-1} (g^{i,j} - \mu_l) \right. \\ & \quad \left. + \alpha \sum_{n=1}^L \partial_{\mu_l} \phi \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) \frac{1}{\sigma_{l,n}^2} \right) \end{aligned}$$

$$\begin{aligned}
\text{Lemma A.1 (b)} &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(\Sigma_l^{-1} (\mu_l - g^{i,j}) + (\Sigma_l^{-1})^T (\mu_l - g^{i,j}) \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \partial_{\mu_l} \phi \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) \frac{1}{\sigma_{l,n}^2} \right) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(2\Sigma_l^{-1} (\mu_l - g^{i,j}) \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \partial_{\mu_l} \phi \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) \frac{1}{\sigma_{l,n}^2} \right) \\
\text{Lemma A.1 (a)} &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(2\Sigma_l^{-1} (\mu_l - g^{i,j}) \right. \\
&\quad \left. - \alpha \sum_{n=1}^L \frac{1}{\sigma_{l,n}^2} \phi' \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) 2 [v_{l,n}^T (g^{i,j} - \mu_l)] v_{l,n} \right) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)} \Sigma_l^{-1} \mu_l - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(\Sigma_l^{-1} g^{i,j} \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \frac{1}{\sigma_{l,n}^2} \phi' \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) [v_{l,n}^T (g^{i,j} - \mu_l)] v_{l,n} \right).
\end{aligned}$$

Similar to the derivation of the fixed point formulations for ϵ AMS, we set the derivative to 0 to isolate a μ_l and obtain a fixed point formulation of the necessary optimality condition. A multiplication by Σ_l from the left and the fact that $\Sigma_l v_{l,n} = \sigma_{l,n}^2 v_{l,n}$ because $v_{l,n}$ is an eigenvector of Σ_l to the eigenvalue $\sigma_{l,n}^2$ lead to

$$\begin{aligned}
\mu_l &= \frac{1}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(g^{i,j} \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \frac{1}{\sigma_{l,n}^2} \phi' \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) [v_{l,n}^T (g^{i,j} - \mu_l)] \Sigma_l v_{l,n} \right) \\
&= \frac{1}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(g^{i,j} \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \phi' \left([v_{l,n}^T (g^{i,j} - \mu_l)]^2 \right) [v_{l,n}^T (g^{i,j} - \mu_l)] v_{l,n} \right). \tag{3.25}
\end{aligned}$$

Fixed point equation of σ_l^2

We compute the derivate of E_ϕ with respect to σ_l^2 element-wise. As before, we use the linearity of the derivative and the chain rule and obtain for the r -th entry $\sigma_{l,r}^2$, $r \in \{1, \dots, L\}$, of σ_l^2

$$\begin{aligned}
& \partial_{\sigma_{l,r}^2} E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\sigma_{l,r}^2} \left[\sqrt{\sum_{n=1}^L \left[[V_l^T (g^{i,j} - \mu_l)]_n^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \right]} \frac{1}{\sigma_{l,n}^2} \right. \\
&\quad \left. + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\sigma_{l,r}^2} \sqrt{\sum_{n=1}^L \left[[V_l^T (g^{i,j} - \mu_l)]_n^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \right]} \frac{1}{\sigma_{l,n}^2} \\
&\quad + \beta \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{\sigma_{l,r}^2} \sum_{n=1}^L \log \sigma_{l,n}^2 \\
&= - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left[[V_l^T (g^{i,j} - \mu_l)]_r^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_r^2 \right) \right] \frac{1}{(\sigma_{l,r}^2)^2} \\
&\quad + \beta \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{\sigma_{l,r}^2}.
\end{aligned}$$

Since the double sums in the expression are independent of r , we obtain an equation in the form of

$$\frac{a}{x^2} + \frac{b}{x} = 0 \quad (3.26)$$

for $a, b \in \mathbb{R}$ with x being $\sigma_{l,r}^2$ when setting the derivate equal to 0. Please note that this ignores that $\sigma_{l,r}^2$ also appears in $h^{i,j}(\mu_l, V_l, \sigma_l^2)$ but, as in the case of ϵ AMS, we only solve for some $\sigma_{l,r}^2$ in the equation. We have that $b > 0$ because we can assume that there is an $(i, j) \in \mathcal{N}$ such that $u_l^{i,j} = 1$, since otherwise the segment was empty, and $\beta > 0$ by assumption. Furthermore, $a > 0$ because the term in the bigger rectangular brackets is constructed in such a way that it is always positive (which we discussed in Section 2.3.1), implied by $\alpha > 0$ by assumption as well as $\phi(t) > 0$ for sufficiently small $t \geq 0$ and $\phi(t) \geq 0$ for any $t \geq 0$. In particular, this term in the brackets also appears under the square root in $h^{i,j}$, which then implies

that $h^{i,j}(\mu_l, V_l, \sigma_l^2) > 0$. Eq. (3.26) is satisfied for $a, b \neq 0$ if $x = -\frac{a}{b}$. Hence, we obtain as the necessary condition, formulated in fixed point form,

$$\sigma_{l,r}^2 = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left[[V_l^T (g^{i,j} - \mu_l)]_r^2 + \alpha \phi \left([V_l^T (g^{i,j} - \mu_l)]_r \right)^2 \right]}{\beta \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}}. \quad (3.27)$$

The argumentation we used to explain why $a, b > 0$ also implies that the expression on the right-hand side is always positive.

Update rule for V_l

As already said, we will derive a gradient-based update rule that preserves the orthogonality constraint. To this end, we use the approach of Wen and Yin [107]. The idea is to define a curve on the manifold of orthogonal matrices that starts at some point and goes in the direction of the projection of the negative gradient of the objective function into the tangent space at the starting point. Hence, the value of the objective function is reduced for a sufficiently small step along the curve. We will now explain the generic update scheme for optimization under an orthogonality constraint introduced in [107] and apply it to our scenario afterward.

Let $n, p \in \mathbb{N}$ with $n \geq p$ and $\mathcal{F}: \mathbb{R}^{n \times p} \rightarrow \mathbb{R}$ be differentiable. We aim to solve the optimization problem

$$\begin{aligned} & \underset{X \in \mathbb{R}^{n \times p}}{\text{minimize}} && \mathcal{F}(X) \\ & \text{subject to} && X^T X = I. \end{aligned}$$

The admissible set is given as

$$\mathcal{M}_n^p := \{X \in \mathbb{R}^{n \times p} \mid X^T X = I\}, \quad (3.28)$$

which is the so-called *Stiefel manifold*. Let $X \in \mathcal{M}_n^p$. Optimization methods are often based on the gradient, like, e.g, conjugate gradients or gradient descent, since the negative gradient points in the direction of the steepest descent. A manifold does not necessarily have a vector space structure and the gradient is therefore not defined on it. However, one usually uses the gradient in the tangent space of the manifold. We denote the tangent space of $\mathcal{M}_n^p$ at X with $\mathcal{T}_X \mathcal{M}_n^p$. The gradient depends on the chosen inner product on $\mathcal{T}_X \mathcal{M}_n^p$ because this defines lengths and angles and, therefore, the notion of the steepest ascent and descent. In this case, the used inner product on $\mathcal{T}_X \mathcal{M}_n^p$ is given as

$$\langle Z_1, Z_2 \rangle_{\mathcal{T}_X \mathcal{M}_n^p} := \text{tr} \left(Z_1^T \left(I - \frac{1}{2} X X^T \right) Z_2 \right),$$

for $Z_1, Z_2 \in \mathcal{T}_X \mathcal{M}_n^p$. We define $G_X := \nabla \mathcal{F}(X)$, which is the gradient of $\mathcal{F}$ at X . Then, we obtain the tangential gradient $\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X) \in \mathcal{T}_X \mathcal{M}_n^p$ in the tangent space at X with the formula [35]

$$\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X) = G_X - X G_X^T X.$$

If we made a step in the direction of $\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X)$ in the classical gradient descent, we generally reach a point that does not lie on the manifold anymore and, therefore, does not fulfill the orthogonality constraint. Hence, we have to get back to the manifold, for instance, by projecting the new point onto $\mathcal{M}_n^p$, before we can take a next gradient descent step. This problem of leaving the manifold, however, can be avoided. To this end, the matrix

$$A_X := G_X X^T - X G_X^T$$

is defined. It is based on the gradient G_X and the considered point X on the manifold. This matrix is skew-symmetric (cf. Definition A.4) because

$$A_X^T = (G_X X^T - X G_X^T)^T = X G_X^T - G_X X^T = -A_X,$$

and, in particular, related to the tangential gradient $\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X)$:

Lemma 3.21 ([107, Lemma 1]).

It holds

$$\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X) = A_X X.$$

In particular, $\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X) = 0$ is equivalent to $A_X = 0$.

This means that $X \in \mathcal{M}_n^p$ fulfills the first order optimality condition if and only if $A_X = 0$. Starting from X on the Stiefel manifold, Wen and Yin [107] propose to reach a new point $Y(\rho)$ on $\mathcal{M}_n^p$ by solving the following equation

$$Y(\rho) = X - \frac{\rho}{2} A_X (X + Y(\rho)) \tag{3.29}$$

with a step size $\rho \geq 0$. As the next lemma will show, this will give us a point on the manifold for any non-negative step size (cf. Lemma 3.22 (b)). We can describe the update also in closed form. It holds

$$\begin{aligned} Y(\rho) &= X - \frac{\rho}{2} A_X X - \frac{\rho}{2} A_X Y(\rho) \\ &= \left(I - \frac{\rho}{2} A_X \right) X - \frac{\rho}{2} A_X Y(\rho). \end{aligned}$$

This is equivalent to

$$\left(I + \frac{\rho}{2}A_X\right) Y(\rho) = \left(I - \frac{\rho}{2}A_X\right) X,$$

and therefore also to

$$Y(\rho) = \left(I + \frac{\rho}{2}A_X\right)^{-1} \left(I - \frac{\rho}{2}A_X\right) X, \quad (3.30)$$

provided that $I + \frac{\rho}{2}A_X$ is invertible. This, however, is indeed always fulfilled by the next lemma (cf. Lemma 3.22 (a)). We set

$$Q_X(\rho) := \left(I + \frac{\rho}{2}A_X\right)^{-1} \left(I - \frac{\rho}{2}A_X\right),$$

which allows writing (3.30) as

$$Y(\rho) = Q_X(\rho)X.$$

This is known as the *Cayley transform* and lets us deduce some properties of $Y(\rho)$: because the inverse matrix is continuous, $Y(\rho)$ is continuous in ρ . Moreover, $Y(0) = X$. More properties are given in the following lemma:

Lemma 3.22 ([107, Lemma 3]).

- (a) For any $\rho \geq 0$, the matrix $I + \frac{\rho}{2}A_X$ is invertible and therefore $Q_X(\rho)$ is well-defined. Moreover, $Q_X(\rho)$ is an orthogonal matrix.
- (b) The matrix $Y(\rho)$, given by (3.29), fulfills $Y(\rho)^T Y(\rho) = X^T X$ for every $\rho \geq 0$. As the derivative of $Y(\rho)$ we have

$$Y'(\rho) = -\left(I + \frac{\rho}{2}A_X\right)^{-1} A_X \left(\frac{X + Y(\rho)}{2}\right),$$

and $Y'(0) = -A_X X$.

- (c) For the derivative of $\mathcal{F}(Y(\rho))$ with respect to ρ , denoted by $\mathcal{F}'_\rho(Y(\rho))$, at $\rho = 0$, it holds

$$\mathcal{F}'_\rho(Y(0)) = \left. \frac{\partial \mathcal{F}(Y(\rho))}{\partial \rho} \right|_{\rho=0} = -\frac{1}{2} \|A_X\|_F^2$$

and for a sufficiently small step size ρ the curve $Y(\rho)$ decreases $\mathcal{F}$.

Hence, the first part of Lemma 3.22 (b) implies that $Y(\rho) \in \mathcal{M}_n^p$ for every $\rho \geq 0$ because $X \in \mathcal{M}_n^p$ and therefore, $Y(\rho)$ describes a path on the Stiefel manifold. The second part and Lemma 3.21 yield

$$Y'(0) = -A_X X = -\nabla_{\mathcal{T}_X \mathcal{M}_n^p} \mathcal{F}(X),$$

which means that the curve described by $Y(\rho)$ at the starting point goes in the direction of the negative tangential gradient at X . From Lemma 3.22 (c) we learn that the slope of the function $\rho \mapsto \mathcal{F}_\rho(Y(\rho))$ is negative at $\rho = 0$ if $A_X \neq 0$, which implies that for a sufficiently small $\rho \geq 0$ the curve decreases the value of $\mathcal{F}$.

The considerations above result for an initial guess $X^{(0)} \in \mathcal{M}_n^p$ in the update rule

$$X^{(m+1)} := \left(I + \frac{\rho_m}{2} A_{X^{(m)}} \right)^{-1} \left(I - \frac{\rho_m}{2} A_{X^{(m)}} \right) X^{(m)} \quad (3.31)$$

for $m \in \mathbb{N}_0$ and $(\rho_m)_{m \in \mathbb{N}_0}$ being a sequence of step sizes with $\rho_m \geq 0$. The most expensive part in the computation of $X^{(m+1)}$ is the inversion of $I + \frac{\rho_m}{2} A_{X^{(m)}}$. Wen and Yin [107] describe methods how to reduce the cost and compute the inverse more efficiently.

A drawback of this method compared to the fixed point iteration schemes that we derived before for the optimization with respect to the other parameters is that we have to choose a step size for each update of the parameters. This is a crucial step since we know from methods like the classical gradient descent that convergence is not guaranteed for a fixed step size. Wen and Yin [107] propose in their article to derive the *Armijo-Wolfe conditions* for their update formula and then use algorithms 3.5 and 3.6 from [81], which yield a step size that satisfies these conditions. In particular, in this case it can be proven that the norm of the tangential gradient $\nabla_{\mathcal{T}_{X^{(m)}} \mathcal{M}_n^p} \mathcal{F}(X^{(m)})$ converges to 0. However, on the other hand, choosing an appropriate step size can also add a significant computational burden. This thesis is meant to give a proof of concept for the proposed methods and not to implement the methods in a productive environment. We, therefore, use the so-called *backtracking line search* algorithm [81, Algorithm 3.1] as a heuristic to find a step size for the next update in (3.31). It is computationally cheaper and easier to implement than the algorithms to meet the Armijo-Wolfe conditions (the algorithms 3.5 and 3.6 in [81]) and although there is no theoretical convergence guarantee, it ensures that the objective functional never increases and our numerical experiments show a converging behavior, as we will see in Chapter 4. The reference for the considerations about the choice of the step size is [81].

The optimal choice for a step size to make maximum progress with the optimization procedure is to select ρ such that the function

$$\varphi(\rho) := \mathcal{F} \left(\left(I + \frac{\rho}{2} A_{X^{(m)}} \right)^{-1} \left(I - \frac{\rho}{2} A_{X^{(m)}} \right) X^{(m)} \right)$$

is minimized because then starting from the m -th iterate $X^{(m)}$ on $\mathcal{M}_n^p$ the minimizing ρ would give the maximum decrease in $\mathcal{F}$. This would require us to solve a minimization problem for each update, which can be extremely expensive. The backtracking line search is a compromise that only aims to find a step size of sufficient decrease. The notion of sufficient decrease is measured by the so-called *Armijo condition* that can be stated in our setting as

$$\mathcal{F}(X^{(m+1)}) \leq \mathcal{F}(X^{(m)}) + c\rho_m \mathcal{F}'_\rho(X^{(m)}) \quad (3.32)$$

for $c \in (0, 1)$. Since by Lemma 3.22 (c) we have $\mathcal{F}'_\rho(X^{(m)}) < 0$ as long as $A_{X^{(m)}} \neq 0$ and $c, \rho_m > 0$ (because for $\rho_m = 0$ we do not move away from $X^{(m)}$ and no decrease can happen), the condition expects a certain reduction of the value of $\mathcal{F}$. To find a step size that fulfills the Armijo condition in (3.32), the backtracking line search starts with an initial step size $\bar{\rho} > 0$ and successively reduces it by a multiplication with a factor $\gamma \in (0, 1)$ until the condition is satisfied. The algorithm is summarized in Algorithm 3.2.

Algorithm 3.2: Backtracking line search [81, Algorithm 3.1]

Input: Initial step size $\bar{\rho} > 0$, $c \in (0, 1)$ and $\gamma \in (0, 1)$;

Result: Step size $\rho_m = \rho$ fulfilling (3.32) for the m -th update;

$\rho \leftarrow \bar{\rho}$;

while $\mathcal{F}(X^{(m+1)}) > \mathcal{F}(X^{(m)}) + c\rho \mathcal{F}'_\rho(X^{(m)})$ **do**

$\rho \leftarrow \gamma\rho$;

end

To sum up, we will use the rule given in (3.31) together with a step size determined by Algorithm 3.2 and apply them iteratively to optimize E_ϕ with respect to V_l . It requires the computation of A_X , which depends on the gradient of the objective function with respect to the parameters that we want to optimize subject to the orthogonality constraint. We therefore need to compute $\partial_{V_l} E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2)$. We will now derive an expression for it. Again, we reformulate $h^{i,j}$ (cf. (3.24)) and E_ϕ (cf. (3.23)) since this simplifies the computations. We define $D_l^2 \in \mathbb{R}^{L \times L}$ again as a diagonal matrix with σ_l^2 on its diagonal. Then, we obtain

$$\begin{aligned} & h^{i,j}(\mu_l, V_l, \sigma_l^2) \\ &= \left[(g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g^{i,j} - \mu_l) + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right]^{\frac{1}{2}} \end{aligned}$$

and

$$\begin{aligned}
& E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\
&= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left[\left((g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g^{i,j} - \mu_l) \right. \right. \\
&\quad \left. \left. + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}
\end{aligned}$$

and observe as before that we only have to consider the l -th summand of the outer sum and can also drop the TV term in the computation of the derivative. It follows with the linearity of the derivate, the chain rule and Lemma A.1 (f) and (g)

$$\begin{aligned}
& \partial_{V_l} E_\phi(u, \mu_1, \dots, \mu_k, V_1, \dots, V_k, \sigma_1^2, \dots, \sigma_k^2) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{V_l} \left[\left((g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g^{i,j} - \mu_l) \right. \right. \\
&\quad \left. \left. + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} + \beta \sum_{n=1}^L \log \sigma_{l,n}^2 \right] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \partial_{V_l} \left((g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g^{i,j} - \mu_l) \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \frac{1}{\sigma_{l,n}^2} \right)^{\frac{1}{2}} \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(\partial_{V_l} \left[(g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} V_l^T (g^{i,j} - \mu_l) \right] \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \frac{1}{\sigma_{l,n}^2} \partial_{V_l} \phi \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) \right) \\
&\stackrel{\text{Lemma A.1 (f), (g)}}{=} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left(2 (g^{i,j} - \mu_l) (g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \frac{1}{\sigma_{l,n}^2} \phi' \left([V_l^T (g^{i,j} - \mu_l)]_n^2 \right) 2 [V_l^T (g^{i,j} - \mu_l)]_n (g^{i,j} - \mu_l) e_n^T \right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l, V_l, \sigma_l^2)} \left((g^{i,j} - \mu_l) (g^{i,j} - \mu_l)^T V_l (D_l^2)^{-1} \right. \\
&\quad \left. + \alpha \sum_{n=1}^L \frac{\phi' \left(\left[V_l^T (g^{i,j} - \mu_l) \right]_n^2 \right)}{\sigma_{l,n}^2} \left[V_l^T (g^{i,j} - \mu_l) \right]_n (g^{i,j} - \mu_l) e_n^T \right).
\end{aligned}$$

When applying Algorithm 3.2 for the step size selection, we need to be able to check whether we make a sufficient decrease with the next step. In our specific case, the function $\mathcal{F}$ in Algorithm 3.2 equals our objective function E_ϕ where all parameters except for V_l are fixed. The function $\mathcal{F}'_\rho$ is given by Lemma 3.22 (c) because $\mathcal{F}'_\rho(X^{(m)})$ means that we evaluate $\mathcal{F}'_\rho$ at the current iterate and have not moved away from it so far. In other words, $X^{(m)}$ is $Y(0)$ in Lemma 3.22 (c).

Joint alternating update scheme

Now, we have all the ingredients to state the update rules for μ_l , V_l and σ_l^2 in our setting. It is essentially also an alternating optimization approach since we use the current iterates to update μ_l , σ_l^2 and V_l separately in an alternating manner. For ease of notation, we define the symbols

$$\mathbf{V} := (V_1, \dots, V_k) \in O(L)^k, \quad \boldsymbol{\sigma}^2 := (\sigma_1^2, \dots, \sigma_k^2) \in (\mathbb{R}_{>0}^L)^k.$$

For $m \in \mathbb{N}_0$, we compute the $(m+1)$ -th iterates motivated by the formulas in Eqs. (3.25), (3.27) and (3.31) via

$$\begin{aligned}
\mu_l^{(m+1)} := & \frac{1}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l^{(m)}, V_l^{(m)}, (\sigma_l^2)^{(m)})}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j}(\mu_l^{(m)}, V_l^{(m)}, (\sigma_l^2)^{(m)})} \left(g^{i,j} \right. \\
& \left. + \alpha \sum_{n=1}^L \phi' \left(\left[\left(V_l^{(m)} \right)^T (g^{i,j} - \mu_l^{(m)}) \right]_n^2 \right) \left[\left(V_l^{(m)} \right)^T (g^{i,j} - \mu_l^{(m)}) \right]_n V_l^{(m)} e_n \right),
\end{aligned} \tag{3.33}$$

$$\begin{aligned}
(\sigma_{l,n}^2)^{(m+1)} := & \\
& \frac{1}{\beta \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{2h^{i,j} \left(\mu_l^{(m)}, V_l^{(m)}, (\sigma_l^2)^{(m)} \right)} \left[\left[\left(V_l^{(m)} \right)^T \left(g^{i,j} - \mu_l^{(m)} \right) \right]_n^2 \right. \\
& \left. + \alpha \phi \left(\left[\left(V_l^{(m)} \right)^T \left(g^{i,j} - \mu_l^{(m)} \right) \right]_n^2 \right) \right],
\end{aligned} \tag{3.34}$$

$$\begin{aligned}
V_l^{(m+1)} &:= \left(I + \frac{\rho_m}{2} A_{V_l^{(m)}} \right)^{-1} \left(I - \frac{\rho_m}{2} A_{V_l^{(m)}} \right) V_l^{(m)}, \\
A_{V_l^{(m)}} &= \partial_{V_l} E_\phi \left(u, \boldsymbol{\mu}^{(m)}, \mathbf{V}^{(m)}, (\boldsymbol{\sigma}^2)^{(m+1)} \right) \left(V_l^{(m)} \right)^T \\
&\quad - V_l^{(m)} \left[\partial_{V_l} E_\phi \left(u, \boldsymbol{\mu}^{(m)}, \mathbf{V}^{(m)}, (\boldsymbol{\sigma}^2)^{(m+1)} \right) \right]^T, \\
\partial_{V_l} E_\phi \left(u, \boldsymbol{\mu}^{(m)}, \mathbf{V}^{(m)}, (\boldsymbol{\sigma}^2)^{(m+1)} \right) &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{u_l^{i,j}}{h^{i,j} \left(\mu_l^{(m)}, V_l^{(m)}, (\sigma_l^2)^{(m)} \right)} \\
&\quad \left(\begin{aligned} & \left(g^{i,j} - \mu_l^{(m)} \right) \left(g^{i,j} - \mu_l^{(m)} \right)^T V_l^{(m)} \left[(D_l^2)^{(m+1)} \right]^{-1} \\ & + \alpha \sum_{n=1}^L \frac{\phi' \left(\left[\left(V_l^{(m)} \right)^T \left(g^{i,j} - \mu_l^{(m)} \right) \right]_n^2 \right) \left[\left(V_l^{(m)} \right)^T \left(g^{i,j} - \mu_l^{(m)} \right) \right]_n}{(\sigma_{l,n}^2)^{(m+1)}} \\ & \left(g^{i,j} - \mu_l^{(m)} \right) e_n^T \end{aligned} \right).
\end{aligned} \tag{3.35}$$

Please note that the expression of $\partial_{V_l} E_\phi \left(u, \boldsymbol{\mu}^{(m)}, \mathbf{V}^{(m)}, (\boldsymbol{\sigma}^2)^{(m+1)} \right)$ is rigorously speaking not correct since we evaluate $\partial_{V_l} E_\phi$ at $(\boldsymbol{\sigma}^2)^{(m+1)}$ but in the formula itself both the old value $(\sigma_l^2)^{(m)}$ and the updated value $(\sigma_l^2)^{(m+1)}$ are used. The reason is that the expression shows how we compute $\partial_{V_l} E_\phi$ in practice during optimization and it, therefore, has to be understood as an approximation to the derivative. The

step size ρ_m is determined by Algorithm 3.2. Please also note that, as in ϵ AMS, we do not always use the updated values of the variables in the update rules, even if updated values are available. As before, the reason for the formulas being essentially based on the old and not the updated values is the attempt to reduce the necessary floating point operations and memory usage in the computation of the new iterates while exploiting the updated and, therefore, refined values when possible. As in case of ϵ AMS, the update rules are defined in such a way that many intermediate values can be reused when computing the updates. The hope is to benefit from the updated and refined values whenever possible while saving memory and computation costs. As it was explained in Chapter 1, the dimensions of the variables when working with hyperspectral images can be very large, making matrix-vector and matrix-matrix products expensive.

We initialize the update scheme similar to the one of ϵ AMS: the starting point is the initial segmentation by k -means, which gives an initial guess for u . Based on that, we set in the first outer iteration $\mu_l^{(0)}$ as the empirical mean of the pixels in the l -th segment, $(\sigma_l^2)^{(0)}$ to be the eigenvalues and $V_l^{(0)}$ as the eigenvectors of the empirical covariance matrix of the l -th segment. This is expressed in formulas as follows:

$$\begin{aligned}\mu_l^{(0)} &:= \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} g^{i,j}}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}}, \\ \Sigma_l^{(0)} &:= \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left(g^{i,j} - \mu_l^{(0)} \right) \left(g^{i,j} - \mu_l^{(0)} \right)^T}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} - 1},\end{aligned}\tag{3.36}$$

which are the same formulas as in the case of ϵ AMS but without the projection onto $\mathcal{P}_\epsilon$. Since $\Sigma_l^{(0)}$ is symmetric, the spectral theorem for symmetric real matrices allows us to compute an eigendecomposition

$$\Sigma_l^{(0)} = V_l^{(0)} (D_l^2)^{(0)} \left(V_l^{(0)} \right)^T.$$

The orthogonal matrix $V_l^{(0)}$ is then the initial guess for the optimization of V_l and the entries on the diagonal of the diagonal matrix $(D_l^2)^{(0)}$ the initial guess for the optimization of σ_l^2 . If an entry equals 0 because the positive semidefiniteness of $\Sigma_l^{(0)}$ only guarantees the non-negativity and we do not have a parameter like ϵ in ϵ AMS that ensures positivity of the diagonal entries, we set that value in $(\sigma_l^2)^{(0)}$ to 1. In the subsequent outer iterations the update scheme is initialed with the result of the last outer iteration.

We stop the update scheme once a maximum number of iterations $m_{\mu, V, \sigma^2}^{\max} \in \mathbb{N}$ is reached or the expression

$$\left\| \mu_l^{(m+1)} - \mu_l^{(m)} \right\|_2 + \left\| (\sigma_l^2)^{(m+1)} - (\sigma_l^2)^{(m)} \right\|_2 + \left\| V_l^{(m+1)} - V_l^{(m)} \right\|_F, \tag{3.37}$$

is smaller than a given threshold $t \geq 0$, which means that the change in the iterates is small. The full update procedure that is run in each outer iteration and for each segment $l \in \{1, \dots, k\}$ is summarized in Algorithm 3.3.

Algorithm 3.3: Update scheme for parameters of $f_{\phi,l}$

Input: Threshold $t \geq 0$, maximum number of iterations $m_{\mu,V,\sigma^2}^{\max} \in \mathbb{N}$, current labeling function u , initial guesses $\mu_l^{(0)} \in \mathbb{R}^L$, $V_l^{(0)} \in O(L)$, $(\sigma_l^2)^{(0)} \in \mathbb{R}_{>0}^L$;

Result: Updated estimates $\mu_l^{(m)}$, $V_l^{(m)}$ and $(\sigma_l^2)^{(m)}$ for μ_l , V_l and σ_l^2 ;
 $m \leftarrow 0$;

while $m < m_{\mu,V,\sigma^2}^{\max}$ or $(3.37) \geq t$ **do**

 compute $\mu_l^{(m+1)}$ via (3.33);

 compute $(\sigma_l^2)^{(m+1)}$ via (3.34);

 compute $V_l^{(m+1)}$ via (3.35) and with a step size determined by Algorithm 3.2;

$m \leftarrow m + 1$;

end

3.3.3 Parameters of kMS

We use the notation from Section 2.4. We choose a symmetric, positive semidefinite kernel K and obtain an associated feature map Φ mapping to a feature space $\mathcal{H}$ such that $\mathcal{H} = \text{span}\{\Phi(g^{i,j}) \mid (i,j) \in \mathcal{N}\}$. Such a feature map and feature space exist by Proposition 2.33. Please note that for ease of notation, $g^{i,j}$ for $(i,j) \in \mathcal{N}$ represents all the information that we process about the pixel at (i,j) ; i.e., in the notation of Section 2.4.2, $g^{i,j}$ is the tuple of spectral and contextual information and contained in $X = X_\omega \times X_s$. The discretized kMS objective functional E_Φ , obtained from plugging the indicator function $f_{\Phi,l}$ in (2.54) into (3.9), is

$$\begin{aligned} E_\Phi(u, \mu_1^\Phi, \dots, \mu_k^\Phi) &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} f_{\Phi,l}^{i,j}(\mu_l^\Phi) + \lambda |u_l|_{\text{BV}(\mathcal{N})_h} \\ &= \sum_{l=1}^k \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \|\Phi(g^{i,j}) - \mu_l^\Phi\|_{\mathcal{H}}^2 + \lambda |u_l|_{\text{BV}(\mathcal{N})_h}. \end{aligned} \quad (3.38)$$

The parameters of $f_{\Phi,l}$, for $l \in \{1, \dots, k\}$, we have to optimize over are the means $\mu_1^\Phi, \dots, \mu_k^\Phi \in \mathcal{H}$ of the embedded spectra $\Phi(g^{i,j})$ in the feature space $\mathcal{H}$. As opposed

to ϵ AMS and ϕ AMS, this model allows for a derivation of a closed-form solution for the embedded mean spectra for a given u .

Let $l \in \{1, \dots, k\}$ be arbitrary but fixed. The mean μ_l^Φ only appears in the l -th summand of the outer sum and we can, therefore, ignore the other summands. Also independent of μ_l^Φ is the TV term, which can be ignored in the following computations, too. Since the induced squared norm on a Hilbert space is always strictly convex (cf. Lemma A.7), the strict convexity of the induced squared norm $\|\cdot\|_{\mathcal{H}}^2$ on $\mathcal{H}$ follows. This implies that for any $(i, j) \in \mathcal{N}$ also the function $\|\Phi(g^{i,j}) - \cdot\|_{\mathcal{H}}^2$ is strictly convex because for $v, w \in \mathcal{H}$ such that $v \neq w$ and $\alpha \in (0, 1)$, we find

$$\begin{aligned} & \|\Phi(g^{i,j}) - (\alpha v + (1 - \alpha)w)\|_{\mathcal{H}}^2 \\ &= \|\alpha\Phi(g^{i,j}) + (1 - \alpha)\Phi(g^{i,j}) - \alpha v - (1 - \alpha)w\|_{\mathcal{H}}^2 \\ &= \|\alpha(\Phi(g^{i,j}) - v) + (1 - \alpha)(\Phi(g^{i,j}) - w)\|_{\mathcal{H}}^2 \\ &< \alpha \|\Phi(g^{i,j}) - v\|_{\mathcal{H}}^2 + (1 - \alpha) \|\Phi(g^{i,j}) - w\|_{\mathcal{H}}^2. \end{aligned}$$

Moreover, a sum of strictly convex functions is still strictly convex and we, therefore, conclude that

$$\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \|\Phi(g^{i,j}) - \cdot\|_{\mathcal{H}}^2$$

is strictly convex because $u_l^{i,j} \in \{0, 1\}$ for any $(i, j) \in \mathcal{N}$ and we assume that $u_l^{i,j} = 1$ for some $(i, j) \in \mathcal{N}$ since otherwise, the segment is empty and the term not present in the functional. Consequently, if we can determine a mean embedded spectrum $\mu_l^\Phi \in \mathcal{H}$ such that E_Φ is minimized with respect to this parameter, this minimizer is unique. We can express μ_l^Φ as

$$\mu_l^\Phi = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} c_l^{i,j} \Phi(g^{i,j}) \quad (3.39)$$

with coefficients $c_l^{i,j} \in \mathbb{R}$ for $(i, j) \in \mathcal{N}$ because $\mathcal{H}$ equals the span of the embedded spectra. The set spanning $\mathcal{H}$ is not a basis and can contain redundant information. Therefore, there might be different choices of coefficients $c_l^{i,j}$, $(i, j) \in \mathcal{N}$, leading to the same mean μ_l^Φ . This implies that, even though a minimizing μ_l^Φ of E_Φ is unique by the strict convexity, there can be several choices of coefficients $c_l^{i,j}$, $(i, j) \in \mathcal{N}$, leading to that μ_l^Φ . Hence, if we plug the expression in (3.39) into E_Φ in (3.38) and interpret the result as a function of the coefficients $c_l^{i,j}$, $(i, j) \in \mathcal{N}$, we cannot expect the strict convexity of E_Φ with respect to μ_l^Φ to be propagated through (3.39) such that we would obtain strict convexity with respect to the coefficients $c_l^{i,j}$, $(i, j) \in \mathcal{N}$.

However, it can be easily shown that such a composition of E_Φ in (3.38) and μ_l^Φ in (3.39) is, in fact, convex in the coefficients $c_l^{i,j}$, $(i,j) \in \mathcal{N}$, which implies that a choice of coefficients that fulfills an optimality condition is a global minimum, even though possibly non-unique. Consequently, if we find a set of coefficients such that the necessary optimality condition is fulfilled, it leads to a unique minimizing μ_l^Φ . To this end, we consider a reformulation of the l -th summand of E_Φ in (3.38) (while omitting the TV term) and a replacement of μ_l^Φ with (3.39), which yields

$$\begin{aligned}
& \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \|\Phi(g^{i,j}) - \mu_l^\Phi\|_{\mathcal{H}}^2 \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} (\langle \Phi(g^{i,j}), \Phi(g^{i,j}) \rangle_{\mathcal{H}} - 2 \langle \Phi(g^{i,j}), \mu_l^\Phi \rangle_{\mathcal{H}} + \langle \mu_l^\Phi, \mu_l^\Phi \rangle_{\mathcal{H}}) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \left(\langle \Phi(g^{i,j}), \Phi(g^{i,j}) \rangle_{\mathcal{H}} - 2 \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} c_l^{m,n} \langle \Phi(g^{i,j}), \Phi(g^{m,n}) \rangle_{\mathcal{H}} \right. \\
&\quad \left. + \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} \sum_{r=1}^{N_1} \sum_{s=1}^{N_2} c_l^{m,n} c_l^{r,s} \langle \Phi(g^{m,n}), \Phi(g^{r,s}) \rangle_{\mathcal{H}} \right) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{i,j}) \rangle_{\mathcal{H}} - 2 \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} c_l^{m,n} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{m,n}) \rangle_{\mathcal{H}} \\
&\quad + \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} \sum_{r=1}^{N_1} \sum_{s=1}^{N_2} c_l^{m,n} c_l^{r,s} \underbrace{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j}}_{=n_l} \langle \Phi(g^{m,n}), \Phi(g^{r,s}) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{i,j}) \rangle_{\mathcal{H}} - 2 \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} c_l^{m,n} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{m,n}) \rangle_{\mathcal{H}} \\
&\quad + n_l \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} \sum_{r=1}^{N_1} \sum_{s=1}^{N_2} c_l^{m,n} c_l^{r,s} \langle \Phi(g^{m,n}), \Phi(g^{r,s}) \rangle_{\mathcal{H}} \\
&=: h_l \left(c_l^{1,1}, c_l^{1,2}, \dots, c_l^{1,N_2}, c_l^{2,1}, c_l^{2,2}, \dots, c_l^{2,N_2}, \dots, c_l^{N_1,N_2} \right).
\end{aligned}$$

The function h_l is a function of the vector consisting of all $c_l^{p,q}$, $(p,q) \in \mathcal{N}$. This function is continuously differentiable. To determine the $c_l^{p,q}$ such that the gradient of h_l vanishes, we compute the entry of the gradient of h_l corresponding to the partial derivative with respect to $c_l^{p,q}$ for an arbitrary but fixed $(p,q) \in \mathcal{N}$, which yields

$$\partial_{c_l^{p,q}} h_l \left(c_l^{1,1}, c_l^{1,2}, \dots, c_l^{1,N_2}, c_l^{2,1}, c_l^{2,2}, \dots, c_l^{2,N_2}, \dots, c_l^{N_1,N_2} \right)$$

$$= -2 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{p,q}) \rangle_{\mathcal{H}} + 2n_l \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} c_l^{m,n} \langle \Phi(g^{m,n}), \Phi(g^{p,q}) \rangle_{\mathcal{H}}.$$

Setting the derivative to 0 leads us to the necessary condition

$$\frac{1}{n_l} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u_l^{i,j} \langle \Phi(g^{i,j}), \Phi(g^{p,q}) \rangle_{\mathcal{H}} = \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} c_l^{m,n} \langle \Phi(g^{m,n}), \Phi(g^{p,q}) \rangle_{\mathcal{H}}.$$

This condition has to hold for every $(p, q) \in \mathcal{N}$ to let the gradient with respect to all $c_l^{p,q}$ vanish. In particular, due to the convexity it is also sufficient for a minimum. One possible way to satisfy the condition is to choose

$$c_l^{m,n} := \frac{u_l^{m,n}}{n_l}$$

for every $(m, n) \in \mathcal{N}$. With a given u and this choice for $c_l^{m,n}$, we obtain for the mean embedded spectrum μ_l^Φ the closed-form solution (cf. (3.39))

$$\mu_l^\Phi = \frac{1}{n_l} \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} u_l^{m,n} \Phi(g^{m,n}),$$

which is simply the arithmetic mean of the embedded spectra corresponding to segment l in the current iteration. This justifies the choice of the embedded mean spectrum in (2.58) in Section 2.4.1. However, as explained in Section 2.4, the idea of kernels is to do the computations in $\mathcal{H}$ without an explicit representation of Φ . Hence, we will not use the derived formula to explicitly compute μ_l^Φ , but the final kernelized formulation of $f_{\Phi,l}(i, j; \mu_l^\Phi)$ in (2.59), in which the derived expression of μ_l^Φ was incorporated and which allows us to compute the indicator values only based on kernel evaluations. The indicator values are enough to update u , which gives us, in the end, the segmentation of the input image g .

3.4 Computational complexity of ϵ AMS

For the computational complexity of the full segmentation framework based on the ϵ AMS model, we make the following considerations. The notation is the same as in the respective sections. Please recall that $\tilde{L}$ describes the number of channels before the MNF is applied (cf. Section 1.5) and $m_{\mu, \Sigma}^{\max}$ and $m_{\text{PDHG}}^{\max}$ denote the maximum number of fixed point iterations in one application of the fixed point iteration scheme and the maximum number of PDHG iterations in one application of the PDHG method in one inner iteration, respectively.

The preprocessing by the MNF involves an eigendecomposition with complexity $\mathcal{O}(\tilde{L}^3)$ and a projection of the spectra onto the determined components which has a complexity of $\mathcal{O}(\tilde{L}^2 N_1 N_2)$. This results in a complexity of the MNF of $\mathcal{O}(\tilde{L}^3 + \tilde{L}^2 N_1 N_2)$ [50]. The initial segmentation is found with one application of the k -means algorithm at the cost of $\mathcal{O}(k N_1 N_2 m_{k\text{-means}}^{\max})$ [71]. The symbol $m_{k\text{-means}}^{\max} \in \mathbb{N}$ describes the maximum number of iterations for the k -means algorithm. The next step is to initialize μ_l and $\Sigma_{\epsilon,l}$, including the regularization of the covariance for every $l \in \{1, \dots, k\}$. This includes the computation of the empirical covariance with complexity $\mathcal{O}(L^2 N_1 N_2)$ and an eigendecomposition for the regularization with $\mathcal{O}(L^3)$ operations, resulting in $\mathcal{O}(L^3 + L^2 N_1 N_2)$ steps. Then, we update μ_l and $\Sigma_{\epsilon,l}$ for any $l \in \{1, \dots, k\}$ with one application of the fixed point iteration scheme. This has complexity of $\mathcal{O}(m_{\mu,\Sigma}^{\max} k [L^3 + L^2 N_1 N_2])$. The reason is that we need to perform in each inner iteration a regularization involving an eigendecomposition of $\mathcal{O}(L^3)$ and in the update formula (cf. (3.22)) of Σ_l an outer product for any spectrum has to be computed which can be done with $\mathcal{O}(L^2 N_1 N_2)$ operations. The update of u consists of an evaluation of the indicator function $f_{\eta,l}$ with $\mathcal{O}(k L^2 N_1 N_2)$ steps and the application of the PDHG algorithm, which requires $\mathcal{O}(m_{\text{PDHG}}^{\max} k N_1 N_2 \log k)$ due to the computations of the proximal operators, i.e., the projection onto the unit simplex with $\mathcal{O}(k N_1 N_2 \log k)$ operations [106] and the projection onto the unit ball with $\mathcal{O}(k N_1 N_2)$ operations. Hence, one update of u costs in total $\mathcal{O}(k L^2 N_1 N_2 + m_{\text{PDHG}}^{\max} k N_1 N_2 \log k)$. We alternate between updating μ_l and Σ_l for every $l \in \{1, \dots, k\}$ and updating u at most $m^{\max}$ times, which is the maximum number of outer iterations. Therefore, when putting everything together, the final computational complexity is

$$\begin{aligned} & \mathcal{O}(\tilde{L}^3 + \tilde{L}^2 N_1 N_2 + m_{k\text{-means}}^{\max} k N_1 N_2 \\ & + m^{\max} [m_{\mu,\Sigma}^{\max} k (L^3 + L^2 N_1 N_2) + m_{\text{PDHG}}^{\max} k N_1 N_2 \log k]) . \end{aligned}$$

We observe that the complexity is logarithmic-linear in the number of segments and linear in the number of pixels. It is cubic in the number of channels.

Chapter 4

Numerical results

In this chapter, we will describe the results of the numerical experiments we carried out to see how our models perform in practice and in comparison to other state-of-the-art unsupervised hyperspectral image segmentation methods. In a nutshell, we will test our models on four hyperspectral datasets that are commonly used in remote sensing. We will see that our models ϵ AMS and ϕ AMS provide segmentations of consistently high quality and beat their competitors. kMS with all three kernels is also able to yield segmentations that achieve high performance scores. However, the experiments show that it is remarkably hard to choose suitable hyperparameters for the kMS models. We will start with an introduction of the used datasets, the performance metrics and an explanation of the experimental setting in Section 4.1. In Section 4.2, we present the results of all competing methods and comment on the effect of the MNF, the convergence of our optimization algorithm and the effects of different parameter choices in the PDHG method. The chapter closes with a presentation of the results of a study in which we applied ϵ AMS to multispectral images captured over the Arctic region to investigate if the model is able to derive complex sea ice states from the data (cf. Section 4.3).

The results of the experiments with ϵ AMS as well as the evaluation of the effect of the MNF have already been published in [26]. The findings about the derivation of complex sea ice states from multispectral images have already been published in [28], which was joint work together with Ekaterina Kim from the Norwegian University of Science and Technology (NTNU) in Norway, who contributed the interpretation of the segmentation results that focus on ice research.

Please note that the segmentations are visualized by converting them into RGB images by assigning a unique color to each of the segments such that we have for each pixel RGB values that specify the color that represents the pixel's segment. The pixel labels and assigned colors themselves only stand for the respective segment but do not have any further meaning. It would also be possible to choose other segment

labels or colors for the segments as long as the pixels of one segment are marked with the same label and color and different segments have different labels and colors. The color coding used in the subsequent sections to visualize the segmentations is adapted from [48].

4.1 Experimental setup

We now describe our experimental setup for investigating the performance of the introduced methods for unsupervised hyperspectral image segmentation. The methods are tested on four publicly available datasets that we briefly introduce in Section 4.1.1. In Section 4.1.2, we describe the performance metrics that we use to quantify the quality of the produced segmentations and to be able to compare the proposed methods with other methods quantitatively. The choices for the hyperparameters of the models that we use are explained in Section 4.1.3.

4.1.1 Datasets

The datasets we use as benchmarks to test our methods and compare them to other models are *Indian Pines (IP)*, *Kennedy Space Center (KSC)*, *Pavia University (PU)* and *Salinas (SA)*. These datasets as well as a corresponding manually annotated segmentation for each of these datasets are publicly available¹. The provided segmentations are considered as the true and correct segmentations and therefore called *ground truths*. They are later used as the reference segmentations we compare the segmentations obtained from the segmentation methods with. Please note that visualizations of ground truths of the benchmark datasets contain black pixels, which are an exception to the comment above that the colors and labels do not have meanings. Because the ground truths of all datasets contain unlabeled pixels that were not annotated by an expert, we are missing the information to which segments these pixels indeed belong. Therefore, these unlabeled pixels in the ground truth always obtain label 0 and are marked in black in visualizations of segmentations. All considered images in this chapter are two-dimensional images. Hence, the dimension of the image domain is $d = 2$.

Indian Pines

This hyperspectral image was acquired airborne in North-western Indiana, USA, with an AVIRIS sensor. It has a size of 145×145 pixels with 220 spectral channels each. In the publicly available version of the dataset we use, the water absorption bands were discarded, which means that we are left with 200 spectra per pixel. Two thirds

¹https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

of the captured scene are fields for agriculture, one third represents forest and other vegetation. A rail line, two major highways and some small roads are also in the image. The image was taken in June and hence, the crops growing on the fields are still in an early stage. The crop types comprise corn, oats, soybean and wheat. The ground truth classifies the pixels into 16 classes or segments of different land cover. The image depicted as a false-color image by interpreting three of its channels as an RGB image and the corresponding ground truth are shown in Fig. 4.1. The channels used to create the false-color image were chosen such that the structure of the image is visible. One can see the highways in the upper part of the image and also the fields

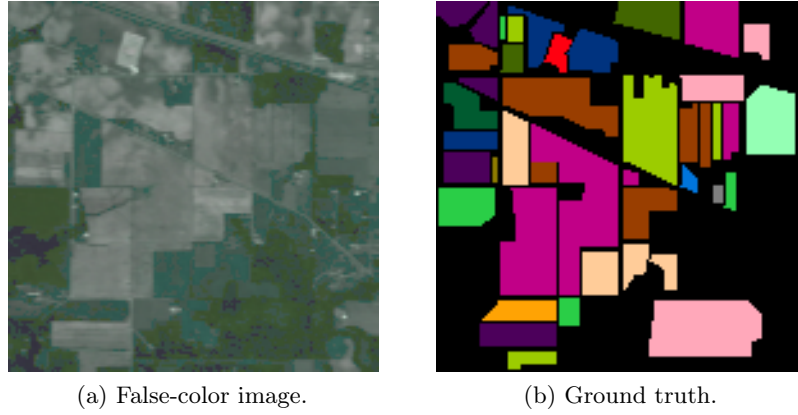


Figure 4.1: The IP dataset.

with the crops in Fig. 4.1a. The different colors in the ground truth correspond to the segments. Since the ground truth is manually labeled, it is known that for this image, the segments correspond to fields of different crop types.

Kennedy Space Center

The KSC dataset was also recorded with an AVIRIS sensor. It was taken over the Kennedy Space Center in Florida in 1996. The image has a size of 512×614 pixels and 224 spectral channels per pixel. Noisy channels and the water absorption channels were rejected in the publicly available version of the dataset, reducing the channels to 176 channels per pixel. A false-color image and the ground truth can be found in Fig. 4.2. Again, the channels for the false-color image were chosen such that the structure of the image is visible. In total, 13 meaningful segments are labeled. We can see in the ground truth that only a noticeably small amount of pixels is labeled, which can be used to evaluate the methods. Therefore, the segmentation results on KSC have a limited expressiveness with respect to the comparison of different segmentation

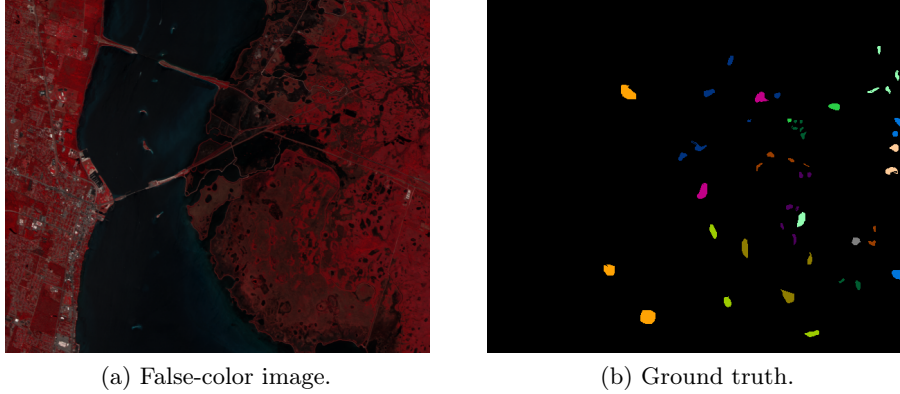


Figure 4.2: The KSC dataset.

methods. Nevertheless, we use this dataset for evaluation of the methods since it is a commonly used dataset in the hyperspectral imaging community.

Pavia University

The third dataset we will use to test our methods is PU. It was acquired airborne with a ROSIS sensor over Pavia in northern Italy. A rural area was captured such that constituents like gravel, trees, bitumen and bare soil are present. The size of the image is 610×340 pixels with 103 spectral channels per pixel. Fig. 4.3 shows a false-color image made from three channels that revealed the image structure and the corresponding manually annotated ground truth that categorizes the pixels into nine segments.

Salinas

The fourth dataset we use in our experiments is SA, which was recorded over the Salinas Valley in California, USA. It was acquired with an AVIRIS sensor. The original image has a size of 512×217 pixels with 224 spectral channels each. In the publicly available version we use, the water absorption bands were removed, resulting in 204 spectral channels. There are mainly agricultural land cover classes in the image, like vineyards, lettuce fields and broccoli fields. A false-color image from three channels showing the image structure and the ground truth are visible in Fig. 4.4. The ground truth distinguishes between 16 land cover types, which are the annotated segments.

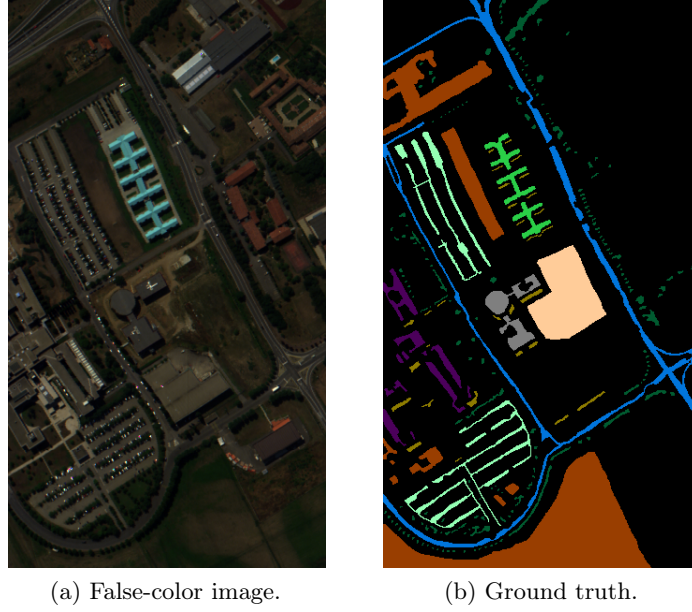


Figure 4.3: The PU dataset.

4.1.2 Performance metrics

To measure the performance of the methods introduced in this thesis and compare it to other methods quantitatively, we use the overall accuracy (OA), the average class accuracy (AA) and the kappa coefficient (KC) [25] as performance metrics. These metrics are commonly used in hyperspectral imaging. They measure essentially the pixel-wise agreement of a given segmentation with a reference segmentation. If a ground truth is given for a dataset, it is usually considered as the true and correct segmentation of the image and, therefore, used as the reference to check how much the segmentation that a method generates agrees with it.

The computation of all three metrics is based on the *confusion matrix*, which is a matrix $C = (c_{i,j})_{(i,j) \in \{1, \dots, k\}^2} \in \mathbb{N}^{k \times k}$ we can think of as a table where the rows carry the true segment labels from 1 to k and the columns the predicted segment labels from 1 to k . Then, the (i, j) -th entry of C is the number of pixels belonging to segment i , given by the ground truth, and predicted as segment j by the method. Hence, this generalizes the idea of a binary confusion matrix where we have the true-positive and true-negative samples on the diagonal and false-positive and false-negative samples below and above the diagonal. Therefore, the confusion matrix tells us how many pixels of a certain segment or class were classified as differing segments and the numbers on the diagonal correspond to the numbers of the correctly classified pixels.



Figure 4.4: The SA dataset.

Every pixel contributing to a number off the diagonal is wrongly classified.

Matching segment labels

A problem that arises in unsupervised segmentation is that the segment labels are only used to group the pixels and do not carry semantic information. If the meanings of the segments are needed, they have to be added to the found segments by an expert afterwards, in general. This procedure is called *a posteriori labeling* and was discussed in Section 1.2. This means that two segmentations are considered equal as long as the pixels sharing a segment label in the first segmentation also share a segment label in the second segmentation, which does not necessarily have to be the same label as in the first segmentation. For example, if we consider the pixels in segment $l \in \{1, \dots, k\}$ in the first segmentation, they all carry segment label l . If we now consider the same pixels in the second segmentation, we consider this segment in the first and second segmentation equal if all of these pixels also have the same label in the second segmentation, say, $m \in \{1, \dots, k\}$ and possibly with $m \neq l$, and there are no other pixels in the second segmentation carrying label m . However, if we want to use the confusion matrix to measure the similarity of two segmentations of the same

image, the described case with $m \neq l$ is not properly reflected in the confusion matrix since the number of pixels in segment l in the first segmentation that are classified as segment m in the second segmentation would appear off the diagonal in the (l, m) -th entry of the confusion matrix. But in the sense of the confusion matrix introduced in Section 4.1.2 we expect the number of pixels where the two segmentations coincide on the diagonal. Hence, when comparing two segmentations using the confusion matrix, we need consistent labels in the two segmentations to make sure that we compare the correct segments of the two segmentations with each other. For two exactly equal segmentations of one image that only use different segment labels, we expect that the confusion matrix has positive values only on the diagonal and is zero off the diagonal because no pixels are wrongly classified in the sense described above. This can be achieved by permuting the labels of one of the segmentations. If the two compared segmentations are not exactly equal, as it is the case in general, we will have non-zero entries off the diagonal in the confusion matrix. In this case, we permute the labels of one segmentation with the permutation leading to the best matching of segment labels with respect to the biggest overlap in terms of pixels between the segments of the true and the predicted segmentation. This permutation can be found based on the confusion matrix with the current labels by solving the linear assignment problem. This problem is about finding a bijection between two finite sets such that a cost function is minimized or maximized. The two sets are, in our case, the true and the predicted segment labels and the cost function is the number of correctly classified pixels. We, therefore, have to permute the columns of the confusion matrix, which correspond to the predicted labels, such that the sum on the diagonal is maximized. This assignment problem can be solved in polynomial time using, e.g., the Hungarian algorithm [60]. Having the confusion matrix with the correct matching of the segment labels, we can compute the performance metrics based on it. Hence, from now on, we assume that we have found the optimal matching of segment labels in the sense that the overlap of the segments is maximized and always compute the performance metrics based on that matching.

Overall accuracy

The OA is the fraction of correctly classified pixels. In formulas, the OA is computed via

$$\text{OA} := \frac{1}{N_1 N_2} \sum_{l=1}^k c_{l,l}. \quad (4.1)$$

Please recall that the number of pixels in the two dimensions of the image domain is given by $N_1 \in \mathbb{N}$ and $N_2 \in \mathbb{N}$. Hence, we sum the diagonal elements of the confusion matrix, which are the numbers of pixels the two compared segmentations agree on

(correctly classified), and divide this sum by the number of all pixels. If all pixels are correctly classified, meaning that off the diagonal in the confusion matrix, we have only zeros, then $OA = 1$. The scenario in which we do not have any correctly classified pixels is not possible. By permuting the columns of the confusion matrix, we can always achieve that there is at least one non-zero entry on the diagonal. The lowest achievable value of the score for a specific image depends on the ground truth. However, one can generally say that the OA cannot go below 0. Therefore, it always attains values in the interval $[0, 1]$ and we want the score ideally to be equal to 1.

A drawback of this performance metric is that it does not take into account the sizes of the segments. For instance, if we imagine an image where 90 % of the pixels are expected to be classified as one segment and the remaining 10 % of the pixels are divided into five other segments, then a segmentation method can achieve 90 % accuracy by putting all pixels into one segment. Such a segmentation is essentially useless as it does not extract any meaningful information from the data. Nevertheless, this metric is often used in the hyperspectral imaging community.

Average class accuracy

The AA is closely related to the OA. Here, we also use the fraction of correctly classified pixels but consider it segment-wise (class-wise) and compute the mean over the segment or class accuracies. Let $n_l := \sum_{j=1}^k c_{l,j}$ be the number of pixels in segment $l \in \{1, \dots, k\}$, given by the reference segmentation. In formulas, we compute the AA with the expression

$$AA := \frac{1}{k} \sum_{l=1}^k \frac{c_{l,l}}{n_l}.$$

Obviously, this performance metric does not suffer from the same drawback as the OA since the cardinalities of the segments are taken into account. If all pixels are correctly classified, we have only zeros off the diagonal in the confusion matrix and therefore $c_{l,l} = n_l$ for every $l \in \{1, \dots, k\}$, leading to a score of 1, which is the maximum possible value. As in the case of the OA, an AA score of 0 is not possible but the lowest achievable value depends on the ground truth. One can merely generally say that the AA score is always non-negative. Hence, the AA score attains values in $[0, 1]$ and a higher value means a better agreement of the segmentation and the reference. In combination with the OA, we can extract more information from the AA. For example, if the OA is high, but the AA rather low, we can deduce that segments with many pixels are well detected, while small segments with only a few pixels could not be correctly detected. In that case, the class accuracy of the small segments is close to 0, letting the AA drop while the OA can still be a high number.

Kappa coefficient

The KC [25] considers not only the observed agreement between the two segmentations but also an agreement by chance that might occur. Let $p_o \in [0, 1]$ describe the fraction of pixels the two compared segmentations agree on and $p_c \in [0, 1]$ the fraction of pixels we expect an agreement by chance on. Please note that p_o coincides with OA. Then, the KC is computed with the formula

$$\text{KC} := \frac{p_o - p_c}{1 - p_c}.$$

The formula describes that we consider the fraction of correctly classified pixels *after* we have removed the pixels the segmentations agree on by chance. If we assume that there is no agreement by chance, i.e., $p_c = 0$, we obtain $\text{KC} = p_o = \text{OA}$. The maximum possible value of KC is 1 but it can also be negative. It becomes negative if the estimated agreement by chance exceeds the observed agreement, i.e., when $p_o < p_c$. In case there is no agreement except for the agreement that occurred by chance, we obtain $p_o = p_c$ and therefore $\text{KC} = 0$. If the observed agreement exceeds the agreement by chance, we obtain positive values for KC. If (and only if) we have a perfect agreement between the two segmentations, i.e., $p_o = 1$, then $\text{KC} = 1$. Hence, the closer the KC to 1, the better the agreement of the compared segmentations. To compute p_o , we can use the formula for the computation of the OA in (4.1). The quantity p_c can be estimated as follows: first, we note that for the KC, it is assumed that the two segmentations are independent. This assumption leads to the fact that we can estimate the probability that both segmentations classify a pixel as segment $l \in \{1, \dots, k\}$, denoted by p_l^{or} , by using the product of the estimated probabilities that the observed segmentation classifies a pixel as l , which we denote by p_l^o and that the reference segmentation classifies a pixel as l , described by p_l^r . These two probabilities are obtained as the marginal distributions of the confusion matrix. Let n_l^o and n_l^r be the number of the pixels classified as segment $l \in \{1, \dots, k\}$ by the observed segmentation and the reference segmentation, respectively. Then, we get from the confusion matrix

$$n_l^o = \sum_{i=1}^k c_{i,l}, \quad n_l^r = \sum_{j=1}^k c_{l,j}.$$

Hence, using the independence of the observed and reference segmentation in the second equality, we obtain

$$p_c \approx \sum_{l=1}^k p_l^{or} = \sum_{l=1}^k p_l^o p_l^r = \sum_{l=1}^k \frac{n_l^o}{N_1 N_2} \frac{n_l^r}{N_1 N_2},$$

which then allows for the computation of the KC.

4.1.3 Settings

We test our segmentation models on different variants of the described datasets because a part of our full segmentation framework is a preprocessing step using the MNF (cf. Sections 1.5 and 3.1.3). Hence, it is necessary to evaluate the effect of the MNF on the segmentation results. To this end, we apply our segmentation models to the raw data, i.e., without any preprocessing, to the data after an application of the MNF according to our segmentation framework and to the data after applying a feature learning method (more details are given below). In all cases, we rescale the image data such that the spectral intensity values of the input image are normalized to the interval $[0, 1]$, i.e., for any $(i, j) \in \mathcal{N}$, we have $g(i, j) \in [0, 1]^L$. This is achieved by subtracting the minimum intensity over all pixels and all channels from all intensities to make the minimum intensity equal to 0. The intensity values of the resulting image are then divided by the maximum intensity over all pixels and channels, making the maximum intensity equal to 1. This linear transformation does not change the structure of the data, only the scaling, but it can help to avoid overflows during numerical optimization. All experiments described in this and the next section, i.e., Sections 4.1 and 4.2, are carried out with float precision.

Competing methods

We compare our three models ϵ AMS, ϕ AMS and kMS to several other methods, including common unsupervised clustering algorithms and state-of-the-art methods specifically designed for unsupervised hyperspectral image segmentation. The competing methods are:

- *Clustering approaches*: k -means, GMM, BGM,
- *Hyperspectral image segmentation approaches*: MS model with squared 2-norm as indicator function (MS-2), sequential spectral clustering with RBF kernel (SSC) [53], MS model with a non-local TV regularizer (NLTV) [116], 3D convolutional autoencoder with clustering layer in the bottleneck layer (3D-CAE) [78].

We use the scikit-learn implementations of k -means, GMM and BGM with their standard configurations, except for the maximum number of iterations of BGM that we increase to 1000 to allow it to converge.

The model MS-2 is given as J_{MS} in (2.9) with the squared Euclidean difference of the spectrum and the mean spectrum as indicator function. We use the described alternating optimization approach (cf. Section 3.1.3) and run it with the same number of PDHG iterations, the same number of outer iterations, thresholds and stopping criterion for the outer iterations (cf. Section 3.1.5) as ϵ AMS, ϕ AMS and kMS. The hyperparameter choices are given in the next section.

The methods SSC, NLTV and 3D-CAE have already been briefly described in the related works section in Section 1.3. Of particular interest is NLTV since it is also based on the MS segmentation model as our models. However, there are some important differences, e.g., that the labeling function u is squared in the case of NLTV, that it uses a non-local gradient of u to obtain a non-local version of the TV and that the indicator function uses only first-order statistics. Unfortunately, the source code to apply SSC and NLTV for segmentation was not publicly available and these methods are quite complex. Hence, we did not reimplement the methods and will only report the scores from the respective articles. This also means that we have only performance scores for SSC and NLTV on the raw data and not after MNF and the feature learning.

We run 3D-CAE with the publicly available code and use the authors' architecture and hyperparameters. The authors train the network for 25 epochs and set the latent dimension in the bottleneck layer to 25. We do the same for the raw data and the data after feature learning. Only in the case of the MNF data, we use the chosen number of MNF channels as the latent dimension since these numbers are between 5 and 8 and, therefore, significantly smaller than the default of 25. We use different batch sizes for the datasets. They are specified in the next section about the hyperparameter choices. A problem that occurs with 3D-CAE and the data after applying the MNF is that 3D-CAE reduces the number of channels by 8. Hence, it needs at least 9 channels in the input image. However, we let the MNF reduce the number of channels of the datasets during the preprocessing to at most 8 channels. Therefore, 3D-CAE is not directly applicable. To solve this problem, we append channels filled with zeros to the datasets after MNF application until they all have 9 channels and 3D-CAE is applicable.

The number of segments k is given as prior information to all methods. Since in every method randomness plays a certain role, e.g., for the initialization, we apply every method (except for SSC and NLTV because we do not have implementations for them) to each variant of the datasets ten times using ten randomly sampled seeds for the random number generator to analyze the stability of the results of the methods against the randomness element.

Unsupervised feature learning

The feature learning method that we apply as a preprocessing to compare it to the MNF preprocessing is an unsupervised feature learning approach based on a spectral-spatial 3D convolutional autoencoder (FL) [73]. The model receives a patch of pixels from an image as the input and processes it with a series of different layers that only use three-dimensional or element-wise operations. Three-dimensional operations mean in this context that the kernels or windows of such layers are three-dimensional in the sense that they cover several pixels and several channels of these pixels at

once and operate along the spatial dimensions *and* spectral dimension. Examples of such layers are 3D convolution, 3D pooling or 3D batch normalization. The layer separating the encoder from the decoder, often called bottleneck layer, outputs the learned features for the central pixel in the input patch, from which the input patch is then reconstructed by the decoder during training. In the case of the datasets considered in this thesis, this layer outputs a vector of 432 features. The learned features for an image are obtained by an unsupervised training of FL on all patches of a predefined size from that image and using the outputs of the bottleneck layer, i.e., ignoring the decoder part, as the new channels for the central pixels in the patches. This means that we have an increase of the number of channels per pixel to 432.

The code for FL is publicly available. We applied the model as it is with the hyperparameters that the authors recommend. We described in Section 4.1.1 that the data providers of the datasets we consider provide the datasets without the noisy and water absorption channels. However, we would like to point out that the data processing pipeline of FL defined by the authors adds these channels back and fills them with zeros before the training. Since we did not change the model but applied it as proposed by the authors, we followed this procedure in our experiments. We use the resulting images with the learned features as the input (without application of the MNF) for all competing models to test them on the learned features.

Because the noisy and water absorption channels are added back to the data to apply FL, we need to know which channels were removed in order to recover the original structure of the image and their channels. This information is only publicly given for the datasets IP, PU and SA. Therefore, we did not apply the feature learning model to KSC.

Hyperparameter choices

We now specify the hyperparameters of the models we used to obtain the results presented in Section 4.2. Some of the parameters are chosen image-independent, meaning that the same choice for these parameters was used across all experiments on the different datasets. The remaining parameters are selected image-dependent, i.e., these parameters were chosen such that we obtain a good performance on one dataset but possibly a bad result on another dataset with the same choice. Tables 4.1 to 4.3 summarize the choices for the data-independent parameters. The contextual information for the contextual kernel in kMS is chosen as the mean spectrum in a local neighborhood of each pixel. The local neighborhood is given as the square around the pixel with the considered pixel in the center. We denote the length of the sides of the square by the odd number $r = 2s + 1$ with $s \in \mathbb{N}$ and specify the length r in terms of the number of pixels defining the side of the square. The local neighborhood of pixels leads to the problem that for boundary pixels, this square neighborhood goes over the image boundary. We use the replication of the spectra

Stopping thresholds		
Outer iterations	PDHG iterations	Fixed point iterations
10^{-6}	10^{-6}	10^{-5}

Maximum number of iterations		
Outer iterations	PDHG iterations	Fixed point iterations
20	1000	20

PDHG parameters		
τ	κ	ϑ
0.125	0.125	1

Model parameters	
η	
10^{-2}	

Table 4.1: Image-independent parameters of ϵ AMS.

at the boundary as the padding strategy, i.e., constant extension with the boundary spectrum in normal direction of the boundary, to extend the image such that each pixel of the original image has a local neighborhood. The parameters $\delta = 1$ and $c = 1$ of the polynomial kernel were chosen in accordance with the choices in [19].

The choices for the image-dependent parameters are given in Table 4.4. The numbers of segments k in the test datasets are given by the numbers of labeled segments in the ground truths. The remaining parameters had to be chosen manually in some way, which was a challenging task since the noise and spectral variability have different strengths in the datasets. Unfortunately, we were not able to find simple heuristics on how to choose these parameters. Therefore, we tuned the number of channels kept by the MNF manually before the experiments and left them fixed across all experiments. The other parameters were found by applying the hyperparameter tuning framework Optuna [3], which uses the ground truth to define an objective function that is then optimized over the hyperparameter search space. We used the OA score as the objective function for Optuna. After Optuna returned a set of hyperparameters, a grid search around these parameters was performed and the best-performing parameters with respect to OA reported in the table. The grid search was performed because Optuna tries to sample the parameter space by balancing exploitation of the previous samples and results and exploration of the whole space. Therefore, there is a chance to find parameters that lead to a better result than the one found by Optuna when a fine-tuning with a grid search is performed. Please note that it rarely happened that the grid search result led to a worse score than the

Stopping thresholds		
Outer iterations	PDHG iterations	Fixed point iterations
10^{-6}	10^{-6}	10^{-5}

Maximum number of iterations		
Outer iterations	PDHG iterations	Fixed point iterations
20	1000	50

PDHG parameters		
τ	κ	ϑ
0.125	0.125	1

Backtracking line search (cf. Algorithm 3.2)		
$\bar{\rho}$	c	γ
1	0.75	0.5

Model parameters		
ϕ		
$\phi(t) = \exp(-t)$ (cf. (2.53))		

Table 4.2: Image-independent parameters of ϕ AMS.

Optuna search. Nevertheless, the grid search result was always reported for consistency reasons. The hyperparameter tuning with Optuna as well as the grid search were applied with a fixed seed for the random number generator that is responsible for the randomness in the methods to ensure reproducibility. This seed is not among the randomly sampled seeds mentioned above but later in the results tables (cf. Tables 4.5 to 4.7) referred to with the word *fixed*. More details are given in Section 4.2.1. It is not ideal to find good parameter choices using the ground truth because a ground truth is usually unavailable in practice. Nevertheless, we went this way to show the potential of the developed methods. Hence, this work can be seen as a proof of concept of the proposed methods. Heuristics for practical use remain open questions for future work. However, if a segmentation method is applied in an unchanging environment in practice, one can tune the parameters manually once with the help of an expert and keep them fixed for the future. Since the environment does not change, there is a good chance that the chosen parameters lead to consistently useful results. This is what we also did in the experiments in Section 4.3 to derive complex sea ice states.

The ϕ AMS model could only be applied to the raw data and the data after an application of the MNF. Due to the complexity of the optimization on the Stiefel

Stopping thresholds		
Outer iterations	PDHG iterations	
10^{-6}	10^{-6}	

Maximum number of iterations		
Outer iterations	PDHG iterations	
20	1000	

PDHG parameters		
τ	κ	ϑ
0.125	0.125	1

Model parameters		
K_ω	K_s	$K_{\omega s}$
Polynomial kernel (cf. (2.63)) $\delta = 1, c = 1$	Gaussian RBF kernel (cf. (2.62))	Polynomial kernel (cf. (2.63)) $\delta = 1, c = 1$
Contextual information		Padding for boundary pixels
Mean spectrum in local neighborhood		Replicate boundary values

Table 4.3: Image-independent parameters of kMS.

manifold (cf. Section 3.3.2), much more memory is needed for the optimization of ϕ AMS than of MS-2 or ϵ AMS. In addition, the feature learning FL increases the number of channels to 432, which also contributes to the higher memory need. Both together exceeded the available computational resources and we, therefore, decided to forgo the experiments on the FL variants of the datasets in the case of ϕ AMS. Furthermore, it was already mentioned that the memory need of a kernel matrix scales quadratically with the number of pixels since the matrix has the size *number of pixels* $\times$ *number of pixels* (cf. Section 2.4.1). This let us also in the case of experiments with the kernel methods on PU and KSC reach the limits of our computational resources. With float precision, the memory demand to store the kernel matrix of PU and KSC is approximately 172 GB and 395 GB, respectively.

The reason why no hyperparameters for the variants of kMS on the FL variants of IP and SA are given is that the kernel methods need an extremely careful choice of hyperparameters to converge to a local minimum without empty segments. The problem with empty segments is that if an empty segment occurs during optimization, we are not able to estimate relevant parameters, like mean spectrum, covariance,

		IP			SA		
	k	16			16		
MNF	L	8			7		
		raw	MNF	FL	raw	MNF	FL
MS-2	λ	0.0021	0.33	0.00023	0.0009	0.135	$3.75 \cdot 10^{-5}$
ϵ AMS	λ	0.078	0.24	0.21	0.1	0.15	0.00042
	ϵ	0.008	0.125	0.0025	0.0156	0.115	0.112
ϕ AMS	λ	0.109	0.296	-	0.156	0.2	-
	α	0.09	7.88	-	0.0066	5.5	-
	β	5.62	9.38	-	8.8	2	-
kMS-	λ	0.00605	0.26	-	0.0078	0.26	-
	r	5	3	-	3	9	-
DSK	p	5	8	-	3	2	-
	γ	43	8.5	-	18.1	46.8	-
kMS-	λ	0.0045	0.29	-	-	0.065	-
	r	5	3	-	-	3	-
	α	0.55	0.2	-	-	0.64	-
WSK	p	1	7	-	-	9	-
	γ	12.125	45	-	-	0.05	-
kMS-	λ	0.013	0.91	-	-	0.67	-
	r	9	7	-	-	5	-
CIK	p	9	7	-	-	10	-
	γ	44	30	-	-	18.5	-
3D-CAE	Batch size	841			992		

		KSC			PU		
	k	13			9		
MNF	L	6			5		
		raw	MNF	FL	raw	MNF	FL
MS-2	λ	0.0003	0.094	-	0.0009	0.0565	$6.85 \cdot 10^{-5}$
ϵ AMS	λ	0.0014	0.0005	-	0.0101	0.008	0.155
	ϵ	0.18	0.5	-	0.03	0.3	$5.46 \cdot 10^{-5}$
ϕ AMS	λ	0.0028	0.066	-	0.06	0.0223	-
	α	0.0046	6.1	-	0.0024	0.037	-
	β	5.096	4.15	-	5	2.95	-
3D-CAE	Batch size	1024			1220		

Table 4.4: Choices of image-dependent model parameters.

etc. for that segment anymore and stop iterating in our implementation. The same problem happened for kMS with WSK and CIK kernel when applied to the raw SA dataset. In all these cases we could not find a hyperparameter setting that did not stop earlier because of an empty segment.

4.2 Results

We will now analyze the segmentation results obtained from the different methods. A quantitative and qualitative comparison of the segmentation results of the competing methods is given in Section 4.2.1. Afterward, in Section 4.2.2, we investigate the effect of the MNF on the data and also the results. Furthermore, to evaluate if our alternating optimization approach used for our developed methods is converging, we collected numerical indications that show a converging behavior. These results are summarized in Section 4.2.3. When we introduced the PDHG algorithm (cf. Section 3.2.1), we also commented on the choice of the step sizes τ and κ that are involved in the PDHG method in order to have a convergence guarantee (cf. Proposition 3.14). We chose in our experiments step sizes and a maximum number of PDHG iterations that are smaller than what is usually done in practice when the PDHG method is used. The reason and the impact of these choices are explained in Section 4.2.4.

4.2.1 Comparison of segmentation results

As mentioned above, in all methods, we have a component that introduces randomness. For example, in the case of MS-2, ϵ AMS, ϕ AMS and kMS, it is the initial segmentation that comes from the k -means method. Therefore, we accounted for this randomness by applying every method we had access to, i.e., except for SSC and NLTV, ten times with different seeds for the random number generator. The scores we report in the results tables below are the minimum (*Min*), mean (*Mean*) and maximum (*Max*) score as well as the standard deviation (*Std*) over the ten trials. Moreover, also the score achieved with the fixed seed, which was used for the hyperparameter tuning and is not among the ten other randomly sampled seeds, is given (*Fixed*). The authors of SSC reported the best result out of ten runs on each dataset with random initialization. These scores are reported in the table. They found the optimal parameters for their model with cross-validation based on the OA. Since the authors of the NLTV model do not mention which scores they report, we consider them as the *mean* scores and list them accordingly. The achieved scores by the segmentation methods on the different datasets are given in Tables 4.5 to 4.7. The score printed in bold in each column is the highest mean score achieved on the respective variant of the dataset. A green cell background is used to highlight the highest maximum score

Method	Image Type Stats	IP			KSC		PU			SA		
		raw	MNF	FL	raw	MNF	raw	MNF	FL	raw	MNF	FL
<i>k</i> -means	Min	0.360	0.485	0.393	0.178	0.616	0.534	0.509	0.603	0.669	0.707	0.636
	Mean	0.368	0.502	0.402	0.178	0.624	0.534	0.515	0.604	0.669	0.723	0.654
	Max	0.378	0.506	0.410	0.178	0.632	0.534	0.550	0.611	0.671	0.732	0.663
	Std	0.007	0.006	0.005	0.000	0.005	0.000	0.012	0.002	0.001	0.011	0.009
GMM	Min	0.376	0.489	0.447	0.178	0.591	0.499	0.495	0.576	0.666	0.683	0.591
	Mean	0.402	0.514	0.467	0.178	0.612	0.528	0.515	0.580	0.691	0.715	0.625
	Max	0.438	0.577	0.493	0.179	0.646	0.556	0.528	0.585	0.715	0.753	0.655
	Std	0.021	0.024	0.018	0.000	0.017	0.025	0.013	0.004	0.018	0.022	0.020
BGM	Min	0.379	0.488	0.357	0.178	0.580	0.443	0.463	0.565	0.668	0.696	0.572
	Mean	0.404	0.514	0.402	0.178	0.612	0.493	0.477	0.569	0.677	0.721	0.632
	Max	0.438	0.590	0.454	0.179	0.626	0.569	0.518	0.571	0.693	0.748	0.685
	Std	0.019	0.028	0.034	0.000	0.014	0.065	0.016	0.003	0.008	0.017	0.040
3D-CAE	Min	0.390	0.456	0.240	0.185	0.330	0.460	0.405	0.363	0.584	0.590	0.415
	Mean	0.414	0.473	0.391	0.213	0.446	0.540	0.477	0.468	0.595	0.625	0.444
	Max	0.436	0.486	0.426	0.309	0.600	0.599	0.561	0.542	0.601	0.657	0.479
	Std	0.015	0.009	0.055	0.051	0.069	0.044	0.055	0.047	0.005	0.026	0.020
SSC	Max	0.429	-	-	-	-	0.682	-	-	0.818	-	-
NLTV	Mean	0.424	-	-	0.415	-	0.440	-	-	-	-	-
MS-2	Min	0.475	0.553	0.442	0.179	0.676	0.564	0.518	0.576	0.733	0.713	0.655
	Mean	0.497	0.588	0.467	0.179	0.676	0.564	0.525	0.576	0.733	0.750	0.683
	Max	0.541	0.631	0.482	0.179	0.676	0.564	0.593	0.578	0.733	0.764	0.709
	Std	0.018	0.038	0.018	0.000	0.000	0.000	0.024	0.000	0.000	0.019	0.019
	Fixed	0.500	0.629	0.482	0.179	0.650	0.564	0.593	0.576	0.733	0.764	0.688
ϵ AMS	Min	0.505	0.582	0.482	0.179	0.666	0.562	0.560	0.605	0.807	0.723	0.646
	Mean	0.554	0.671	0.494	0.180	0.668	0.562	0.567	0.633	0.809	0.828	0.667
	Max	0.639	0.697	0.499	0.180	0.669	0.562	0.615	0.636	0.810	0.896	0.690
	Std	0.051	0.034	0.006	0.000	0.001	0.000	0.017	0.010	0.001	0.074	0.016
	Fixed	0.571	0.697	0.523	0.179	0.716	0.562	0.615	0.636	0.810	0.896	0.689
ϕ AMS	Min	0.496	0.574	-	0.178	0.619	0.603	0.558	-	0.739	0.709	-
	Mean	0.556	0.633	-	0.178	0.653	0.635	0.566	-	0.739	0.720	-
	Max	0.620	0.685	-	0.178	0.678	0.642	0.577	-	0.739	0.740	-
	Std	0.056	0.039	-	0.000	0.030	0.011	0.005	-	0.000	0.008	-
	Fixed	0.615	0.695	-	0.178	0.678	0.641	0.565	-	0.739	0.730	-
kMS-	Min	0.447	0.556	-	-	-	-	-	-	0.838	0.676	-
	Mean	0.503	0.579	-	-	-	-	-	-	0.838	0.718	-
	Max	0.555	0.588	-	-	-	-	-	-	0.838	0.729	-
DSK	Std	0.036	0.011	-	-	-	-	-	-	0.000	0.019	-
	Fixed	0.550	0.573	-	-	-	-	-	-	0.838	0.736	-
kMS-	Min	0.441	0.566	-	-	-	-	-	-	-	0.674	-
	Mean	0.509	0.590	-	-	-	-	-	-	-	0.742	-
	Max	0.543	0.631	-	-	-	-	-	-	-	0.755	-
WSK	Std	0.033	0.032	-	-	-	-	-	-	-	0.025	-
	Fixed	0.560	0.630	-	-	-	-	-	-	-	0.736	-
kMS-	Min	0.495	0.566	-	-	-	-	-	-	-	0.676	-
	Mean	0.532	0.646	-	-	-	-	-	-	-	0.699	-
	Max	0.609	0.669	-	-	-	-	-	-	-	0.701	-
CIK	Std	0.031	0.039	-	-	-	-	-	-	-	0.008	-
	Fixed	0.555	0.666	-	-	-	-	-	-	-	0.737	-

Table 4.5: OA of the competing methods on the four datasets.

Method	Image Type Stats	IP			KSC		PU			SA		
		raw	MNF	FL	raw	MNF	raw	MNF	FL	raw	MNF	FL
<i>k</i> -means	Min	0.375	0.457	0.268	0.079	0.405	0.526	0.582	0.474	0.687	0.631	0.565
	Mean	0.404	0.464	0.291	0.079	0.415	0.526	0.585	0.478	0.687	0.644	0.626
	Max	0.440	0.473	0.310	0.079	0.441	0.526	0.598	0.505	0.688	0.653	0.656
	Std	0.027	0.005	0.019	0.000	0.011	0.000	0.005	0.010	0.000	0.009	0.035
GMM	Min	0.361	0.481	0.301	0.078	0.396	0.420	0.466	0.398	0.547	0.609	0.461
	Mean	0.386	0.531	0.343	0.079	0.410	0.483	0.528	0.404	0.615	0.654	0.527
	Max	0.418	0.595	0.396	0.079	0.453	0.546	0.587	0.412	0.650	0.695	0.585
	Std	0.021	0.036	0.030	0.000	0.019	0.044	0.058	0.006	0.041	0.029	0.037
BGM	Min	0.357	0.469	0.216	0.078	0.393	0.406	0.429	0.383	0.536	0.615	0.533
	Mean	0.378	0.531	0.264	0.079	0.422	0.409	0.480	0.393	0.575	0.662	0.566
	Max	0.405	0.603	0.340	0.079	0.436	0.412	0.582	0.398	0.604	0.688	0.611
	Std	0.017	0.041	0.042	0.000	0.017	0.003	0.051	0.007	0.028	0.024	0.030
3D-CAE	Min	0.191	0.317	0.088	0.101	0.201	0.554	0.306	0.148	0.396	0.405	0.251
	Mean	0.224	0.355	0.196	0.115	0.274	0.573	0.356	0.209	0.410	0.473	0.303
	Max	0.255	0.409	0.237	0.161	0.388	0.598	0.405	0.297	0.419	0.540	0.363
	Std	0.019	0.032	0.042	0.024	0.056	0.018	0.036	0.052	0.007	0.039	0.035
MS-2	Min	0.405	0.529	0.300	0.078	0.459	0.539	0.548	0.424	0.724	0.631	0.539
	Mean	0.471	0.540	0.309	0.078	0.459	0.539	0.552	0.426	0.724	0.671	0.590
	Max	0.527	0.552	0.360	0.079	0.459	0.539	0.588	0.440	0.724	0.685	0.648
	Std	0.042	0.010	0.019	0.000	0.000	0.000	0.013	0.005	0.000	0.021	0.039
	Fixed	0.449	0.488	0.360	0.079	0.436	0.539	0.588	0.424	0.724	0.685	0.597
ϵ AMS	Min	0.497	0.538	0.293	0.078	0.454	0.601	0.613	0.452	0.671	0.639	0.568
	Mean	0.541	0.607	0.331	0.079	0.454	0.601	0.616	0.491	0.675	0.730	0.614
	Max	0.585	0.633	0.355	0.079	0.455	0.601	0.636	0.495	0.676	0.768	0.676
	Std	0.028	0.030	0.018	0.000	0.001	0.000	0.007	0.014	0.002	0.045	0.039
	Fixed	0.539	0.607	0.392	0.078	0.530	0.601	0.636	0.495	0.676	0.768	0.629
ϕ AMS	Min	0.461	0.477	-	0.079	0.440	0.689	0.606	-	0.677	0.633	-
	Mean	0.502	0.574	-	0.079	0.452	0.713	0.611	-	0.677	0.652	-
	Max	0.540	0.630	-	0.079	0.462	0.723	0.614	-	0.677	0.671	-
	Std	0.030	0.047	-	0.000	0.010	0.009	0.002	-	0.000	0.010	-
	Fixed	0.536	0.571	-	0.079	0.462	0.722	0.611	-	0.677	0.660	-
kMS-	Min	0.464	0.464	-	-	-	-	-	-	0.739	0.614	-
	Mean	0.509	0.508	-	-	-	-	-	-	0.739	0.678	-
	Max	0.574	0.536	-	-	-	-	-	-	0.739	0.691	-
DSK	Std	0.037	0.030	-	-	-	-	-	-	0.000	0.026	-
	Fixed	0.594	0.476	-	-	-	-	-	-	0.739	0.672	-
kMS-	Min	0.498	0.537	-	-	-	-	-	-	-	0.611	-
	Mean	0.558	0.544	-	-	-	-	-	-	-	0.670	-
	Max	0.590	0.548	-	-	-	-	-	-	-	0.686	-
WSK	Std	0.027	0.003	-	-	-	-	-	-	-	0.026	-
	Fixed	0.596	0.548	-	-	-	-	-	-	-	0.673	-
kMS-	Min	0.496	0.533	-	-	-	-	-	-	-	0.614	-
	Mean	0.534	0.561	-	-	-	-	-	-	-	0.673	-
	Max	0.605	0.589	-	-	-	-	-	-	-	0.684	-
CIK	Std	0.034	0.014	-	-	-	-	-	-	-	0.023	-
	Fixed	0.588	0.563	-	-	-	-	-	-	-	0.673	-

Table 4.6: AA of the competing methods on the four datasets.

Method	Image Type Stats	IP			KSC		PU			SA		
		raw	MNF	FL	raw	MNF	raw	MNF	FL	raw	MNF	FL
k-means	Min	0.289	0.426	0.314	0.002	0.560	0.434	0.421	0.482	0.632	0.669	0.590
	Mean	0.298	0.442	0.319	0.002	0.570	0.434	0.428	0.484	0.632	0.689	0.610
	Max	0.307	0.446	0.321	0.003	0.579	0.434	0.467	0.493	0.634	0.699	0.621
	Std	0.008	0.006	0.002	0.000	0.006	0.000	0.014	0.003	0.001	0.013	0.011
GMM	Min	0.305	0.439	0.359	0.001	0.531	0.383	0.400	0.437	0.623	0.644	0.538
	Mean	0.328	0.464	0.381	0.002	0.558	0.407	0.425	0.444	0.651	0.680	0.577
	Max	0.360	0.521	0.415	0.003	0.597	0.424	0.437	0.451	0.679	0.722	0.611
	Std	0.020	0.023	0.023	0.001	0.020	0.016	0.013	0.006	0.021	0.024	0.023
BGM	Min	0.309	0.439	0.250	0.001	0.525	0.320	0.356	0.426	0.626	0.658	0.529
	Mean	0.330	0.463	0.306	0.002	0.558	0.364	0.373	0.432	0.635	0.686	0.590
	Max	0.358	0.534	0.371	0.003	0.573	0.430	0.420	0.435	0.654	0.717	0.650
	Std	0.016	0.026	0.044	0.000	0.015	0.057	0.020	0.005	0.010	0.019	0.042
3D-CAE	Min	0.270	0.377	0.019	0.049	0.228	0.373	0.105	0.159	0.524	0.535	0.314
	Mean	0.302	0.395	0.258	0.080	0.366	0.446	0.296	0.217	0.536	0.573	0.347
	Max	0.335	0.416	0.305	0.191	0.543	0.501	0.377	0.363	0.543	0.612	0.391
	Std	0.021	0.013	0.086	0.058	0.080	0.041	0.082	0.057	0.006	0.030	0.023
SSC	Max	0.365	-	-	-	-	0.622	-	-	0.798	-	-
MS-2	Min	0.405	0.501	0.375	0.002	0.630	0.465	0.423	0.436	0.698	0.675	0.606
	Mean	0.431	0.537	0.386	0.002	0.630	0.465	0.432	0.437	0.698	0.719	0.641
	Max	0.487	0.580	0.399	0.003	0.630	0.466	0.509	0.437	0.698	0.735	0.671
	Std	0.022	0.038	0.009	0.000	0.000	0.000	0.027	0.000	0.000	0.023	0.023
	Fixed	0.428	0.577	0.392	0.003	0.603	0.466	0.509	0.437	0.698	0.735	0.646
ϵAMS	Min	0.445	0.531	0.392	0.002	0.617	0.472	0.479	0.485	0.785	0.689	0.601
	Mean	0.501	0.631	0.413	0.003	0.619	0.472	0.486	0.519	0.787	0.807	0.625
	Max	0.587	0.659	0.426	0.003	0.620	0.472	0.539	0.523	0.788	0.884	0.650
	Std	0.051	0.037	0.010	0.000	0.001	0.000	0.019	0.012	0.001	0.084	0.018
	Fixed	0.511	0.659	0.446	0.002	0.680	0.472	0.539	0.523	0.788	0.884	0.648
ϕAMS	Min	0.430	0.527	-	0.002	0.571	0.521	0.475	-	0.711	0.677	-
	Mean	0.499	0.587	-	0.002	0.607	0.556	0.484	-	0.712	0.685	-
	Max	0.570	0.644	-	0.003	0.633	0.565	0.501	-	0.712	0.706	-
	Std	0.063	0.042	-	0.000	0.031	0.013	0.007	-	0.000	0.008	-
	Fixed	0.564	0.655	-	0.002	0.633	0.564	0.482	-	0.711	0.695	-
kMS-	Min	0.383	0.503	-	-	-	-	-	-	0.819	0.636	-
	Mean	0.446	0.527	-	-	-	-	-	-	0.819	0.687	-
	Max	0.502	0.538	-	-	-	-	-	-	0.819	0.701	-
DSK	Std	0.040	0.012	-	-	-	-	-	-	0.000	0.024	-
	Fixed	0.502	0.520	-	-	-	-	-	-	0.819	0.702	-
kMS-	Min	0.376	0.519	-	-	-	-	-	-	-	0.634	-
	Mean	0.455	0.541	-	-	-	-	-	-	-	0.711	-
	Max	0.492	0.580	-	-	-	-	-	-	-	0.725	-
WSK	Std	0.040	0.029	-	-	-	-	-	-	-	0.029	-
	Fixed	0.516	0.579	-	-	-	-	-	-	-	0.702	-
kMS-	Min	0.444	0.520	-	-	-	-	-	-	-	0.636	-
	Mean	0.481	0.601	-	-	-	-	-	-	-	0.666	-
	Max	0.562	0.625	-	-	-	-	-	-	-	0.671	-
CIK	Std	0.033	0.039	-	-	-	-	-	-	-	0.011	-
	Fixed	0.511	0.621	-	-	-	-	-	-	-	0.703	-

Table 4.7: KC of the competing methods on the four datasets.

on the respective variant of the dataset. The rows with the standard deviations have a gray background to differentiate them from the scores.

Visualizations of the obtained segmentations can be found in Figs. 4.5 to 4.15. Because the scores show some variation across the ten trials, we show for every method the segmentation that corresponds to the seed that achieved the fifth best OA score. The OA scores of these segmentations are given in parentheses in the figure captions. The segmentation methods provide segmentations for the full images. However, since the ground truths contain unlabeled pixels, we mask these pixels also in the shown segmentations for easier comparison. The unlabeled pixels are shown in black.

We start with some general observations. When considering the scores in Tables 4.5 to 4.7, we can generally say that our proposed segmentation framework with one of the two anisotropic MS-based models ϵ AMS and ϕ AMS, that take into account first- and second-order statistics, and including the preprocessing by the MNF yields excellent results across all datasets. In particular, this shows that these two models, which were explicitly designed to tackle the spectral variability of hyperspectral data, are indeed able to deal with this phenomenon effectively. We can furthermore observe that the application of the MNF has a significant effect on the segmentation result in many cases. Both models ϵ AMS and ϕ AMS in combination with the MNF are not much affected by spectral outliers. The clustering methods k -means, GMM and BGM have the biggest problems achieving useful segmentations, which is what was expected since the three methods rely only on spectral information. Moreover, GMM as well as BGM often achieve slightly better scores than k -means since k -means estimates only the means of the spectral clusters, while the other two methods also estimate the covariance and are, therefore, more flexible. Surprisingly, although explicitly proposed for unsupervised hyperspectral image segmentation and informed about spectral and spatial information, the scores of 3D-CAE are significantly worse than the results of the other methods that are designed for hyperspectral imaging. The method is only able to compete with the clustering algorithms. Our proposed anisotropic MS models ϵ AMS and ϕ AMS show, in total, the best performance of all methods. On most datasets and their variants, they achieve either the highest mean or the highest maximum score and beat the clustering methods as well as the state-of-the-art methods significantly. The kernel methods also show their potential and achieve good scores on the datasets they were tested on. However, the clustering methods seem to tend to have a slightly lower standard deviation in their scores and, therefore, a little bit more stable segmentation results than the other algorithms. Additionally, the visualizations of the segmentations in Figs. 4.5 to 4.15 show that the MS-based methods yield homogeneous segmentations due to the regularization of the perimeter and are even able to separate entangled spectral clusters. Since all MS-based models (except for NLTV) use the k -means result as the initial segmentation, we would expect them to always beat the k -means score. Surprisingly, ϕ AMS and the kMS variants

have problems doing so on SA after MNF. A simple reason could be that we have not found hyperparameter choices that lead to good results.

Indian Pines

We now take a closer look on the segmentation results of the competing methods on IP. The dataset is a highly challenging one. Many of the segments are spectrally

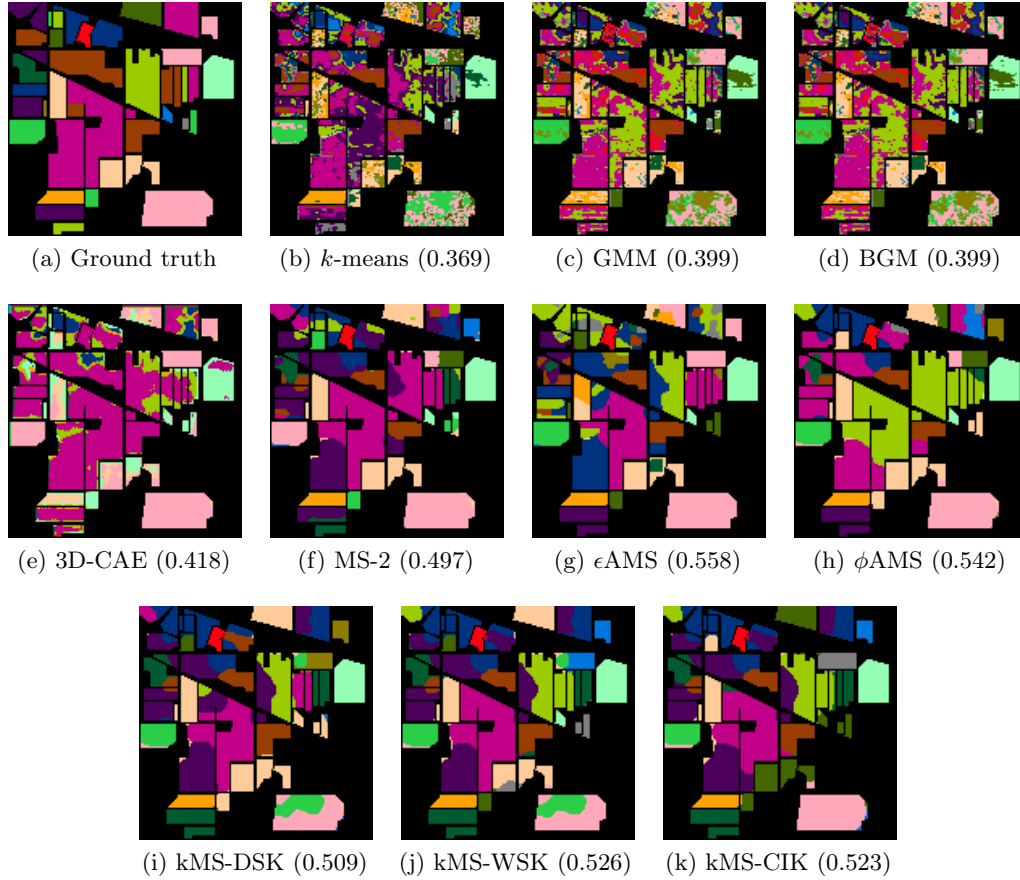


Figure 4.5: Ground truth and segmentations of raw IP data produced by the competing methods.

similar and their spectral clusters overlap. Some spectral clusters of segments given by the ground truth consist of multiple modes. Considering the scores, we see that ϵ AMS and ϕ AMS beat the other methods significantly. In particular, their mean OA and KC scores are even better than the maximum scores of the other methods (except

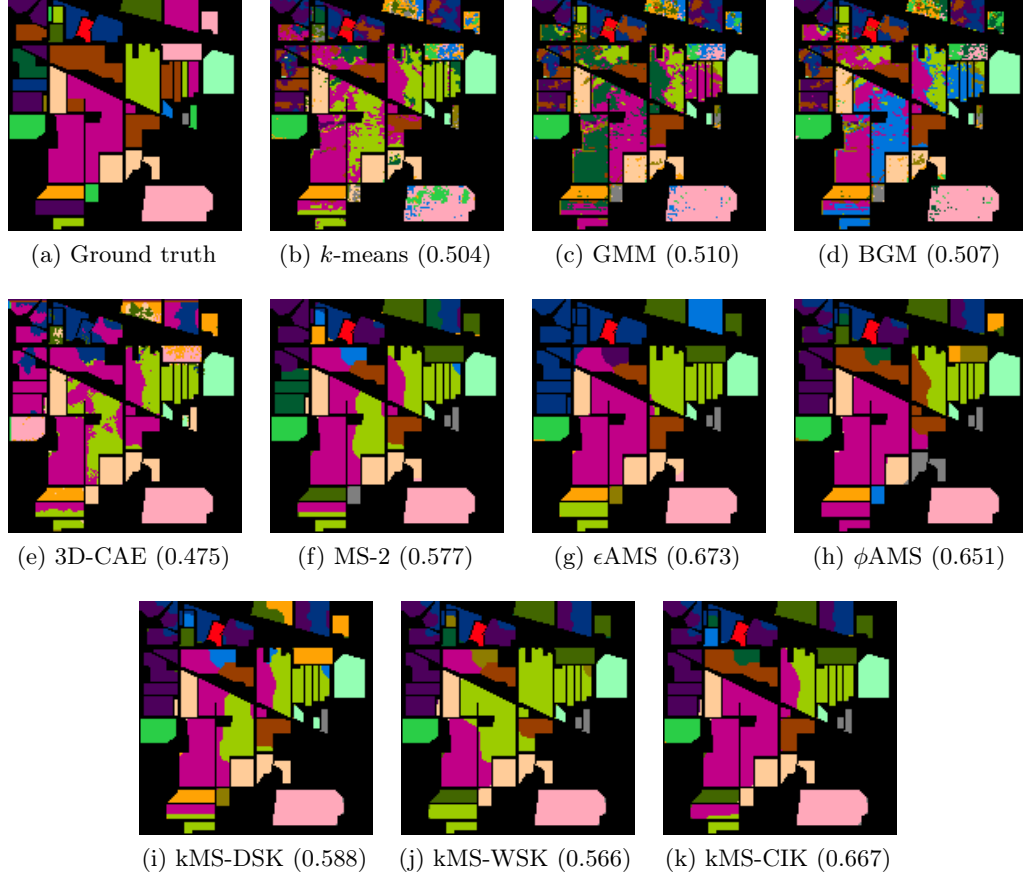


Figure 4.6: Ground truth and segmentations of MNF-transformed IP data produced by the competing methods.

for kMS).

On raw IP, the anisotropic MS models ϵ AMS and ϕ AMS have the highest standard deviations in the OA scores. The kernel methods also achieved good results, while kMS-CIK is the best of the kernel methods in terms of OA. Its maximum score ranges between the mean and maximum scores of ϵ AMS and ϕ AMS. There is hardly a difference between the OA scores of the two other variants of kMS, namely DSK and WSK. When we compare the AAs of the methods, we see that the kMS variants perform slightly better than the anisotropic MS models. This indicates that they could detect small segments more accurately. The isotropic MS model MS-2 performed significantly worse than its anisotropic counterparts ϵ AMS and ϕ AMS. Considering the segmentations in Fig. 4.5, it can be seen that the clustering methods produced

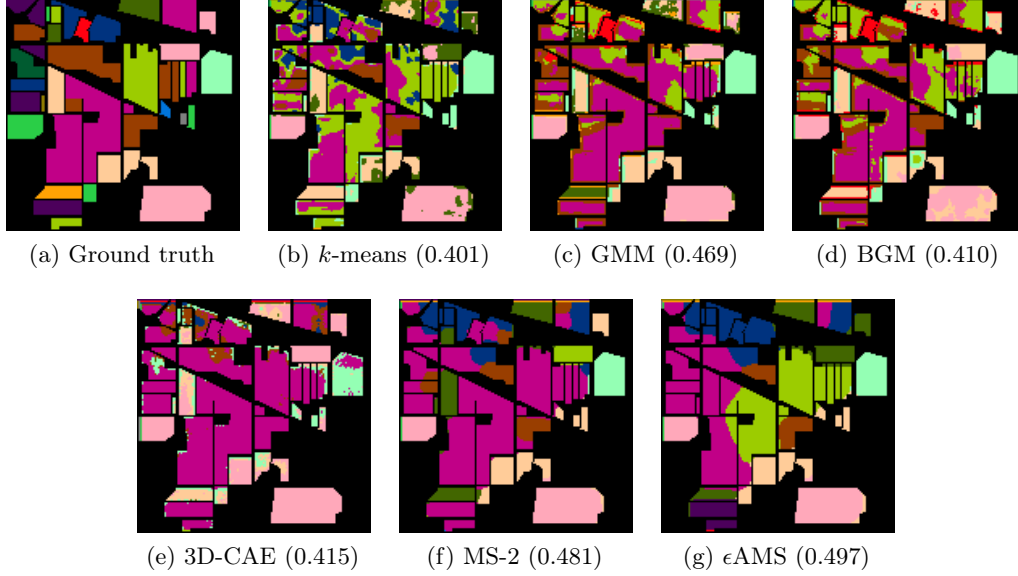


Figure 4.7: Ground truth and segmentations of the learned FL features of the IP data produced by the competing methods.

extremely inhomogeneous segmentations, which means that they had problems to handle the overlapping spectral clusters that correspond to the mixed segments. The deep learning model 3D-CAE is not able to recognize the structure of the segments, either. However, all MS-based methods (except for NLTV) are able to correctly classify some bigger segments.

On IP after an application of the MNF, we observe a significant performance boost in terms of OA for all methods. The largest leap in the OA scores made k -means with around 13 percentage points. Also here, ϵ AMS and ϕ AMS achieve significantly higher OA scores than the other methods, but k MS-CIK gets close. The standard deviations of ϵ AMS and ϕ AMS with respect to OA are lower on the MNF variant of IP than on the raw data. The biggest performance boost through the MNF in terms of AA happened for GMM and BGM. Both methods were better able to correctly classify small segments. This indicates a potential disentanglement of spectral clusters through the MNF because the clustering methods, that rely purely on the spectra, could improve their results. The fact that GMM and BGM could boost their AA scores much more than k -means hints that the spectral variability in the spectral clusters is still present and the reason why the isotropic description of the spectral clusters by k -means could not improve the result as much as an anisotropic description in the case of GMM and BGM can. Surprisingly, the kernel-based methods have

bigger problems classifying small segments than on the raw data, which results in lower AA scores. Considering the visualizations of the segmentations in Fig. 4.6, we see that the clustering methods still yield inhomogeneous segmentations and also the problems of 3D-CAE to identify the segments remain. The results of our MS-based models are quite good, in particular, the ones of ϵ AMS, ϕ AMS and kMS-CIK.

From the OA scores of the methods applied to IP after FL, we learn that k -means and GMM benefit from FL, while the OA scores of BGM are in the same order as on the raw IP data. Interestingly, the remaining methods deliver worse results than on the raw data or with MNF preprocessing. However, the variation in the OAs of the MS models is lower than before. The AA scores of all methods are much worse compared to the AA scores on the raw and MNF data. A possible reason could be that FL makes the spectra of small segments spectrally more similar to the spectra of bigger segments. This could be a consequence of the feature learning model that processes patches of pixels as input, possibly resulting in a smoothing of the spectral information across segment boundaries and, therefore, the loss of the sharp difference between spectra of different segments. Hence, spectral clusters of segments are smeared out towards other spectral clusters, which makes it particularly difficult to detect small segments. From the KC numbers, we learn that many of the correct pixel classifications in the case of the clustering methods probably happened by chance since the scores are quite low. This observation is supported by the visualizations of the segmentations in Fig. 4.7. We can see that the pink segment is a large segment in the ground truth and that many methods have put many pixels into that segment.

Kennedy Space Center

In the case of KSC, we have to have in mind that only extremely few labeled pixels are available in the ground truth and, hence, the scores and visualizations of the segmentations have a limited expressiveness. The performance on the raw KSC is bad for all methods. The best OA score is achieved by NLTV. However, its score of 0.415 still means that less than every second pixel could be correctly classified and we, therefore, cannot consider this result as a good one. From the low KC scores, we see that there is a considerably high agreement with the ground truth by chance. All variants of the MS model (except for NLTV) base their initialization on the k -means result. Since already the k -means segmentation is essentially useless, the initial guesses of the model parameters derived from that k -means segmentation are also of particularly low quality. The algorithms are not able to recover from that and find a local minimum that corresponds to a useful and meaningful segmentation due to the high non-convexity of the objective functions. The visualizations of the segmentations in Fig. 4.8 confirm the low quality of the segmentations. We can see that every method essentially puts all labeled pixels into only one segment.

If we apply the MNF to preprocess the data, the situation changes dramatically.

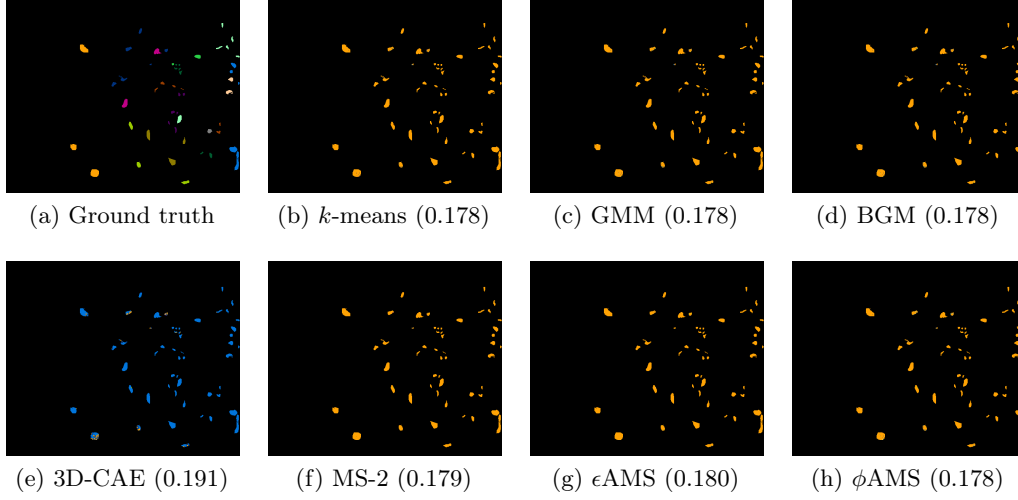


Figure 4.8: Ground truth and segmentations of raw KSC data produced by the competing methods.

3D-CAE benefits less from this step but is also able to improve its result compared to the result on the raw data. In particular, the standard deviation in the scores of 3D-CAE is rather high. The performance boost for k -means also has a huge impact on the MS variants that use the k -means outcome as their initialization. While MS-2 achieves the highest mean score in terms of OA, the highest maximum score is produced by ϕ AMS. Please note that this maximum score is even beaten by ϵ AMS in the run with the fixed seed, meaning that this model was able to yield the best description of the data, but when started with the ten randomly chosen seeds, it converged to local minima that correspond to segmentations of lower quality. The best mean score in terms of AA was achieved by MS-2, but the scores of ϵ AMS and ϕ AMS are very close to it. The same holds essentially for the KC scores. In terms of AA and KC, the best maximum scores were obtained from ϕ AMS. The visualizations of the segmentations in Fig. 4.9 for this case reflect these findings.

Pavia University

In contrast to the other three test datasets that were captured with the AVIRIS sensor, PU was acquired with a ROSIS sensor. Therefore, the dataset has a special role in the experiments. SSC showed the best performance in terms of maximum OA and KC. The application of the MNF as well as FL boosts the performance measured with OA only of some methods, but not enough to beat SSC. This suggests that this data type is better handled by a graph-based spectral clustering algorithm.

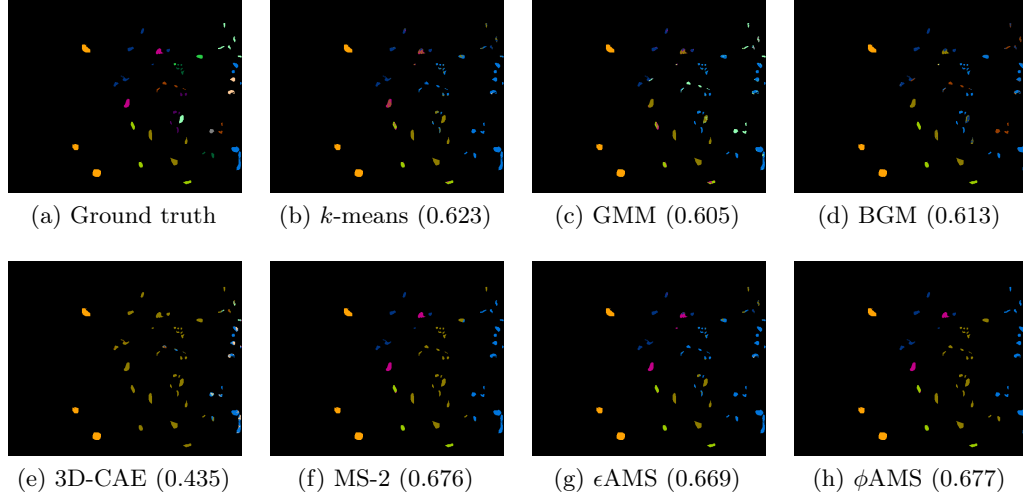


Figure 4.9: Ground truth and segmentations of MNF-transformed KSC data produced by the competing methods.

Nevertheless, also ϵ AMS on the FL variant and ϕ AMS on the raw data perform well in terms of OA. Their maximum scores are about 4.6 percentage points and 4 percentage points, respectively, lower than the score by SSC. A reason why ϕ AMS beats ϵ AMS significantly on the raw data, even though the models are particularly similar, could be that the fixed lower bound for the eigenvalues of the covariance matrices defined by ϵ in the case of ϵ AMS was chosen too high and, therefore, local minima that correspond to a better description of the data were excluded, while ϕ AMS allowed smaller estimates for the eigenvalues. We can further observe that ϵ AMS on FL PU and ϕ AMS on raw PU achieve quite similar mean OA scores (0.633 and 0.635). However, the AA scores tell us that ϵ AMS on FL PU performs much worse than ϕ AMS on raw PU (0.491 and 0.713). We can deduce that ϵ AMS on FL PU has problems detecting small segments and providing a detailed segmentation, while ϕ AMS on raw PU is able to do so. This could be again due to the fact that the feature learning method not only considers the separate spectra but for each pixel a patch of pixels and concentrates the information in the patch in one vector of learned features. Of course, there are patches that cover pixels of different segments and, therefore, smear out the spectral clusters of the segments. This observation is also confirmed by the visualizations of the segmentations in Figs. 4.10 to 4.12. If we consider in the ground truth (cf. Fig. 4.3b), for example, the small structures (marked with a dark green color) next to the streets (narrow lines marked in blue), we see that these structures are put into the blue segment in the case of ϵ AMS on

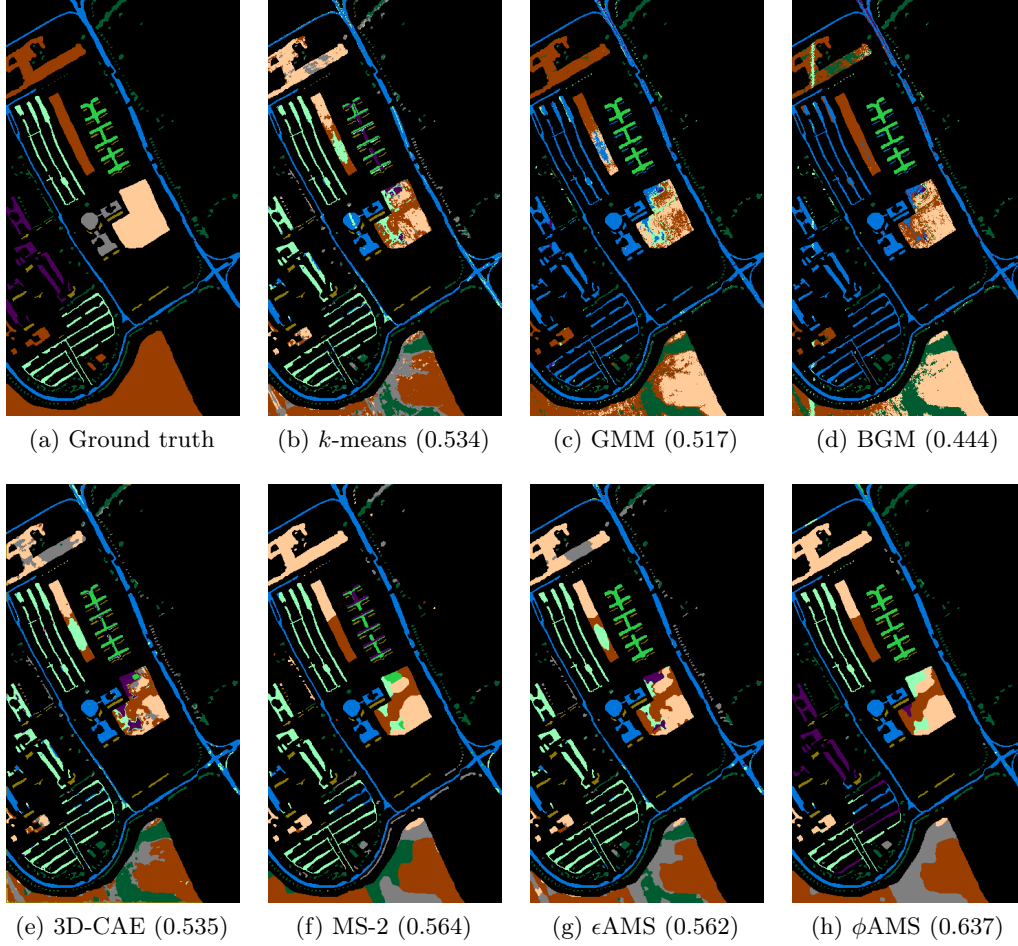


Figure 4.10: Ground truth and segmentations of raw PU data produced by the competing methods.

FL PU (cf. Fig. 4.12g), while ϕ AMS on raw PU was able to correctly classify them (cf. Fig. 4.10h). Furthermore, ϕ AMS could also detect the purple segment on raw PU well. Something similar can be observed when considering the AA scores of the other methods. In every case, except for 3D-CAE and ϕ AMS, the MNF boosts the AAs. Hence, also here, the MNF shows a clear positive effect and helps to detect small structures. Unfortunately, we do not have access to the AA score of SSC on PU to obtain an idea if the model is able to detect small structures.

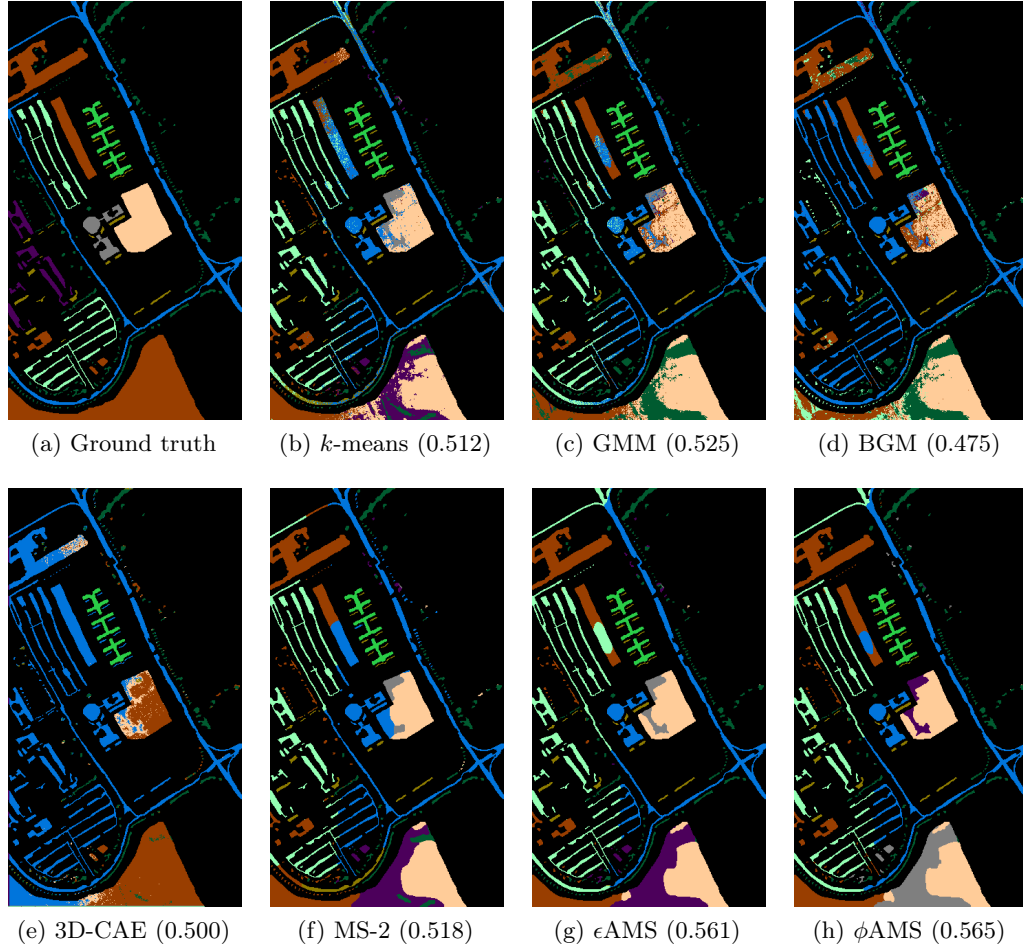


Figure 4.11: Ground truth and segmentations of MNF-transformed PU data produced by the competing methods.

Salinas

In the case of SA, there are also some complicated challenges to solve since also here, some segments that are given by the ground truth correspond to spectral clusters of multiple modes or their spectral clusters overlap. On the raw SA dataset, kMS-DSK yields the segmentation of the highest quality in terms of OA and beats SSC by two percentage points. The standard deviation of the OAs of the methods is rather low. The OA scores of the clustering methods are similar and lie around 0.68. Our model ϵ AMS also achieves a quite good OA score that is close to the one from SSC. Although

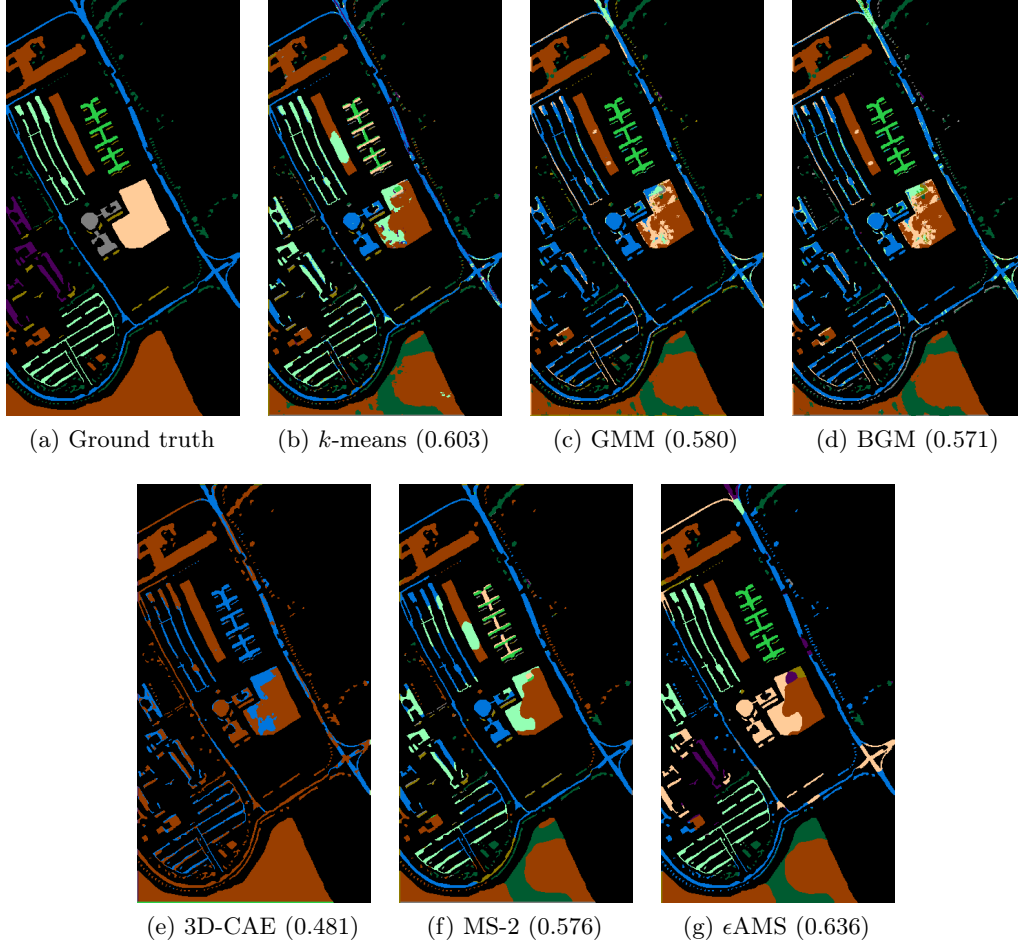


Figure 4.12: Ground truth and segmentations of the learned FL features of the PU data produced by the competing methods.

ϕ AMS is a model that is highly similar to ϵ AMS, it surprisingly yields much lower scores with respect to OA that are only marginally higher than the ones of MS-2. If we consider the AA scores, however, ϵ AMS and ϕ AMS perform quite similarly on the raw SA dataset. This could be again due to the different regularization techniques of the covariance matrices. Our choice for ϕ used in the experiments (cf. Table 4.2) allows usually much lower eigenvalues than the concrete choices for ϵ (cf. Table 4.4). Hence, local minima could be excluded by the choice of ϵ also here that could correspond to segmentations with lower AA. However, on the raw SA data, again kMS-DSK yields the best segmentation in terms of AA, but also MS-2 achieves a good score that

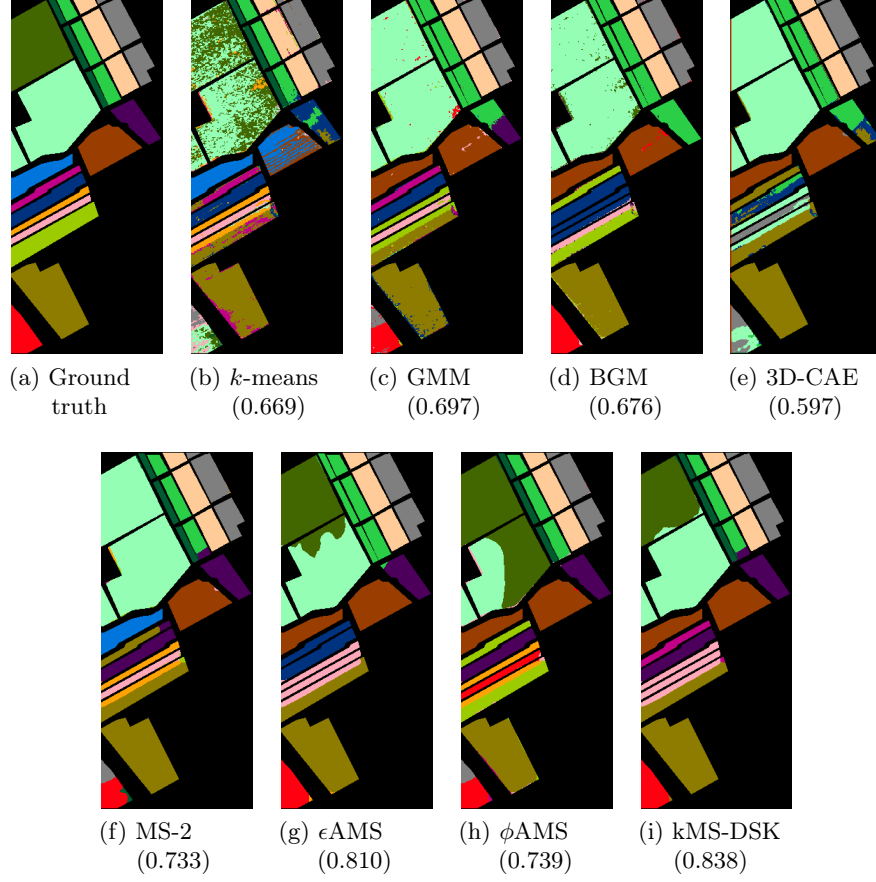


Figure 4.13: Ground truth and segmentations of raw SA data produced by the competing methods.

is higher than the one for ϵ AMS and ϕ AMS. Notably, also k -means has a slightly better AA score than ϵ AMS and ϕ AMS. Hence, the anisotropic MS models did not detect small structures and segments well. However, one has to have in mind that the parameters of the models were tuned to give high OA scores because the remote sensing community mainly focuses on this score, and it is, therefore, likely that other parameter choices would be able to beat k -means in terms of AA. In the case of the raw SA data, the OA benefited most from a strongly weighted perimeter by a high value of λ in the MS functional that put many pixels into big segments. When we inspect the visualizations of the segmentations of the raw SA data in Fig. 4.13, we see that kMS-DSK almost perfectly segmented the olive green and mint green segments in the upper left part of the image (cf. Fig. 4.13i). These two classes correspond

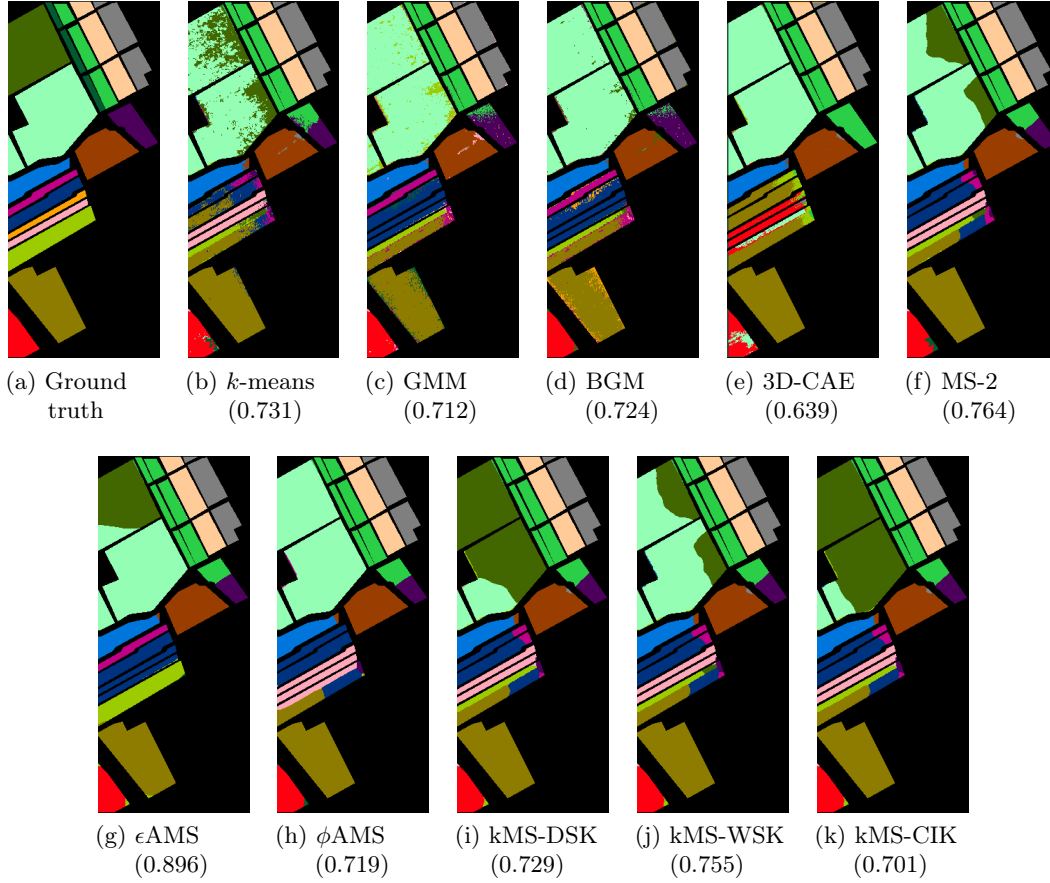


Figure 4.14: Ground truth and segmentations of MNF-transformed SA data produced by the competing methods.

to *vineyard untrained* and *grapes untrained* and are spectrally extremely similar. Their spectral clusters are completely mixed and overlapping. Therefore, most of the methods could not distinguish the two segments and classified the corresponding pixels as being from the same segment. The model ϵ AMS could also separate these two segments to some extent (cf. Fig. 4.13g). An advantage that kMS-DSK has over the other methods is that it also uses contextual information in its indicator function. This could be one reason why it was able to almost perfectly classify the two segments. The model is also able to identify the narrow, long, dark green segment to the right of the olive and mint green segments. Many other methods also had problems with this segment. The good OA and AA scores, together with the visual inspection of the segmentations, show the great potential of the kernelized MS model.

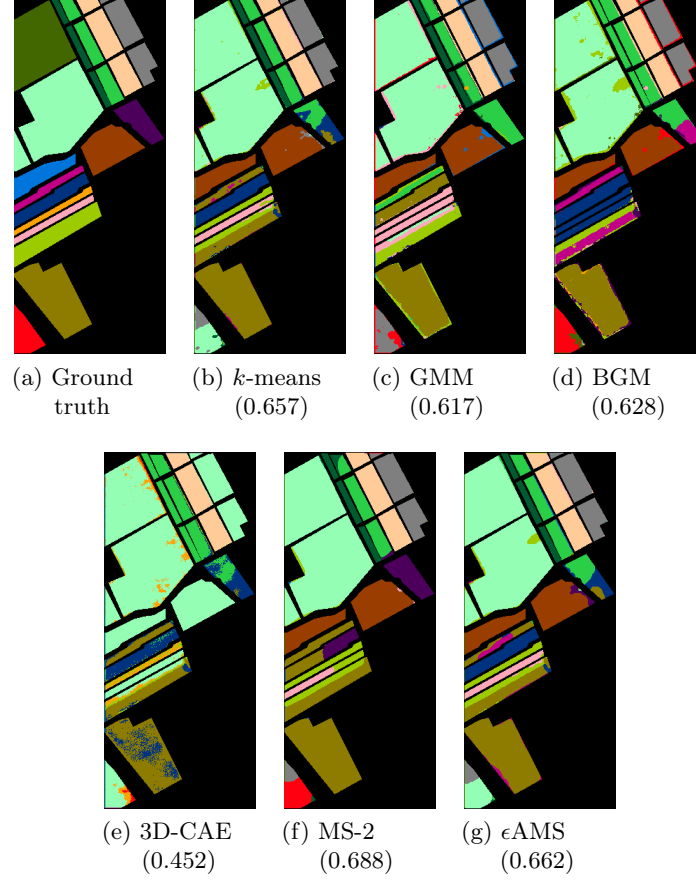


Figure 4.15: Ground truth and segmentations of the learned FL features of the SA data produced by the competing methods.

In the case of the MNF-transformed SA data, we observe an increase in the OA scores compared to the results on the raw data for all methods except for ϕ AMS and kMS-DSK. However, also the standard deviations increased in all cases, in particular, for ϵ AMS that shows a noticeably high fluctuation in the scores of the ten trials. Nevertheless, ϵ AMS could find a remarkable description of the data that achieves the highest maximum OA score and also beats the scores of SSC and kMS-DSK on the raw data. When we consider the AA scores on SA after MNF, we see that ϵ AMS was also able to achieve an excellent score in this category and beat all its competitors. Notably, kMS-DSK obtained an even better AA score on the raw data than ϵ AMS on the MNF-transformed data when comparing the mean AA scores. However, the other anisotropic MS model ϕ AMS did not benefit from the MNF in terms of AA and got

slightly worse scores. The kernelized methods had problems finding segmentations with a quality that is high enough to compete with the segmentations of the other methods. Equipped with the kernels DSK and CIK, the results are even slightly worse than the results of k -means with respect to OA, although the k -means output is used as the initialization. In terms of AA, however, all kMS variants beat k -means, even though the parameters were chosen to maximize the OA. Considering the visualizations of the segmentations of the MNF-transformed SA data in Fig. 4.14, we observe that only ϵ AMS was able to come close to the correct separation of the olive green and mint green segment in the upper left part (cf. Fig. 4.14g). The other MS-based methods basically classify these two segments either as the mint green or olive green segment. None of the methods was able to detect the narrow, long, dark green segment to the right of the olive and mint green segments. This could indicate that the MNF considers the difference of the spectra of this segment and its neighboring segment to the upper right as noise and, therefore, reduces this difference in the spectra, making it hard for the methods to correctly classify the corresponding pixels.

The application of FL to SA lets the OA and AA scores decrease compared to the results on the raw and MNF-transformed SA data. While MS-2 achieves the highest mean OA score and slightly beats ϵ AMS in terms of OA, ϵ AMS beats MS-2 in terms of mean AA score. The segmentation of highest mean AA score was produced by k -means. The score is slightly higher than the one achieved by ϵ AMS. The visualizations of the segmentations in Fig. 4.15 also show segmentations of lower quality. When having a closer look, we see that none of the methods is able to separate the olive green and mint green segment.

The KC scores essentially reflect the OA scores, telling us that the agreement by chance is negligible.

Summary

In summary, ϵ AMS and ϕ AMS showed the best performances across the four datasets. Their OA, AA and KC scores were consistently among the highest scores and the visualizations of the segmentations showed segmentations of consistently high quality. They could beat all their competitors in terms of segmentation quality, including current state-of-the-art methods for unsupervised hyperspectral image segmentation, and showed their high potential. The lowest scores are mainly achieved by the clustering methods and 3D-CAE. The kernelized MS model also yields segmentations of good quality. In particular, on the raw SA dataset, kMS-DSK demonstrates the great potential of the kernelized MS model. However, it remains a hard challenge to find hyperparameters with which the method is able to converge to good local minima without empty segments, that cause the optimization procedure to stop in its current implementation. In some cases, it even happened that the grid search after the Optuna hyperparameter optimization returned hyperparameters that led to

worse results than the hyperparameters from Optuna. This problem also prevented us from finding suitable parameters for kMS with WSK and CIK on the raw SA dataset. Generally speaking, an application of the MNF improves the results. In contrast to that, the feature learning method leads to rather worse results and, in particular, prevents the models from correctly detecting small structures.

4.2.2 Effect of MNF

We have already seen that an application of the MNF as a preprocessing has an effect on the behavior and results of the methods. In most cases, we obtained segmentations that achieved significantly higher accuracies than the segmentations obtained from the raw data. In the case of the KSC dataset, the MNF actually enabled the methods to provide useful segmentations. On PU, we have observed that the MNF leads in many cases to lower OA scores compared to the results on the raw data (cf. Table 4.5). However, the AA scores increased after an application of the MNF (cf. Table 4.6). This is an indicator that the MNF leads to a disentanglement of the spectral clusters, which makes it easier for the segmentation methods to also correctly detect smaller spectral clusters that correspond to small structures and segments in the image. The MNF also helped to make the results of the models less sensitive to different hyperparameter choices. It was much easier to find suitable parameters after an application of the MNF. The MNF enabled us to find useful hyperparameters for every method. However, one downside of the MNF is that the standard deviation of the scores seems to increase. This could be caused by the used noise estimation in the MNF (cf. Section 1.5). When we estimate the noise in the data, we consider the difference of the spectra of neighboring pixels without the information of segment boundaries, that we do not have at that moment. Hence, also the difference of spectra from pixels of different segments that lie at the segment boundaries is considered to some extent as noise, which leads to an overestimation of the noise covariance. Consequently, the MNF has a stronger flattening effect on the data than wanted and, in particular, also reduces the difference of spectra at the segment boundaries. In terms of the spectra, this results in some pixels whose spectra are dragged into the regions between the spectral clusters of different segments. These pixels are potentially misclassified, which in turn leads to worse estimates of the model parameters and, therefore, to worse segmentation results. Of course, also the initialization by k -means is affected in that way and passes on the effect to the models.

We also carried out a visual inspection of the data, in which we analyzed the data before and after the application of the MNF. To this end, we used the PCA and projected the four datasets before and after the MNF onto their first three PCs to compare the spectral clusters of the segments given by the ground truth visually. The subjective impression suggests that the MNF brings the spectral clusters

closer to a Gaussian distribution. This would mean that the description of the spectral clusters with mean and covariance as in our anisotropic indicator functions $f_{\eta,l}$ (cf. Section 2.2.1) and $f_{\phi,l}$ (cf. Section 2.3.1) of the models ϵ AMS and ϕ AMS, respectively, becomes more suitable by the MNF. Henze and Zirkler [55] described a test statistic that can be used to test whether the distribution of data samples is approximately a multivariate Gaussian distribution. The test uses the fact that a probability distribution is uniquely determined by its so-called *characteristic function* (not to be confused with the characteristic function χ_A of a set A that we used in this thesis). It computes the difference between the empirical characteristic function of the data samples and the theoretical characteristic function of a multivariate Gaussian distribution. The smaller this difference, the closer is the data distribution to a multivariate Gaussian distribution. We applied this test to quantify the subjective perception from the visual inspection. To this end, we considered for each dataset each segment that is given by the corresponding ground truth and applied the test to the distribution of the segment's spectra before and after the application of the MNF. In this way, we obtained for each segment two numbers; one quantifies how close the raw data distribution is to the multivariate Gaussian and the other one how close the MNF data distribution of the respective segment is to the multivariate Gaussian. To ensure that both numbers are comparable, the raw data had to have the same dimension as the MNF data. We achieved this by applying the PCA to the raw data and keeping the same number of components as we did with the MNF (cf. Table 4.4). The results are plotted in Fig. 4.16. Each of the Figs. 4.16a to 4.16d has on the x -axis the segment number and on the y -axis the score from the Henze–Zirkler test. The two curves correspond to the raw and the MNF data. Please note that the values of the segments 10, 12 and 13 for the raw KSC data (cf. Fig. 4.16b) are not visible because we had to cut the value range to small numbers to be able to compare the test statistic of the other segments. The values of the segments 10, 12 and 13 are much higher. We observe that merely in few cases, the raw data is closer to the multivariate Gaussian distribution than the MNF data. Consequently, the test results coincide with the subjective perception and support our conjecture that the application of the MNF is particularly beneficial for our anisotropic MS models ϵ AMS and ϕ AMS.

4.2.3 Convergence of alternating optimization

To reliably obtain useful results from a model, it is crucial that the applied optimization algorithm converges. We have not proven a convergence guarantee for our alternating optimization approach (cf. Section 3.1.3) but collected numerical indications that the optimization procedure applied to our developed segmentation models indeed converges. To this end, we plotted the values of the MS segmentation functional (cf.

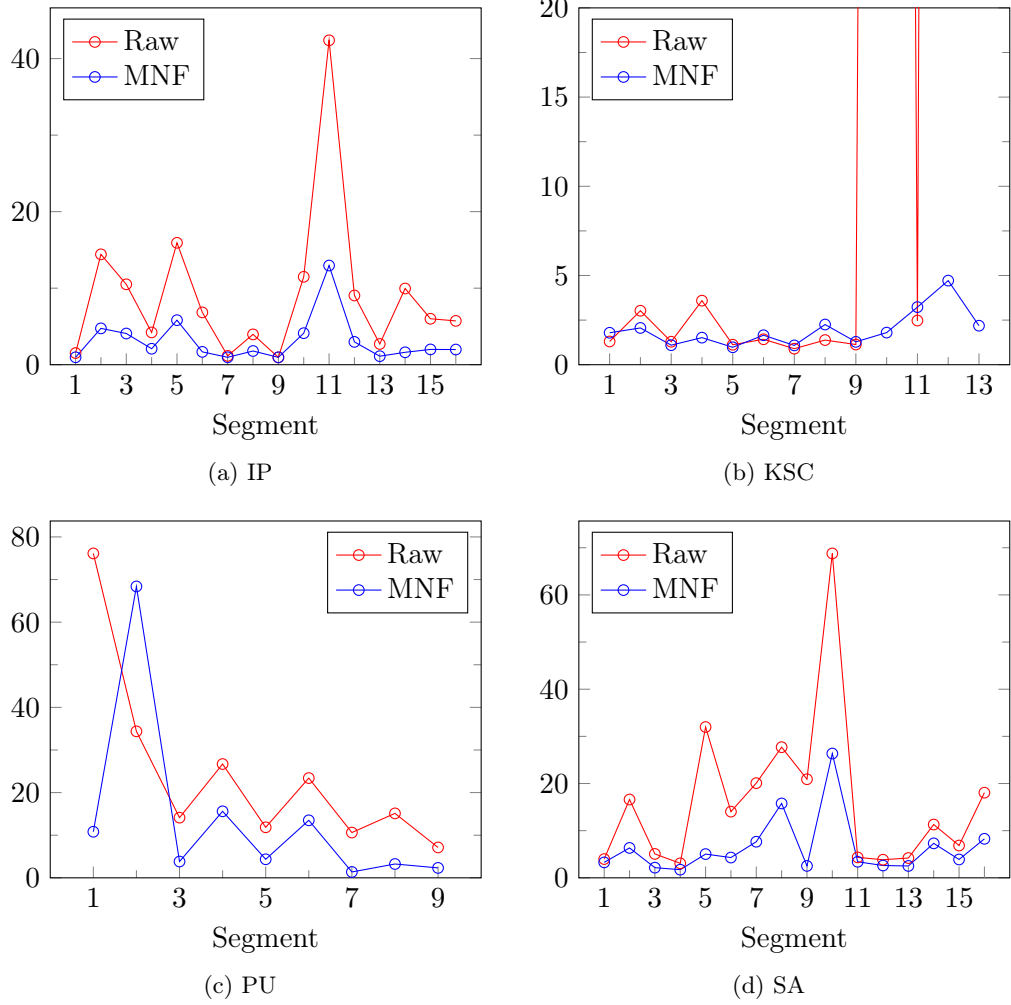


Figure 4.16: Henze-Zirkler statistics for each dataset and each segment.

(2.9)) with the corresponding indicator function, the values of the objective functional and the values of the stopping criterion against the outer iterations. The plots for ϵ AMS, ϕ AMS and kMS with DSK, WSK and CIK on the raw IP dataset can be found in Figs. 4.17 to 4.21. They indicate a converging behavior. Please note that from the plots, one can also see that most of the methods did not need the full 20 outer iterations to stop, which we set as the maximum number of outer iterations. We can further observe that the optimization approach that we use to optimize our methods leads to a rapid decay of the plotted values such that the values are not changing

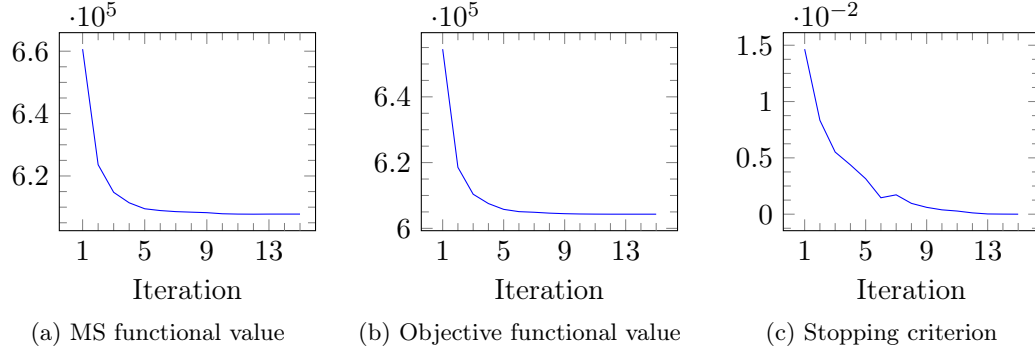


Figure 4.17: Plots of functional values and stopping criterion against outer iterations of ϵ AMS on raw IP.

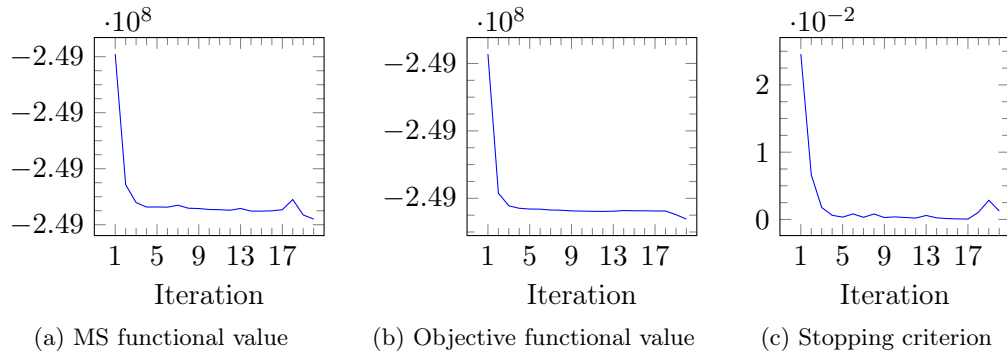


Figure 4.18: Plots of functional values and stopping criterion against outer iterations of ϕ AMS on raw IP.

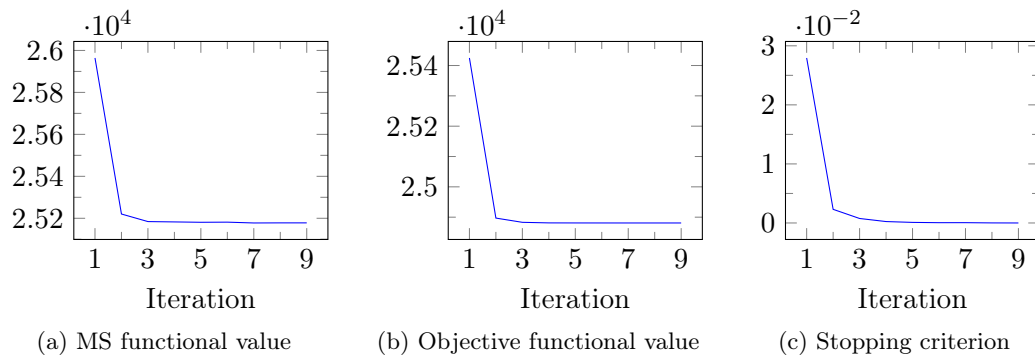


Figure 4.19: Plots of functional values and stopping criterion against outer iterations of kMS-DSK on raw IP.

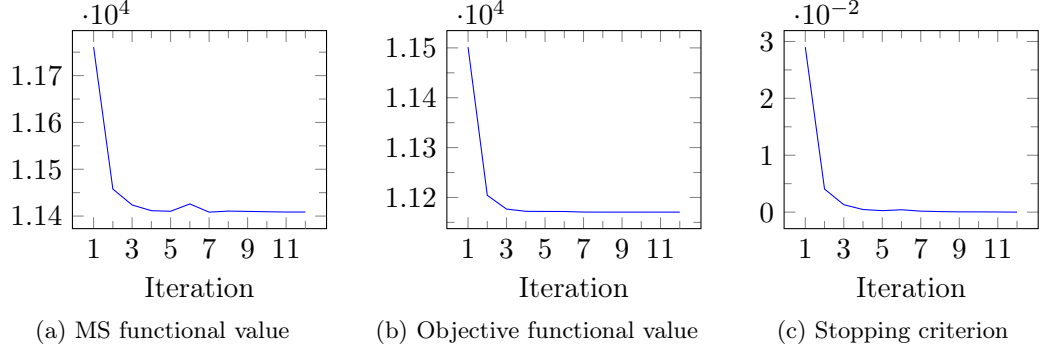


Figure 4.20: Plots of functional values and stopping criterion against outer iterations of kMS-WSK on raw IP.

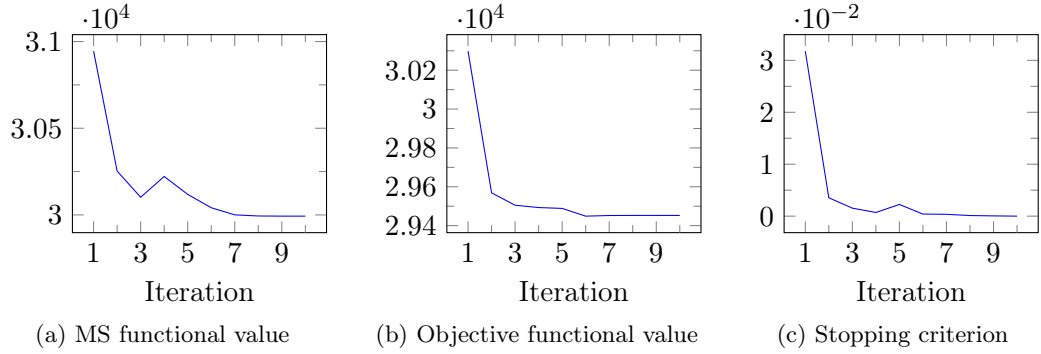


Figure 4.21: Plots of functional values and stopping criterion against outer iterations of kMS-CIK on raw IP.

significantly anymore in the last iterations. Even though the functional values and stopping criterion show a decreasing behavior also in the case of the other datasets, the converging behavior is in a few cases not as clearly visible as in the shown plots for raw IP. Further investigation is needed in this regard. Nevertheless, the results are an indicator that the optimization procedures show a converging behavior.

4.2.4 Effects of early stopping in PDHG algorithm

The experience shows that the PDHG algorithm needs many iterations to converge. We, therefore, set the maximum number of PDHG iterations in our experiments to 1000 (cf. Tables 4.1 to 4.3). However, in our experiments, we observed that these iterations were almost always needed and that the PDHG method usually

stopped iterating because it reached its maximum number of iterations. Moreover, Proposition 3.14 gives the condition $\tau\kappa L_0^2 < 1$ on the step sizes τ and κ to have guaranteed convergence of the PDHG algorithm. From Remark 3.20, we know that L_0^2 is slightly below 8, resulting in reasonable choices $\tau = \kappa = 1/\sqrt{8}$. In our experiments, however, we accidentally used step sizes $\tau = \kappa = 1/8$ (cf. Tables 4.1 to 4.3), which means that we used smaller step sizes than it would have been possible to choose them. Hence, we applied the algorithm with a limited number of iterations and also with step sizes τ and κ that are smaller than the optimal choices. This can be understood as an early stopping in the PDHG application, which can influence the path to and also the resulting local minimum of the objective function that we find with our alternating optimization.

To evaluate the effect of the early stopping, we reran the experiments carried out with ϵ AMS on the raw and MNF-transformed IP dataset with the same choices for λ and ϵ (cf. Table 4.4) but with larger step sizes ($\tau = \kappa = 1/\sqrt{8}$) and a higher maximum number of iterations (maximum 100 000 PDHG iterations), that enables the PDHG method to always terminate due to the stopping criterion, and compared the results. We could observe that the algorithm converges indeed to a local minimum with a lower functional value than with early stopping but this minimum corresponds to a worse segmentation in terms of OA than the one obtained with early stopping. However, the choices for the model parameters λ and ϵ we found with early stopping are not optimal when we run the segmentation without early stopping. We, therefore, made another Optuna study as described in Section 4.1.3 in which we determined parameter choices leading to good segmentation results with respect to OA without early stopping. These segmentations achieve OA and AA scores in the same range as the segmentations achieved with early stopping. The segmentations are of the same quality. We, therefore, decided not to change our setting and rerun all experiments with larger step sizes and more PDHG iterations but to keep the maximum number of PDHG iterations of 1000 and the smaller step sizes since the early stopping leads to comparable results.

4.3 Derivation of complex sea ice states from multispectral images

As the last part of the thesis and before concluding it, we will present the results of tests of ϵ AMS on multispectral Sentinel-2 data to show that the method is able to provide useful segmentations of high quality in a broad range of applications and is not restricted to hyperspectral data from AVIRIS or ROSIS sensors.

The multispectral sensors of the Sentinel-2 satellites deliver multispectral images of the Earth surface. A brief introduction of these sensors and the produced multispectral images is given in Section 1.1. In particular, images taken of the Arctic with the

multispectral Sentinel-2 sensors are used to monitor and analyze the extent of Arctic sea ice cover over the year. There are different states in which sea ice can appear, like new (young) or thin ice and old or thick ice. Furthermore, brash ice, crushed ice and slurry mixture states of ice can also occur. We used two Sentinel-2 images taken over the Arctic area to investigate whether different sea ice states reflect their own spectral signatures and if we can derive these sea ice states from Sentinel-2 images. To this end, we applied ϵ AMS to these two images and manually analyzed the resulting segmentations and inferred mean spectra of the segments.

4.3.1 Experimental setup

The two chosen datasets cover freezing conditions during winter and melting conditions during summer. They are described in the following²:

- (D1) *Winter conditions*: taken over the Barents Sea, Victoria Island, on April 4, 2018. Since we have freezing conditions, we expect no melting.
- (D2) *Summer conditions*: taken over the Fram Strait between Greenland and Svalbard on June 20, 2021. In the summer conditions, we expect melt ponds.

In Fig. 4.22, the two images are depicted as RGB images. In contrast to the false-color images we used to show the hyperspectral images, these images are true RGB images since the multispectral Sentinel-2 sensor samples the spectrum of visible light as normal photo cameras and, therefore, the data contains channels for the intensities of red, green and blue light. Please note that the contrast was enhanced by histogram equalization. The bright regions are mainly ice cover, while the dark regions correspond to water. However, the ice cover is not only ice floes but also land that is covered by ice. For example, in the lower mid part of D1 in Fig. 4.22a, there is an island that is covered by ice. Clouds can affect the reflected spectra, too. D1 and D2 are not free of clouds. For instance, in the upper right corner of D2 in Fig. 4.22c, where the image is slightly blurry, a cloud is present. Therefore, to evaluate the effect of clouds on the spectra and the derivation of the sea ice states, we also applied ϵ AMS to a crop of D1 that is free of clouds and can be seen in Fig. 4.22b. The crop was manually selected by an expert.

²The datasets are publicly available in the *Copernicus Data Space Ecosystem* web-portal, which is accessible under <https://dataspace.copernicus.eu/>. The file names of the datasets are S2B_MSIL1C_20180404T114639_N0206_R023_T37XDK_20180404T121607 (D1) and S2B_MSIL1C_20210620T133729_N0300_R067_T29XNG_20210620T154321 (D2). The Sentinel-2 mission divides the Earth surface into tiles and the file names allow for a unique identification of the recorded tiles and give information about the conditions. E.g., D1 is from the Sentinel-2B mission (S2B) and of Level-1C type (MSIL1C). It was recorded on April 4, 2018 at 11:46:39 AM (20180404T114639). The code N0206 tells us that it was processed with Baseline 02.06. The image was taken while the satellite was on the relative orbit 023 (R023) over tile 37XDK (T37XDK).

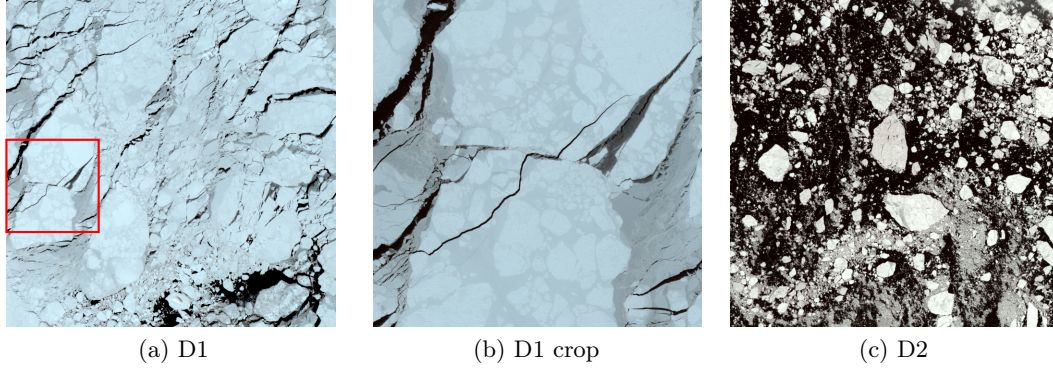


Figure 4.22: RGB versions of the two considered Sentinel-2 images and the D1 crop (region of D1 covered by the crop marked in D1 with a red box).

The Sentinel-2 multispectral sensor records 13 channels at three different spatial resolutions: 10 m, 20 m and 60 m. However, we need all channels to have the same spatial resolution to be able to segment an image. For Sentinel-2 images, Lanaras et al. [62] proposed a deep neural network, called Deep Sentinel-2 (DSen2), that is trained to *super-resolve* the images, i.e., to increase the spatial resolutions of channels with 20 m or 60 m to the highest resolution of 10 m. DSen2 is composed of two convolutional neural networks. One super-resolves the channels with 20 m spatial resolution, the other one the channels with 60 m. Channels with 10 m spatial resolution are not altered. The authors provide pretrained weights for the network. We applied the network as it is using the provided weights. DSen2 excludes the channel for cirrus cloud detection. Hence, we were left with the super-resolved images D1 and D2 that both have a size of 10980×10980 pixels containing twelve channels. From the super-resolved version of D1, we took the cloud-free crop, that was mentioned above, which contains the pixels of the rows 4813-7812 and columns 0-2999, indexed with a 0-based indexing.

As in the case of hyperspectral images, we also scaled the input data such that the intensity values lay in the interval $[0, 1]$. Since the choice of the number of channels to keep by the MNF is a difficult task and also because almost all of the twelve channels in the data contain important information, we did not apply the MNF to reduce the dimensionality but applied ϵ AMS directly to the rescaled data.

The high number of pixels posed an extra challenge. It turned out that only with double precision we obtained reasonable results. However, the memory demand of the image in double precision is about 11.57 GB. Additionally, the PDHG method uses some variables that have at least the size of the image, which increases the used memory even further. Moreover, the higher the number of segments, the higher

the needed memory because the dimensions of the indicator values and the labeling function depend on the number of segments. All this required us to optimize the code with respect to memory and not runtime. The experiments were run on a compute server that is equipped with two AMD EPYC 7443 CPUs with 128 GB memory and two Nvidia A40 with 48 GB memory each. The computations of the indicator functions had to be performed on the CPUs, while the update of the labeling function could be performed on the GPUs. From four segments up, the GPUs had to share their memory using peer-to-peer memory access.

Segmentations were computed on both images D1 and D2 as well as the D1 crop for $k \in \{2, 3, 4, 5\}$ segments and $\lambda \in \{5 \cdot 10^{-5}, 10^{-4}, 5 \cdot 10^{-4}, 0.001, 0.0025\}$. We fixed the regularization parameter ϵ to 10^{-6} , the remaining parameters were chosen as given in Table 4.1.

4.3.2 Results

With this study, we tried to investigate if different sea ice states reflect spectral signatures that enable us to derive these states from Sentinel-2 multispectral image data and if the corresponding spectral clusters are well separated in the spectral space under consideration of the spectral variability. Figs. 4.23 to 4.25 show the

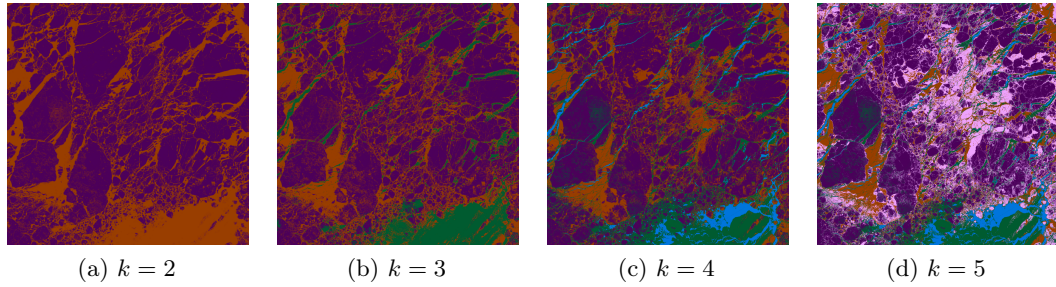


Figure 4.23: Segmentation results of D1 with $\lambda = 5 \cdot 10^{-5}$.

segmentation results of D1, the D1 crop and D2 obtained from ϵ AMS with $\lambda = 5 \cdot 10^{-5}$. We show the results only for $\lambda = 5 \cdot 10^{-5}$ here because these also reveal small structures and have the highest degree of detail. Unfortunately, there is no ground truth data available. We, therefore, base our analysis on qualitative aspects. The analysis was done with the help of an expert from the ice research community. The color coding that is used in the visualizations of the segmentations is chosen such that the colors of the segments with the biggest overlap match across the different numbers of segments for both images D1 and D2 separately to improve comparability. It was done for each image D1 and D2 by fixing the colors for the segments in the visualizations of the segmentations for $k = 5$. These were then used as references and the colors of the

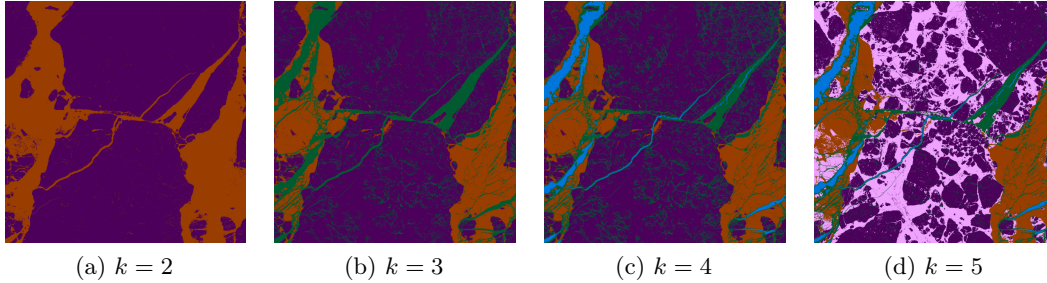


Figure 4.24: Segmentation results of D1 crop with $\lambda = 5 \cdot 10^{-5}$.

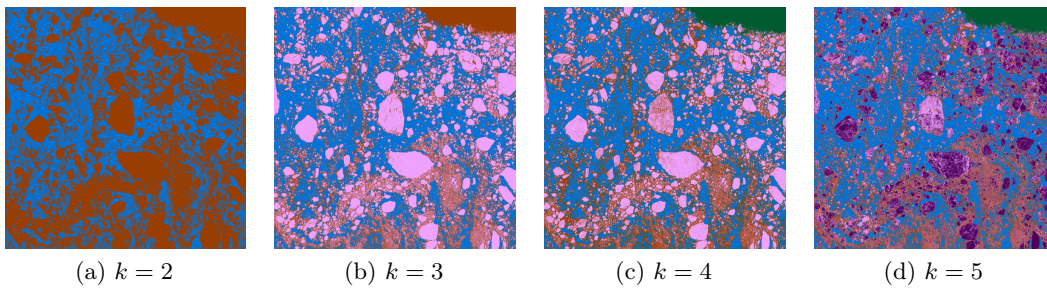


Figure 4.25: Segmentation results of D2 with $\lambda = 5 \cdot 10^{-5}$.

segments in the visualizations where $k < 5$ were aligned by choosing for each segment the color of the segment in the reference that had the biggest overlap. However, the colors do not always mean exactly the same for different k . If the number of segments is altered, also the corresponding groupings of the spectra have to change. Spectral clusters are being split up or merged. For example, if the number of segments is increased, the algorithm is allowed to divide the spectral clusters into more groups. Hence, it could divide a cluster that was formed by spectra representing young ice and cloud into two clusters, one for young ice and one for cloud, but one of these segments would still obtain the same color in the visualization.

Interpretation of segmentation results

The binary segmentation, i.e., $k = 2$, seems to approximately distinguish between water and ice. We computed the sea ice concentration for both images, which is estimated by our method as the fraction of pixels assigned to the segment that covers most of the ice. The decision, which of the two segments represents ice, was done by an expert. The sea ice concentration is relevant, e.g., for navigation. These values could be compared to the estimated sea ice concentration values of the Norwegian

Meteorological Institute (MET) [82] and Danish Meteorological Institute (DMI) [31] for D1 and D2, respectively, that estimated the sea ice concentrations in the same geographical areas on the same days but with different methods. Moreover, Zhang and Hughes estimated the sea ice concentration of D2 based on their supervised segmentation result of that image [113]. The sea ice concentration estimates are listed in Table 4.8. Please note that the estimated sea ice concentration values are not entirely comparable since the estimations of MET and DMI are from the exact same days as D1 and D2 but only roughly the same times. Moreover, MET and DMI do not give single values as estimates but only intervals. Because ice is drifting, the sea ice concentration in a fixed area is constantly changing. In the case of D1, it can

Data-set	MET (D1)/ DMI (D2)	Zhang, Hughes [113]	λ				
			$5 \cdot 10^{-5}$	10^{-4}	$5 \cdot 10^{-4}$	0.001	0.0025
D1	[0.625, 1]	-	0.649	0.654	0.670	0.675	0.681
D2	[0.4, 0.6]	0.515	0.595	0.595	0.597	0.599	0.605

Table 4.8: Comparison of estimated sea ice concentrations in D1 and D2.

be observed that the estimated sea ice concentration values from the segmentations produced by ϵ AMS are quite stable across the different values of λ . Moreover, they agree with the range that was estimated by the MET on the same day. Please note that the sea ice concentration estimate of the MET is a broad interval since the MET estimates the sea ice concentration in a larger area than a Sentinel-2 image covers. The true sea ice concentration is probably higher than estimated by our algorithm because a manual inspection by an expert showed that the model merged newer ice states with the water segment. Hence, the reflectance spectra of newer sea ice states resemble more the water spectra than the spectra of older ice. Additionally, also ice covered by a thin cloud was put into the water segment. In D2, the segment representing ice also contains pixels of some regions between the ice floes. Additionally, the thin cloud in the upper right corner is counted towards that segment. When we inspect the sea ice concentration values of D2 in Table 4.8, we see that the sea ice concentration is at the upper boundary of the interval estimated by the DMI and significantly higher than the value that was estimated by Zhang and Hughes. A reason for this overestimation is the thin cloud in the upper right corner. In particular, also water covered by the cloud was included in the ice segment, meaning that the cloud is shifting the reflectance spectra of water towards the sea ice spectra.

It is challenging to obtain well-separated spectral clusters when clouds are present and the number of segments k is small. With an increasing number of segments, the meaningfulness and interpretability of the classes improve. The water class, for example, stabilized and saturated with an increasing k in case of both images. In the case of D2 with $k \geq 4$, we obtained an extra class for the thin cloud, which allowed

for a derivation of well-separated segments also for the ice states (cf. Figs. 4.25c and 4.25d). If we consider the segmentation of D2 into $k = 5$ classes in Fig. 4.25d, a plausible meaning of the segments could be: water and liquid states (blue), thin cloud (green), different states of ice floe surfaces with different reflectance spectra (purple and pink), a mixture of ice objects in water, e.g., surface slush and brash ice (thin red-brown boundary of the purple and pink segment). The interpretation that the pink and purple segments are ice floe surfaces with different reflectance spectra is supported by a visual inspection of the RGB image after the use of gamma correction, where one can see a faint difference of the two segments in brightness.

Effect of clouds

The investigation of the effect of clouds is based on the segmentation of the D1 crop compared to the corresponding crop of the segmentation of the full D1 image. Both segmentations are shown in Fig. 4.26. In Fig. 4.26a, the obtained segmentation of

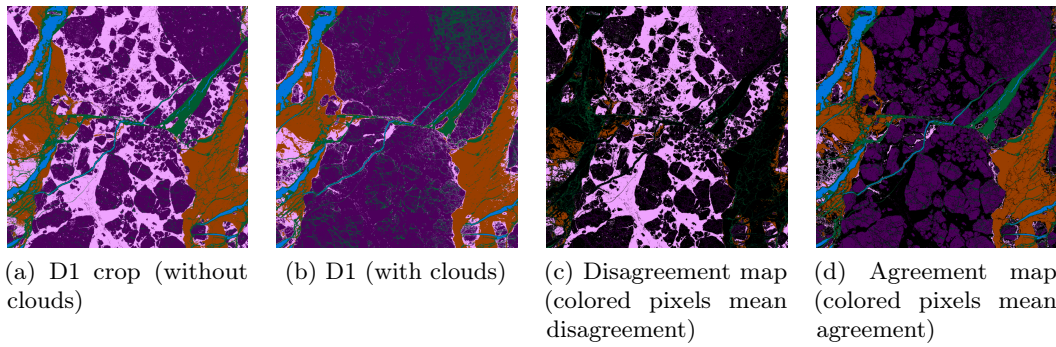


Figure 4.26: Comparison of the segmentation result of D1 crop and the corresponding crop of the full D1 segmentation with $\lambda = 5 \cdot 10^{-5}$ and $k = 5$.

the D1 crop with $\lambda = 5 \cdot 10^{-5}$ and $k = 5$ is shown next to the corresponding crop from the segmentation of the full D1 image (cf. Fig. 4.26b). Comparing the image in Fig. 4.22b with the segmentation in Fig. 4.26a, the segmentation of the crop seems to be more meaningful than the full segmentation in Fig. 4.26b. The ice structures that one can observe through the different brightness levels in Fig. 4.22b are not properly detected in the case of the full image segmentation. This is also supported by the agreement and disagreement maps in Figs. 4.26c and 4.26d. A difference between the complete D1 image and the D1 crop is that the crop comprises only a part of the spectra of the image. Some spectral clusters are potentially missing, in particular, the spectra affected by the cloud in the upper right corner. Hence, the method does not have to decide which clusters it merges with the cloud spectra. This allows for better

estimates of mean and covariance of the segments since there is less perturbation present that is caused by other spectra.

We also computed the agreement scores, i.e., the fraction of pixels where the two segmentations coincide. They can be found in Table 4.9. One can see that the

$k \backslash \lambda$	$5 \cdot 10^{-5}$	10^{-4}	$5 \cdot 10^{-4}$	0.001	0.0025
2	0.92	0.93	0.96	0.97	0.96
3	0.84	0.84	0.88	0.91	0.92
4	0.92	0.92	0.92	0.94	0.96
5	0.63	0.64	0.78	0.76	0.65

Table 4.9: Agreement scores of the segmentation of the D1 crop and the crop of the full D1 segmentation.

agreement drops when the number of segments k is increased to 5. The reason is that in the case of the D1 crop (cf. Figs. 4.24c and 4.24d) the model splits mainly the spectral cluster of the purple segment into two clusters that are marked in pink and purple when we increase the number of segments from 4 to 5. In the full D1 image (cf. Figs. 4.23c and 4.23d), it is mainly the spectral cluster of the red-brown segment that is split up. We see the effect of the spectra that are present in the full image but absent in the crop also here. Only a small part of the spectra that were split in the case of the full image are actually present in the crop; in particular, they are not affected by clouds. Hence, the algorithm can split other spectral clusters in the case of the crop to form a fifth segment. This means that some pixels of the same segment in the full image could be placed into different segments in the crop. This results in low agreement scores.

The obtained preliminary results have shown that our method ϵ AMS is able to derive plausible complex sea ice states and also sea ice concentration values. Moreover, it could also be used to enhance other methods for sea ice concentration estimation and for space-borne observation of the characteristic reflectance spectra of sea ice states. However, the effect of clouds has to be reduced, for example, by masking the clouds.

Conclusions

In this thesis, we investigated and proposed models for unsupervised hyperspectral image segmentation that are able to handle the special characteristics of hyperspectral images. To describe the hyperspectral data, we used two different approaches: (1) modeling the spectral distributions with first- and second-order statistics that are estimated from the data (ϵ AMS and ϕ AMS) and (2) lifting the data to a potentially high-dimensional Hilbert space where the data is ideally disentangled (kMS). This resulted in three different MS-based segmentation functionals. Furthermore, we introduced two different regularizations for the covariance matrix to ensure its positive definiteness and, consequently, the feasibility of our models ϵ AMS and ϕ AMS.

In a theoretical analysis of ϵ AMS and ϕ AMS, we proved the existence of minimizers for both models and, in particular, evaluated the importance of the ϵ -regularization for ϵ AMS. It turned out that the regularization parameter ϵ plays a crucial role for the model, in particular, when proving the existence of minimizers of ϵ AMS. Since the functional is not coercive, the covariance estimate can be pushed out of the admissible set of positive definite matrices. The regularization parameter fixes this by making the functional domain closed with respect to the covariance matrix, which ensures that the covariance estimate always stays in the admissible set. This is essential for the existence of minimizers. Moreover, we proved that ϵ AMS Γ -converges to a Γ -limit functional when ϵ tends to zero. In the Γ -limit, we lose the closedness of the functional domain and, consequently, the guarantee for the existence of minimizers of ϵ AMS. We furthermore showed in a counterexample that the Γ -limit functional does not have a minimizer in general.

We proposed a segmentation workflow for hyperspectral image segmentation, including an application of the MNF as a preprocessing and an alternating optimization with a novel stopping criterion to optimize the models with respect to their degrees of freedom. The introduced models required techniques to optimize the model-specific parameters in practice. To this end, we proposed fixed point iteration schemes for ϵ AMS and ϕ AMS; in the case of kMS, we derived a closed-form solution.

The numerical results showed the great potential and effectiveness of the three models. In particular, the two models ϵ AMS and ϕ AMS performed consistently better

than all the competing methods and provided segmentations of high quality on the used test datasets. An evaluation of the preprocessing step with the MNF showed that the MNF is important to improve the results but the crucial contribution comes from the models themselves. Moreover, we collected indications that our alternating optimization approach indeed shows a converging behavior.

Finally, we tested ϵ AMS also on real-world multispectral Sentinel-2 images and could show that the model is even able to provide useful segmentations of such images that allow for a derivation of complex sea ice states. This indicates that our method has the potential to perform well in a broad range of applications.

There remain open questions for future work. A frequent problem of models for hyperspectral image segmentation is the lack of useful heuristics for the choice of parameters that need to be set before the application of the models. This has not been investigated in this thesis for our models but would be of interest for practical applications. Another aspect that would have to be investigated in the future is how to deal with segments that do not contain pixels during optimization. A problem is that without any pixels in a segment, we are not able to estimate the model parameters for that segment in that iteration, even though it might make sense to have an empty segment, for example, if we try to find a segmentation of a constant image into two segments.

As we have seen in the theoretical analysis, a regularization of the covariance estimates is of utter importance for the success of the models ϵ AMS and ϕ AMS. However, the proposed regularizations for ϵ AMS and ϕ AMS yield data-independent lower bounds for the eigenvalues of the covariance matrices. Ideally, a model would choose the lower bounds for the eigenvalues itself based on the data. Therefore, it would be of great interest to find answers to the questions (1) which practical conditions on the data ensure a data-dependent lower bound for the eigenvalues without any extra regularization of the covariance matrices such that a minimizer for the functional exists and (2) how we have to modify the model ϵ AMS such that it always has a data-dependent lower bound for the eigenvalues and a minimizer, independent of any condition on the data and without any extra regularization of the eigenvalues.

Even though kMS could show its great potential in the numerical experiments, the required memory is tremendous. Hence, it has to be investigated how the kernel matrix can effectively be approximated, e.g., with low rank approximations. The huge number of channels of hyperspectral images poses extra challenges to memory management. It would be interesting to investigate streaming models that load only chunks of the data into memory for hyperspectral image processing.

We could find numerical indications that our alternating optimization consisting of a fixed point iteration scheme in the case of ϵ AMS and ϕ AMS or a closed-form solution in the case of kMS and the PDHG method shows a converging behavior.

Hence, as a future direction, further analysis and a proof of convergence of the alternating optimization is desirable.

Finally, another question that is relevant for the practical application and needs investigation is the dependence of the results on the initialization.

We believe that the proposed segmentation models in combination with the introduced segmentation workflow have a great potential to improve segmentation-based algorithms and workflows in a broad range of real-world applications by delivering accurate segmentation results.

Appendix A

Mathematical foundations and results

Please note that the notation described in Section 1.4 also applies here. Several statements are stated in a form we need for the thesis, but could be stated in a more general context. Measurability is meant with respect to the Borel σ -algebra and $|\cdot|$ denotes the Lebesgue measure. Definitions A.8 to A.11 follow the corresponding definitions in [4].

Lemma A.1.

Let $M \in \mathbb{R}^{n \times n}$ and $a, b, x \in \mathbb{R}^n$.

(a) $h(x) := a^T x \quad \Rightarrow \quad \nabla h(x) = a$

(b) $h(x) := x^T M x \quad \Rightarrow \quad \nabla h(x) = Mx + M^T x$

(c) Let M be additionally invertible.

$$h(M) := a^T M^{-1} b \quad \Rightarrow \quad \nabla h(M) = - (M^{-1})^T a b^T (M^{-1})^T$$

(d) If $\det M > 0$, then we have

$$h(M) := \log \det M \quad \Rightarrow \quad \nabla h(M) = (M^{-1})^T.$$

(e) $h(M) := a^T M^T b \quad \Rightarrow \quad \nabla h(M) = b a^T$

(f) Let $r \in \{1, \dots, n\}$.

$$h(M) := [M^T a]_r \quad \Rightarrow \quad \nabla h(M) = a e_r^T$$

(g) Let $\Lambda \in \mathbb{R}^{n \times n}$ be a diagonal matrix.

$$h(M) := a^T M \Lambda M^T a \quad \Rightarrow \quad \nabla h(M) = 2 a a^T M \Lambda$$

$$h(M) = [M^T a]_r = e_r^T M^T a$$
$$h(M) = (M^T a)^T \Lambda (M^T a) = \sum_{i=1}^n \lambda_i (M^T a)_i^2 = \sum_{i=1}^n \lambda_i \left(\sum_{j=1}^n m_{ji} a_j \right)^2.$$
$$\begin{aligned} \frac{\partial h(M)}{\partial m_{rs}} &= \frac{\partial}{\partial m_{rs}} \sum_{i=1}^n \lambda_i \left(\sum_{j=1}^n m_{ji} a_j \right)^2 \\ &= \lambda_s \frac{\partial}{\partial m_{rs}} \left(\sum_{j=1}^n m_{js} a_j \right)^2 \\ &= 2 \lambda_s \left(\sum_{j=1}^n m_{js} a_j \right) a_r \\ &= 2 a_r \left(\sum_{j=1}^n m_{js} a_j \right) \lambda_s. \end{aligned}$$
$$\begin{aligned} (2aa^T M \Lambda)_{rs} &= 2 \left(aa^T \begin{pmatrix} | & & | \\ \lambda_1 m_1 & \dots & \lambda_n m_n \\ | & & | \end{pmatrix} \right)_{rs} \\ &= 2 \left(a \left(\sum_{i=1}^n \lambda_1 a_i m_{i1}, \dots, \sum_{i=1}^n \lambda_n a_i m_{in} \right) \right)_{rs}. \end{aligned}$$

The vertical bars in the expression after the first equality sign indicate that $\lambda_i m_i$ for every $i \in \{1, \dots, n\}$ is a column vector. The last expression describes the (r, s) -th

entry of the outer product of the two vectors a and

$$\left(\sum_{i=1}^n \lambda_1 a_i m_{i1}, \dots, \sum_{i=1}^n \lambda_n a_i m_{in} \right)^T \in \mathbb{R}^n.$$

Having in mind that the (r, s) -th entry of an outer product of two vectors $b, x \in \mathbb{R}^n$ is

$$(bx^T)_{rs} = b_r x_s,$$

we get

$$(2aa^T M \Lambda)_{rs} = 2a_r \sum_{i=1}^n \lambda_s a_i m_{is} = 2a_r \left(\sum_{i=1}^n a_i m_{is} \right) \lambda_s.$$

This expression coincides with $\frac{\partial h(M)}{\partial m_{rs}}$ from above and we conclude

$$\nabla h(M) = 2aa^T M \Lambda. \quad \square$$

Definition A.2 (Subgradient, subdifferential [8, Def. 3.1, Def. 3.2]).

Let X be a finite-dimensional vector space with inner product $\langle \cdot, \cdot \rangle_X$ and let the function $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ be proper (cf. Definition 3.2). Then, we call the element $u \in X$ a subgradient of f at $x \in \text{dom}(f)$ if

$$f(x) + \langle u, y - x \rangle_X \leq f(y)$$

for any $y \in X$. We denote the set of all subgradients of f at x with $\partial f(x)$ and call it the subdifferential of f at x .

Please note that one can introduce the subgradient and subdifferential for functionals on more general (possibly infinite-dimensional) vector spaces.

Proposition A.3 (Simultaneous diagonalization [45, Th. 12.2.13 (1)]).

Let $M_1, M_2 \in \mathbb{R}^{d \times d}$ be symmetric and let M_2 be additionally positive definite. Then, the existence of an invertible matrix $V \in \mathbb{R}^{d \times d}$ such that $V M_1 V^T = \Lambda$ and $V M_2 V^T = I$ with a diagonal matrix $\Lambda \in \mathbb{R}^{d \times d}$ is guaranteed. The diagonal entries of Λ are given as the roots of the polynomial $p(\lambda) = \det(M_1 - \lambda M_2)$.

Definition A.4 (Skew-symmetric matrix).

Let $M \in \mathbb{R}^{n \times n}$ be a matrix. We call M skew-symmetric if it fulfills $M^T = -M$.

Definition A.5 (Convexity of functionals [15, Def. 6.20]).

Let X be a normed vector space and $J: X \rightarrow \mathbb{R} \cup \{\infty\}$ a functional. If for any $x, y \in X$ and $\alpha \in [0, 1]$ it follows that

$$J[\alpha x + (1 - \alpha)y] \leq \alpha J[x] + (1 - \alpha)J[y],$$

the functional J is said to be convex. If we have that even the strict inequality

$$J[\alpha x + (1 - \alpha)y] < \alpha J[x] + (1 - \alpha)J[y]$$

holds for any $x, y \in X$ with $x \neq y$ and $\alpha \in (0, 1)$, then J is said to be strictly convex.

Please note that this definition needs the usual conventions for arithmetics in $\mathbb{R} \cup \{\infty\}$.

Lemma A.6 (Convexity and continuity of norms).

Let X be a normed vector space with norm $\|\cdot\|_X$. Then, $\|\cdot\|_X$ is convex and continuous.

Proof. The convexity follows directly from the homogeneity and triangle inequality of norms. The continuity is a consequence of the negative triangle inequality because for $x, y \in X$ with $x \rightarrow y$ we have

$$0 \leq \|x\|_X - \|y\|_X \leq \|x - y\|_X \rightarrow 0. \quad \square$$

Lemma A.7.

Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Then, the squared induced norm $\|\cdot\|_{\mathcal{H}}^2 = \langle \cdot, \cdot \rangle_{\mathcal{H}}$ is strictly convex.

Proof. Let $\alpha \in (0, 1)$ and $x, y \in \mathcal{H}$ such that $x \neq y$. We define the function

$$h: \mathbb{R} \rightarrow \mathbb{R}, \quad t \mapsto \|tx + (1 - t)y\|_{\mathcal{H}}^2,$$

which can be reformulated with the linearity and symmetry of the inner product to

$$\begin{aligned} h(t) &= \langle tx + (1 - t)y, tx + (1 - t)y \rangle_{\mathcal{H}} \\ &= t^2 \langle x, x \rangle_{\mathcal{H}} + 2t(1 - t) \langle x, y \rangle_{\mathcal{H}} + (1 - t)^2 \langle y, y \rangle_{\mathcal{H}}. \end{aligned}$$

We observe that h is a polynomial and, therefore, differentiable. The first derivative is

$$\begin{aligned} h'(t) &= 2t \langle x, x \rangle_{\mathcal{H}} + 2(1 - 2t) \langle x, y \rangle_{\mathcal{H}} - 2(1 - t) \langle y, y \rangle_{\mathcal{H}} \\ &= 2t \langle x, x \rangle_{\mathcal{H}} + 2 \langle x, y \rangle_{\mathcal{H}} - 4t \langle x, y \rangle_{\mathcal{H}} - 2 \langle y, y \rangle_{\mathcal{H}} + 2t \langle y, y \rangle_{\mathcal{H}} \\ &= 2(\langle x, x \rangle_{\mathcal{H}} + \langle y, y \rangle_{\mathcal{H}})t + 2 \langle x, y \rangle_{\mathcal{H}} - 4t \langle x, y \rangle_{\mathcal{H}} - 2 \langle y, y \rangle_{\mathcal{H}}. \end{aligned}$$

As the second derivative, we obtain with the linearity and symmetry of the inner product

$$\begin{aligned} h''(t) &= 2(\langle x, x \rangle_{\mathcal{H}} + \langle y, y \rangle_{\mathcal{H}}) - 4\langle x, y \rangle_{\mathcal{H}} \\ &= 2(\langle x, x \rangle_{\mathcal{H}} - 2\langle x, y \rangle_{\mathcal{H}} + \langle y, y \rangle_{\mathcal{H}}) \\ &= 2\|x - y\|_{\mathcal{H}}^2. \end{aligned}$$

Since $x \neq y$, it follows $h''(t) > 0$ for any $t \in \mathbb{R}$, which is an equivalent characterization for strict convexity, and hence, h is strictly convex and fulfills for any $\beta \in (0, 1)$ and $t_1, t_2 \in \mathbb{R}$ with $t_1 \neq t_2$

$$h(\beta t_1 + (1 - \beta)t_2) < \beta h(t_1) + (1 - \beta)h(t_2).$$

Choosing $\beta = \alpha$, $t_1 = 1$ and $t_2 = 0$, it follows

$$\begin{aligned} \|\alpha x + (1 - \alpha)y\|_{\mathcal{H}}^2 &= h(\alpha) \\ &= h(\alpha t_1 + (1 - \alpha)t_2) \\ &< \alpha h(t_1) + (1 - \alpha)h(t_2) \\ &= \alpha h(1) + (1 - \alpha)h(0) \\ &= \alpha \|x\|_{\mathcal{H}}^2 + (1 - \alpha) \|y\|_{\mathcal{H}}^2, \end{aligned}$$

and therefore the strict convexity of $\|\cdot\|_{\mathcal{H}}^2$. □

Definition A.8 (Bounded domain with Lipschitz boundary).

A set $\Omega \subseteq \mathbb{R}^d$ that is non-empty, open and connected is called a domain. We refer to Ω as a bounded domain if it is additionally bounded. If Ω is a bounded domain, we call it bounded Lipschitz domain or bounded domain with Lipschitz boundary if there exists an open cover for the boundary $\partial\Omega$ of Ω of finitely many open sets $U_1, \dots, U_n \subseteq \mathbb{R}^d$ such that for any $i \in \{1, \dots, n\}$, the set $\partial\Omega \cap U_i$ is the graph of a Lipschitz continuous function and $\Omega \cap U_i$ lies only on one side of this graph.

Definition A.9 (Continuously differentiable functions).

Let $\Omega \subseteq \mathbb{R}^d$ be open and let $m, n \in \mathbb{N}$. Then, we define the space of m -times continuously differentiable functions as

$$C^m(\Omega, \mathbb{R}^n) := \{f: \Omega \rightarrow \mathbb{R}^n \mid f \text{ } m\text{-times continuously differentiable}\}$$

and the space of m -times continuously differentiable functions with compact support as

$$C_c^m(\Omega, \mathbb{R}^n) := \{f \in C^m(\Omega, \mathbb{R}^n) \mid \text{supp } f \subseteq \Omega \text{ is compact}\},$$

where the support of a function $f \in C^m(\Omega, \mathbb{R}^n)$ is

$$\text{supp } f := \overline{\{x \in \Omega \mid f(x) \neq 0\}}.$$

If $n = 1$, we write $C^m(\Omega)$ and $C_c^m(\Omega)$, respectively.

Definition A.10 (Lebesgue spaces).

Let $\Omega \subseteq \mathbb{R}^d$. We define the Lebesgue space $L^p(\Omega, \mathbb{R}^n)$ for $n \in \mathbb{N}$ and $1 \leq p \leq \infty$ as

$$L^p(\Omega, \mathbb{R}^n) := \left\{ u: \Omega \rightarrow \mathbb{R}^n \mid u \text{ measurable, } \|u\|_{L^p(\Omega, \mathbb{R}^n)} < \infty \right\}$$

with the equivalence relation

$$u = v \text{ in } L^p(\Omega, \mathbb{R}^n) :\Leftrightarrow u = v \text{ a.e.}$$

The norm $\|\cdot\|_{L^p(\Omega, \mathbb{R}^n)}$ is defined as

$$\|u\|_{L^p(\Omega, \mathbb{R}^n)} := \left(\int_{\Omega} \|u(x)\|_p^p \, dx \right)^{1/p}$$

for $p < \infty$ and

$$\|u\|_{L^\infty(\Omega, \mathbb{R}^n)} := \inf_{\substack{N \subseteq \Omega, \\ |N|=0}} \left(\sup_{x \in \Omega \setminus N} \|u(x)\|_\infty \right)$$

for $p = \infty$. If $n = 1$, we write $L^p(\Omega)$.

Definition A.11 (Sobolev spaces).

Let $\Omega \subseteq \mathbb{R}^d$ be open and bounded with Lipschitz boundary. For $1 \leq p \leq \infty$, the Sobolev space of order $m \in \mathbb{N}$ with exponent p is defined as

$$H^{m,p}(\Omega) := \left\{ f \in L^p(\Omega) \mid \forall \alpha \in \mathbb{N}_0^d: |\alpha| \leq m \exists \partial^\alpha f \in L^p(\Omega) \forall \varphi \in C_c^\infty(\Omega): \right. \\ \left. \int_{\Omega} \partial^\alpha f(x) \varphi(x) \, dx = (-1)^{|\alpha|} \int_{\Omega} f(x) \partial^\alpha \varphi(x) \, dx \right\}.$$

The index α is a d -dimensional multi-index, that is, $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| := |\alpha_1| + \dots + |\alpha_d|$ indexing the partial derivative of higher order of φ at $x \in \Omega$

$$\partial^\alpha \varphi(x) := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} \varphi(x).$$

The functions $\partial^\alpha f \in L^p(\Omega)$ are the so-called weak derivatives of $f \in H^{m,p}(\Omega)$. The space $H^{m,p}(\Omega)$ is endowed with the norm

$$\|f\|_{H^{m,p}(\Omega)} := \sum_{|\alpha| \leq m} \|\partial^\alpha f\|_{L^p(\Omega)}.$$

Proposition A.12 (Integration by parts for C^1 -functions [40, § 10, Satz 2 b])).

Let $\Omega \subseteq \mathbb{R}^d$ be open. For $i \in \{1, \dots, d\}$ we have

$$\int_{\Omega} \partial_{x_i} f(x) \varphi(x) \, dx = - \int_{\Omega} f(x) \partial_{x_i} \varphi(x) \, dx$$

for every $f \in C^1(\Omega)$ and $\varphi \in C_c^1(\Omega)$.

Theorem A.13 ($C_c^\infty(\Omega)$ dense in $L^p(\Omega)$ [4, Th. 4.15 (3)]).

Let $\Omega \subseteq \mathbb{R}^d$ be open. For $1 \leq p < \infty$, $C_c^\infty(\Omega)$ is a dense subset of $L^p(\Omega)$.

Proposition A.14.

Let $\Omega \subseteq \mathbb{R}^d$ be open and bounded. Furthermore, let $\Omega' \subseteq \mathbb{R}^d$ be open and such that $\Omega \subseteq \Omega'$. Then, we have that for $\varphi \in C_c^\infty(\Omega')$ it holds $\varphi|_{\Omega} \in H^{1,p}(\Omega)$ for $1 \leq p \leq \infty$.

Proof. Let $\varphi \in C_c^\infty(\Omega')$. Since φ is continuous and has compact support, we have that φ is bounded, i.e., $|\varphi(x)| < \infty$ for every $x \in \Omega'$. Moreover, for each $i \in \{1, \dots, d\}$ the function $\partial_{x_i} \varphi$ exists, is continuous on Ω' and its support is bounded since it is contained in the support of φ , which is bounded due to its compactness. It follows that $\partial_{x_i} \varphi$ is also bounded, i.e., $|\partial_{x_i} \varphi(x)| < \infty$ for any $x \in \Omega'$. The boundedness of φ and $\partial_{x_i} \varphi$ for every $i \in \{1, \dots, d\}$ on Ω' implies the boundedness on $\Omega \subseteq \Omega'$. Since continuous functions are measurable, it follows that φ and $\partial_{x_i} \varphi$ for every $i \in \{1, \dots, d\}$ are measurable. This implies $\|\varphi\|_{L^\infty(\Omega)} < \infty$ and $\|\partial_{x_i} \varphi\|_{L^\infty(\Omega)} < \infty$, and φ and $\partial_{x_i} \varphi$ are therefore elements of $L^\infty(\Omega)$ for any $i \in \{1, \dots, d\}$. For $p \neq \infty$, we have

$$\|\varphi\|_{L^p(\Omega)}^p = \int_{\Omega} |\varphi(x)|^p \, dx \leq |\Omega| \sup_{x \in \Omega} |\varphi(x)|^p < \infty$$

by the boundedness of φ on Ω and the boundedness of Ω , and hence, $\|\varphi\|_{L^p(\Omega)} < \infty$. Doing the same computation for $\partial_{x_i} \varphi$, we can deduce $\|\partial_{x_i} \varphi\|_{L^p(\Omega)} < \infty$ for any $i \in \{1, \dots, d\}$. We obtain $\varphi \in L^p(\Omega)$ and $\partial_{x_i} \varphi \in L^p(\Omega)$. By integration by parts for continuous functions (cf. Proposition A.12), it follows that φ fulfills the integral condition in Definition A.11. We conclude that $\varphi \in H^{1,p}(\Omega)$ for $1 \leq p \leq \infty$. $\square$

Theorem A.15 (Gauß's theorem [4, A8.8]).

Let $\Omega \subseteq \mathbb{R}^d$ be an open and bounded set with Lipschitz boundary.

(1) Let $u \in H^{1,1}(\Omega)$. We have for every $i \in \{1, \dots, d\}$ the equality

$$\int_{\Omega} \partial_{x_i} u(x) \, dx = \int_{\partial\Omega} u(x) \nu_i(x) \, d\mathcal{H}^{d-1}(x),$$

with ν being the outer normal of $\partial\Omega$.

(2) For $p \in [1, \infty]$ and $q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, let $u \in H^{1,p}(\Omega)$ and $v \in H^{1,q}(\Omega)$. Then, it holds for every $i \in \{1, \dots, d\}$

$$\int_{\Omega} (u(x)\partial_{x_i}v(x) + v(x)\partial_{x_i}u(x)) \, dx = \int_{\partial\Omega} u(x)v(x)\nu_i(x) \, d\mathcal{H}^{d-1}(x),$$

and again with ν being the outer normal of $\partial\Omega$.

Proposition A.16 ([72, Cor. 2.6.20]).

If X is a separable normed space and $A \subseteq X^*$ a bounded subset of the dual space X^* of X , then there is a metric that induces the weak* topology of A .

Bibliography

- [1] N. Acito, G. Corsini, and M. Diani. “An unsupervised algorithm for hyperspectral image segmentation based on the Gaussian mixture model.” In: *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*. Vol. 6. IGARSS-03. 2003. DOI: 10.1109/IGARSS.2003.1295256.
- [2] N. Agrawal and K. Verma. “Dimensionality Reduction on Hyperspectral Data Set.” In: *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*. 2020. DOI: 10.1109/ICPC2T48082.2020.9071461.
- [3] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. “Optuna: a next-generation hyperparameter optimization framework.” In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.
- [4] H. W. Alt. *Linear Functional Analysis: An Application-Oriented Introduction*. Universitext 1. Springer London, 2016. DOI: 10.1007/978-1-4471-7280-2.
- [5] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. Oxford University Press, 2000.
- [6] L. Bar, T. F. Chan, G. Chung, M. Jung, N. Kiryati, R. Mohieddine, N. Sochen, and L. A. Vese. “Mumford and Shah Model and its Applications to Image Segmentation and Image Restoration.” In: *Handbook of Mathematical Methods in Imaging*. 2011. Chap. 25, pp. 1097–1157. DOI: 10.1007/978-0-387-92920-0.
- [7] M. P. Barbato, P. Napoletano, F. Piccoli, and R. Schettini. “Unsupervised segmentation of hyperspectral remote sensing images with superpixels.” In: *Remote Sensing Applications: Society and Environment* 28 (2022), p. 100823. DOI: 10.1016/j.rsase.2022.100823.

- [8] A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2017. DOI: 10.1137/1.9781611974997.
- [9] C. Bench, J. Nallala, C.-C. Wang, H. Sheridan, and N. Stone. “Unsupervised segmentation of biomedical hyperspectral image data: tackling high dimensionality with convolutional autoencoders.” In: *Biomedical Optics Express* 13.12 (2022), pp. 6373–6388. DOI: 10.1364/boe.476233.
- [10] M. Berman, A. Phatak, and A. Traylen. “Some invariance properties of the minimum noise fraction transform.” In: *Chemometrics and Intelligent Laboratory Systems* 117 (2012), pp. 189–199. DOI: 10.1016/j.chemolab.2012.02.005.
- [11] J. M. 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.” In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.2 (2012), pp. 354–379. DOI: 10.1109/JSTARS.2012.2194696.
- [12] C. M. Bishop. *Pattern Recognition and Machine Learning*. Vol. 1. Springer-Verlag New York Inc., 2006.
- [13] E. Bozzi, L. Crocetti, and R. Lencioni. “US Image Acquisition.” In: *Image Processing in Radiology: Current Applications*. Medical Radiology. 2008, pp. 3–13. DOI: 10.1007/978-3-540-49830-8_1.
- [14] A. Braides. *Gamma-convergence for Beginners*. Oxford University Press, 2002. DOI: 10.1093/acprof:oso/9780198507840.001.0001.
- [15] K. Bredies and D. Lorenz. *Mathematical Image Processing*. Applied and Numerical Harmonic Analysis 1. Birkhäuser Cham, 2018. DOI: 10.1007/978-3-030-01458-2.
- [16] O. Burggraaff, N. Schmidt, J. Zamorano, K. Pauly, S. Pascual, C. Tapia, E. Spyrakos, and F. Snik. “Standardized spectral and radiometric calibration of consumer cameras.” In: *Optics Express* 27.14 (2019), pp. 19075–19101. DOI: 10.1364/OE.27.019075.
- [17] X. Cai, R. Chan, M. Nikolova, and T. Zeng. “A Three-Stage Approach for Segmenting Degraded Color Images: Smoothing, Lifting and Thresholding (SLaT).” In: *Journal of Scientific Computing* 72 (2017), pp. 1313–1332. DOI: 10.1007/s10915-017-0402-2.
- [18] X. Cai, R. Chan, and T. Zeng. “A Two-Stage Image Segmentation Method Using a Convex Variant of the Mumford–Shah Model and Thresholding.” In: *SIAM Journal on Imaging Sciences* 6.1 (2013), pp. 368–390. DOI: 10.1137/120867068.

- [19] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla. “Composite kernels for hyperspectral image classification.” In: *IEEE Geoscience and Remote Sensing Letters* 3.1 (2006), pp. 93–97. DOI: 10.1109/LGRS.2005.857031.
- [20] C. Cariou and K. Chehdi. “Unsupervised Nearest Neighbors Clustering With Application to Hyperspectral Images.” In: *IEEE Journal of Selected Topics in Signal Processing* 9.6 (2015), pp. 1105–1116. DOI: 10.1109/JSTSP.2015.2413371.
- [21] A. Chambolle. “An Algorithm for Total Variation Minimization and Applications.” In: *Journal of Mathematical Imaging and Vision* 20 (2004), pp. 89–97. DOI: 10.1023/B:JMIV.0000011325.36760.1e.
- [22] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. “An Introduction to Total Variation for Image Analysis.” In: *Theoretical Foundations and Numerical Methods for Sparse Recovery*. 2010, pp. 263–340. DOI: 10.1515/9783110226157.263.
- [23] A. Chambolle and T. Pock. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging.” In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145. DOI: <https://doi.org/10.1007/s10851-010-0251-1>.
- [24] M. Cholewa and P. Glomb. “Two Stage SVM Classification for Hyperspectral Data.” In: *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*. INSTICC. 2016, pp. 387–391. DOI: 10.5220/0005828103870391.
- [25] J. Cohen. “A coefficient of agreement for nominal scales.” In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.
- [26] J.-C. Cohrs, C. Bajaj, and B. Berkels. “A Distribution-Dependent Mumford–Shah Model for Unsupervised Hyperspectral Image Segmentation.” In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–21. DOI: 10.1109/TGRS.2022.3227061.
- [27] J.-C. Cohrs and B. Berkels. “On the importance of the ε -regularization of the distribution-dependent Mumford–Shah model for hyperspectral image segmentation.” In: *Multiscale, Nonlinear and Adaptive Approximation II*. 2024, pp. 201–222. DOI: 10.1007/978-3-031-75802-7_10.
- [28] J.-C. Cohrs, E. Kim, and B. Berkels. “Sea-Ice States: Unsupervised Segmentation from Multispectral Data Based on Mumford–Shah Functional.” In: *Proceedings of the 27th IAHR International Symposium on Ice*. 2024.

- [29] J. B. Conway. *A course in functional analysis*. Second edition. Graduate texts in mathematics 96. Springer, 1990.
- [30] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Sixteenth print. Cambridge University Press, 2014.
- [31] Danish Meteorological Institute, Ice Service. *Greenland ice chart from 06 June 2021*. 2021. URL: https://ocean.dmi.dk/arctic/images/MODIS/Greenland_WA/202106061200.ISK0.pdf (visited on 09/08/2024).
- [32] G. David. *Singular Sets of Minimizers for the Mumford-Shah Functional*. Progress in Mathematics. Birkhäuser Basel, 2005. DOI: 10.1007/b137039.
- [33] E. De Giorgi, M. Carriero, and A. Leaci. “Existence theorem for a minimum problem with free discontinuity set.” In: *Archive for Rational Mechanics and Analysis* 108 (1989), pp. 195–218. DOI: 10.1007/BF01052971.
- [34] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services.” In: *Remote Sensing of Environment* 120 (2012), pp. 25–36. DOI: 10.1016/j.rse.2011.11.026.
- [35] A. Edelman, T. A. Arias, and S. T. Smith. “The Geometry of Algorithms with Orthogonality Constraints.” In: *SIAM Journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353. DOI: 10.1137/S0895479895290954.
- [36] R. F. Egerton. *Electron Energy-Loss Spectroscopy in the Electron Microscope*. Third. Springer Science+Business Media, LLC, 2011. DOI: 10.1007/978-1-4419-9583-4.
- [37] S. Esedoglu and J. Shen. “Digital inpainting based on the Mumford–Shah–Euler image model.” In: *European Journal of Applied Mathematics* 13.4 (2002), pp. 353–370. DOI: 10.1017/S0956792502004904.
- [38] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. 1996, pp. 226–231. DOI: 10.5555/3001460.3001507.
- [39] M. Fornasier, R. March, and F. Solombrino. “Existence of minimizers of the Mumford-Shah functional with singular operators and unbounded data.” In: *Annali di Matematica Pura ed Applicata* 192.3 (2011), pp. 361–391. DOI: 10.1007/s10231-011-0228-8.
- [40] O. Forster. *Analysis 3*. Springer Fachmedien Wiesbaden, 2017. DOI: 10.1007/978-3-658-16746-2.

- [41] L. Gao, J. Li, M. Khodadadzadeh, A. Plaza, B. Zhang, Z. He, and H. Yan. “Subspace-Based Support Vector Machines for Hyperspectral Image Classification.” In: *IEEE Geoscience and Remote Sensing Letters* 12.2 (2015), pp. 349–353. DOI: 10.1109/lgrs.2014.2341044.
- [42] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarablaka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi, S. B. Serpico, and J. A. Benediktsson. “New Frontiers in Spectral-Spatial Hyperspectral Image Classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning.” In: *IEEE Geoscience and Remote Sensing Magazine* 6.3 (2018), pp. 10–43. DOI: 10.1109/MGRS.2018.2854840.
- [43] D. B. Gillis and J. H. Bowles. “Hyperspectral image segmentation using spatial-spectral graphs.” In: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*. Vol. 8390. International Society for Optics and Photonics. 2012, pp. 527–537. DOI: 10.1117/12.919743.
- [44] J. Gomez, J. Blasco, E. Molto, and G. Camps-Valls. “Hyperspectral detection of citrus damage with Mahalanobis kernel classifier.” In: *Electronics Letters* 43 (2007), pp. 1082–1084. DOI: 10.1049/e1:20070906.
- [45] F. A. Graybill. *Matrices with Applications in Statistics*. Second edition. The Wadsworth Statistics/Probability Series. Wadsworth International Group, 1983.
- [46] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. “A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal.” In: *IEEE Transactions on Geoscience and Remote Sensing* 26.1 (1988), pp. 65–74. DOI: 10.1109/36.3001.
- [47] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, and O. Williams. “Imaging Spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS).” In: *Remote Sensing of Environment* 65.3 (1998), pp. 227–248. DOI: 10.1016/S0034-4257(98)00064-9.
- [48] P. Green-Armytage. “A Colour Alphabet and the Limits of Colour Coding.” In: *JAIC - Journal of the International Colour Association* 5 (2010).
- [49] P. R. Griffiths and J. A. de Haseth. “Introduction to Vibrational Spectroscopy.” In: *Fourier Transform Infrared Spectrometry*. Second. 2007. Chap. 1, pp. 1–18. DOI: <https://doi.org/10.1002/9780470106310.ch1>.
- [50] S. Gupta, S. Mittal, A. Kajdacsy-Balla, R. Bhargava, and C. Bajaj. “A fully automated, faster noise rejection approach to increasing the analytical capability of chemical imaging for digital histopathology.” In: *PLOS ONE* 14.4 (2019), pp. 1–15. DOI: 10.1371/journal.pone.0205219.

- [51] A. Halim, A. Rohim, B. V. R. Kumar, and R. Saha. *A Fractional Image Inpainting Model Using a Variant of Mumford–Shah Model*. 2022. DOI: 10.48550/arXiv.2208.01887.
- [52] A. Hammer, J. Dumoulin, B. Vozel, and K. Chehdi. “Deblurring of UAV aerial images for civil structures inspections using Mumford-Shah / Total variation regularisation.” In: *2007 5th International Symposium on Image and Signal Processing and Analysis*. 2007, pp. 262–267. DOI: 10.1109/ISPA.2007.4383702.
- [53] A. Hassanzadeh, A. Kaarna, and T. Kauranne. “Sequential spectral clustering of hyperspectral remote sensing image over bipartite graph.” In: *Applied Soft Computing* 73 (2018), pp. 727–734. DOI: 10.1016/j.asoc.2018.09.015.
- [54] A. Hassanzadeh, A. Kaarna, and T. Kauranne. “Unsupervised Multi-manifold Classification of Hyperspectral Remote Sensing Images with Contractive Autoencoder.” In: *Image Analysis*. Vol. 10270. SCIA 2017. 2017, pp. 169–180. DOI: 10.1007/978-3-319-59129-2_15.
- [55] N. Henze and B. Zirkler. “A class of invariant consistent tests for multivariate normality.” In: *Communications in Statistics - Theory and Methods* 19.10 (1990), pp. 3595–3617. DOI: 10.1080/03610929008830400.
- [56] G. F. Hughes. “On the mean accuracy of statistical pattern recognizers.” In: *IEEE Transactions on Information Theory* 14.1 (1968), pp. 55–63. DOI: 10.1109/tit.1968.1054102.
- [57] M. Imani. “Difference-based target detection using Mahalanobis distance and spectral angle.” In: *International Journal of Remote Sensing* 40.3 (2018), pp. 811–831. DOI: 10.1080/01431161.2018.1519280.
- [58] M. Jung, X. Bresson, T. F. Chan, and L. A. Vese. “Nonlocal Mumford-Shah Regularizers for Color Image Restoration.” In: *IEEE Transactions on Image Processing* 20.6 (2011), pp. 1583–1598. DOI: 10.1109/TIP.2010.2092433.
- [59] J. P. Kerekes and J. R. Schott. “Hyperspectral Imaging Systems.” In: *Hyperspectral Data Exploitation*. 2007. Chap. 2, pp. 17–45. DOI: 10.1002/9780470124628.ch2.
- [60] H. W. Kuhn. “The Hungarian method for the assignment problem.” In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109.
- [61] B. Kunkel, F. Blechinger, D. Viehmann, H. van der Piepen, and R. Doerffer. “ROSIS imaging spectrometer and its potential for ocean parameter measurements (airborne and space-borne).” In: *International Journal of Remote Sensing* 12.4 (1991), pp. 753–761. DOI: 10.1080/01431169108929691.

- [62] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler. “Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 146 (2018), pp. 305–319. DOI: 10.1016/j.isprsjprs.2018.09.018.
- [63] J. B. Lee, A. S. Woodyatt, and M. Berman. “Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform.” In: *IEEE Transactions on Geoscience and Remote Sensing* 28.3 (1990), pp. 295–304. DOI: 10.1109/36.54356.
- [64] L. Li, C. Sun, L. Lin, J. Li, S. Jiang, and J. Yin. “A dual-kernel spectral-spatial classification approach for hyperspectral images based on Mahalanobis distance metric learning.” In: *Information Sciences* 429 (2018), pp. 260–283. DOI: <https://doi.org/10.1016/j.ins.2017.11.005>.
- [65] J. Liu and Y. Qiao. “Mahalanobis distance-based kernel supervised machine learning in spectral dimensionality reduction for hyperspectral imaging remote sensing.” In: *International Journal of Distributed Sensor Networks* 16.11 (2020). DOI: 10.1177/1550147720968467.
- [66] L. Liu, G. Delnevo, and S. Mirri. “Unsupervised hyperspectral image segmentation of films: a hierarchical clustering-based approach.” In: *Journal of Big Data* 10.1 (2023). DOI: 10.1186/s40537-023-00713-8.
- [67] Y. Liu, H. Pu, and D.-W. Sun. “Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications.” In: *Trends in Food Science & Technology* 69 (2017), pp. 25–35. DOI: 10.1016/j.tifs.2017.08.013.
- [68] G. Lu and B. Fei. “Medical hyperspectral imaging: a review.” In: *Journal of Biomedical Optics* 19.1 (2014), p. 010901. DOI: 10.1117/1.JBO.19.1.010901.
- [69] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Third edition. Wiley series in probability and statistics. Wiley, 2019. DOI: 10.1002/9781119541219.
- [70] P. C. Mahalanobis. “On the generalized distance in statistics.” In: *Proceedings of the National Institute of Sciences (Calcutta)* 2 (1936), pp. 49–55.
- [71] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
- [72] R. E. Megginson. *An Introduction to Banach Space Theory*. Graduate Texts in Mathematics. Springer New York, 1998. DOI: 10.1007/978-1-4612-0603-3.

- [73] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du. “Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 57.9 (2019), pp. 6808–6820. DOI: 10.1109/TGRS.2019.2908756.
- [74] L. Mou, P. Ghamisi, and X. X. Zhu. “Deep Recurrent Neural Networks for Hyperspectral Image Classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 55.7 (2017), pp. 3639–3655. DOI: 10.1109/tgrs.2016.2636241.
- [75] A. Mughees, X. Chen, and L. Tao. “Unsupervised hyperspectral image segmentation: Merging spectral and spatial information in boundary adjustment.” In: *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. 2016, pp. 1466–1471. DOI: 10.1109/SICE.2016.7749195.
- [76] D. Mumford and J. Shah. “Optimal approximations by piecewise smooth functions and associated variational problems.” In: *Communications on Pure and Applied Mathematics* 42.5 (1989), pp. 577–685. DOI: 10.1002/cpa.3160420503.
- [77] J. M. Murphy and M. Maggioni. “Unsupervised Clustering and Active Learning of Hyperspectral Images With Nonlinear Diffusion.” In: *IEEE Transactions on Geoscience and Remote Sensing* 57.3 (2019), pp. 1829–1845. DOI: 10.1109/TGRS.2018.2869723.
- [78] J. Nalepa, M. Myller, Y. Imai, K.-I. Honda, T. Takeda, and M. Antoniuk. “Unsupervised Segmentation of Hyperspectral Images Using 3-D Convolutional Autoencoders.” In: *IEEE Geoscience and Remote Sensing Letters* 17.11 (2020), pp. 1948–1952. DOI: 10.1109/LGRS.2019.2960945.
- [79] J. Nalepa, M. Myller, and M. Kawulok. “Validating Hyperspectral Image Segmentation.” In: *IEEE Geoscience and Remote Sensing Letters* 16.8 (2019), pp. 1264–1268. DOI: 10.1109/LGRS.2019.2895697.
- [80] A. A. Nielsen. “An Extension to a Filter Implementation of a Local Quadratic Surface for Image Noise Estimation.” In: *10th International Conference on Image Analysis and Processing, ICIAP’99*. 1999, pp. 119–124.
- [81] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2006. DOI: 10.1007/978-0-387-40065-5.
- [82] Norwegian Meteorological Institute, Ice Service. *Barents ice chart on 4th April 2018*. 2018. URL: https://cryo.met.no/archive/ice-service/icecharts/quicklooks/2018/20180404/barent_20180404_col.png (visited on 09/08/2024).

- [83] A. Obeid, I. M. Elfadel, and N. Werghi. “Unsupervised Land-Cover Segmentation Using Accelerated Balanced Deep Embedded Clustering.” In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022), pp. 1–5. DOI: 10.1109/LGRS.2021.3079710.
- [84] J. M. Ortega. *Numerical analysis: a second course*. Classics in applied mathematics ; 3. Society for Industrial and Applied Mathematics, 1990. DOI: 10.1137/1.9781611971323.
- [85] R. Palmieri, R. Gasbarrone, and L. Fiore. “Hyperspectral Imaging for Sustainable Waste Recycling.” In: *Sustainability* 15.10 (2023). DOI: 10.3390/su15107752.
- [86] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza. “Deep learning classifiers for hyperspectral imaging: A review.” In: *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), pp. 279–317. DOI: 10.1016/j.isprsjprs.2019.09.006.
- [87] K. R. Parthasarathy and K. Schmidt. *Positive Definite Kernels, Continuous Tensor Products, and Central Limit Theorems of Probability Theory*. Ed. by A. Dold and B. Eckmann. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 1972. DOI: 10.1007/bfb0058340.
- [88] K. Pearson. “LIII. On lines and planes of closest fit to systems of points in space.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (6 1901), pp. 559–572. DOI: 10.1080/14786440109462720.
- [89] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. 2012.
- [90] J. Qin, O. Monje, M. R. Nugent, J. R. Finn, A. E. O’Rourke, K. D. Wilson, R. F. Fritsche, I. Baek, D. E. Chan, and M. S. Kim. “A hyperspectral plant health monitoring system for space crop production.” In: *Frontiers in Plant Science* 14 (2023). DOI: 10.3389/fpls.2023.1133505.
- [91] R. Ramlau and W. Ring. “A Mumford–Shah level-set approach for the inversion and segmentation of X-ray tomography data.” In: *Journal of Computational Physics* 221.2 (2007), pp. 539–557. DOI: 10.1016/j.jcp.2006.06.041.
- [92] W. Rawat and Z. Wang. “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review.” In: *Neural Computation* 29.9 (2017), pp. 2352–2449. DOI: 10.1162/neco_a_00990.
- [93] J. A. Richards and X. Jia. *Remote sensing digital image analysis : an introduction*. Fourth. Springer-Verlag Berlin Heidelberg, 2006. DOI: 10.1007/3-540-29711-1.

- [94] L. J. Rickard, R. W. Basedow, E. F. Zalewski, P. R. Silverglate, and M. Landers. “HYDICE: an airborne system for hyperspectral imaging.” In: *Imaging Spectrometry of the Terrestrial Environment*. Vol. 1937. International Society for Optics and Photonics. 1993, pp. 173–179. DOI: 10.1117/12.157055.
- [95] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1997.
- [96] L. Rondi and F. Santosa. “Enhanced electrical impedance tomography via the Mumford–Shah functional.” In: *ESAIM: Control, Optimisation and Calculus of Variations* 6 (2001), pp. 517–538.
- [97] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri. “HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification.” In: *IEEE Geoscience and Remote Sensing Letters* 17.2 (2020), pp. 277–281. DOI: 10.1109/lgrs.2019.2918719.
- [98] S. K. Roy, S. Manna, T. Song, and L. Bruzzone. “Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 59.9 (2021), pp. 7831–7843. DOI: 10.1109/tgrs.2020.3043267.
- [99] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [100] B. M. Shahshahani and D. A. Landgrebe. “The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon.” In: *IEEE Transactions on Geoscience and Remote Sensing* 32.5 (1994), pp. 1087–1095. DOI: 10.1109/36.312897.
- [101] C. Shi, J. Sun, and L. Wang. “Hyperspectral Image Classification Based on Spectral Multiscale Convolutional Neural Network.” In: *Remote Sensing* 14.8 (2022), p. 1951. DOI: 10.3390/rs14081951.
- [102] W. Sun and Q. Du. “Hyperspectral Band Selection: A Review.” In: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), pp. 118–139. DOI: 10.1109/MGRS.2019.2911100.
- [103] A. Tsai, A. Yezzi, and A. S. Willsky. “Curve evolution implementation of the Mumford–Shah functional for image segmentation, denoising, interpolation, and magnification.” In: *IEEE Transactions on Image Processing* 10.8 (2001), pp. 1169–1186. DOI: 10.1109/83.935033.
- [104] Q. Wang, Q. Li, H. Liu, Y. Wang, and J. Zhu. “An improved ISODATA algorithm for hyperspectral image classification.” In: *2014 7th International Congress on Image and Signal Processing*. 2014. DOI: 10.1109/cisp.2014.7003861.

- [105] R. Wang, F. Nie, and W. Yu. “Fast Spectral Clustering With Anchor Graph for Large Hyperspectral Images.” In: *IEEE Geoscience and Remote Sensing Letters* 14.11 (2017), pp. 2003–2007. DOI: 10.1109/LGRS.2017.2746625.
- [106] W. Wang and M. Á. Carreira-Perpiñán. *Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application*. 2013. DOI: 10.48550/ARXIV.1309.1541.
- [107] Z. Wen and W. Yin. “A feasible method for optimization with orthogonality constraints.” In: *Mathematical Programming* 142.1–2 (2012), pp. 397–434. DOI: 10.1007/s10107-012-0584-1.
- [108] J. Xie, R. Girshick, and A. Farhadi. “Unsupervised Deep Embedding for Clustering Analysis.” In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. Proceedings of Machine Learning Research. 2016, pp. 478–487.
- [109] Z. Xue, M. Zhang, Y. Liu, and P. Du. “Attention-Based Second-Order Pooling Network for Hyperspectral Image Classification.” In: *IEEE Transactions on Geoscience and Remote Sensing* 59.11 (2021), pp. 9600–9615. DOI: 10.1109/tgrs.2020.3048128.
- [110] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer. “Fast Global Labeling for Real-Time Stereo Using Multiple Plane Sweeps.” In: *Proceedings of the Vision, Modeling, and Visualization Conference 2008, VMV 2008*. 2008, pp. 243–252.
- [111] H. Zhang, H. Zhai, L. Zhang, and P. Li. “Spectral-Spatial Sparse Subspace Clustering for Hyperspectral Remote Sensing Images.” In: *IEEE Transactions on Geoscience and Remote Sensing* 54.6 (2016), pp. 3672–3684. DOI: 10.1109/TGRS.2016.2524557.
- [112] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao. “Hyperspectral image unsupervised classification by robust manifold matrix factorization.” In: *Information Sciences* 485 (2019), pp. 154–169. DOI: 10.1016/j.ins.2019.02.008.
- [113] Q. Zhang and N. Hughes. “Ice floe segmentation and floe size distribution in airborne and high-resolution optical satellite images: towards an automated labelling deep learning approach.” In: *The Cryosphere* 17.12 (2023), pp. 5519–5537. DOI: 10.5194/tc-17-5519-2023.
- [114] Y. Zhang, B. Du, L. Zhang, and S. Wang. “A low-rank and sparse matrix decomposition-based Mahalanobis distance method for hyperspectral anomaly detection.” In: *IEEE Transactions on Geoscience and Remote Sensing* 54.3 (2016), pp. 1376–1389. DOI: 10.1109/TGRS.2015.2479299.

- [115] Y. Zhao, Y. Yuan, and Q. Wang. “Fast Spectral Clustering for Unsupervised Hyperspectral Image Classification.” In: *Remote Sensing* 11.4 (2019). DOI: 10.3390/rs11040399.
- [116] W. Zhu, V. Chayes, A. Tiard, S. Sanchez, D. Dahlberg, A. L. Bertozzi, S. Osher, D. Zosso, and D. Kuang. “Unsupervised Classification in Hyperspectral Imagery With Nonlocal Total Variation and Primal-Dual Hybrid Gradient Algorithm.” In: *IEEE Transactions on Geoscience and Remote Sensing* 55.5 (2017), pp. 2786–2798. DOI: 10.1109/tgrs.2017.2654486.

List of Abbreviations

3D-CAE	3D convolutional autoencoder
AVIRIS	airborne visible/infrared imaging spectrometer
AA	average class accuracy
BGM	Bayesian Gaussian mixture model
BV	bounded variation
CIK	cross-information kernel
DMI	Danish Meteorological Institute
DSen2	Deep Sentinel-2
DSK	direct summation kernel
ϵ AMS	ϵ -regularized anisotropic Mumford–Shah model
EELS	electron energy-loss spectroscopy
FL	feature learning
FT-IR	Fourier-transform infrared spectroscopy
GMM	Gaussian mixture model
HYDICE	hyperspectral digital imagery collection experiment
IP	Indian Pines
KC	kappa coefficient
KSC	Kennedy Space Center
kMS	kernelized Mumford–Shah model
MRI	magnetic resonance imaging
MNF	minimum noise fraction transform
MS	Mumford–Shah
MS-2	MS model with squared 2-norm as indicator function

MIR	mid-infrared
NIR	near-infrared
NLTV	MS model with a non-local TV regularizer
MET	Norwegian Meteorological Institute
OA	overall accuracy
PU	Pavia University
PCA	principal components analysis
PC	principal component
PDHG	primal-dual hybrid gradient
ϕ AMS	ϕ -regularized anisotropic Mumford–Shah model
RBF	radial basis function
ROSIS	reflective optics system imaging spectrometer
SA	Salinas
SSC	sequential spectral clustering with RBF kernel
SNR	signal-to-noise ratio
SVM	support vector machine
TV	total variation
WSK	weighted summation kernel

List of Symbols

Thesis-specific symbols

d	Dimension of image domain
Ω	Continuous image domain, a d -dimensional bounded domain with Lipschitz boundary
$\mathcal{N}$	Discrete image domain, a d -dimensional finite regular grid
N_i	Grid size of discrete image domain in component i
g	Image on a continuous or discrete domain
L	Number of channels of an image (dimension of value range)
W	Value range of an image, subset of $\mathbb{R}^L$
k	Number of segments
J_{MS}	MS segmentation functional (cf. (2.9))
J_ϵ	Functional of ϵ AMS (cf. (2.20))
J_ϕ	Functional of ϕ AMS (cf. (2.46) and (2.47))
J_Φ	Functional of kMS (cf. (2.55))
$f_{\eta,l}$	Indicator function of segment l in J_ϵ of ϵ AMS (cf. (2.16))
$f_{\phi,l}$	Indicator function of segment l in J_ϕ of ϕ AMS (cf. (2.48))
$f_{\Phi,l}$	Indicator function of segment l in J_Φ of kMS (cf. (2.54))
E	General notation for discretized and relaxed objective functional (cf. (3.9))
E_ϵ	Discretized and relaxed objective functional of ϵ AMS (cf. (3.19))

E_ϕ	Discretized and relaxed objective functional of ϕ AMS (cf. (3.23))
E_Φ	Discretized and relaxed objective functional of kMS (cf. (3.38))
$\mathcal{A}$	Set of all partitions of Ω into k measurable subsets with finite perimeter in Ω (cf. (2.7))
$\mathcal{P}$	Set of symmetric and positive definite matrices (cf. (2.13))
$\mathcal{P}_\epsilon$	Set of symmetric and positive definite matrices with eigenvalues greater or equal to ϵ^2 (cf. (2.19))
$\text{Per}(\mathcal{O}, \Omega)$	Perimeter of $\mathcal{O}$ in Ω
λ	Regularization parameter balancing the data term and the perimeter in MS-based segmentation functionals
ϵ	Regularization parameter for positive definiteness of covariance matrices in ϵ AMS
η	Regularization parameter for differentiability of J_ϵ
$\Sigma_{\epsilon,l}$	Covariance matrix of segment l regularized with ϵ in J_ϵ
$\ \cdot\ _{\Sigma_l^{-1}, \eta}$	Square root term in indicator function $f_{\eta,l}$ of J_ϵ (cf. (2.17))
σ^2	Operator returning the eigenvalues of a symmetric matrix (cf. (2.18)), or vector of eigenvalues of covariance matrix
ϕ	Regularization function for positive definiteness of covariance matrices in and differentiability of J_ϕ (cf. (2.44) and (2.45))
Φ	General notation for a feature map in kMS
K	General notation for a kernel in kMS
u	Labeling function (in the context of optimization in Chapter 3)
$m^{\max}$	Maximum number of outer iterations
$m_{\mu, \Sigma}^{\max}$	Maximum number of fixed point iterations in ϵ AMS
$m_{\text{PDHG}}^{\max}$	Maximum number of PDHG iterations

Derivatives

$\partial_{x_i} f(x), \frac{\partial f(x)}{\partial x_i}$	Partial derivative of a function f with respect to the variable x_i at a position x
---	---

$\nabla f(x)$	Gradient of a real-valued function f at a position x
$\operatorname{div} f(x)$	Divergence of a function f at a position x (cf. (1.2))
$\partial f(x)$	Subdifferential of a function f at a position x (cf. Definition A.2)

Groups, manifolds, sets and spaces

$\mathbb{N}$	Natural numbers $(\{1, 2, 3, \dots\})$
$\mathbb{N}_0$	Natural numbers including 0
$\mathbb{R}$	Real numbers
$\mathbb{R}_{>0}$	Positive real numbers
$\overline{\mathbb{R}}$	Real numbers extended with $-\infty$ and ∞ , i.e., $\mathbb{R} \cup \{-\infty, \infty\}$
$O(n)$	Orthogonal group in dimension n
$\mathcal{M}_n^p$	Stiefel manifold (p -tuples of n -dimensional orthonormal vectors) (cf. (3.28))
$C^m(\Omega), C^m(\Omega, \mathbb{R}^n)$	Space of m -times continuously differentiable functions on Ω with values in $\mathbb{R}$ and $\mathbb{R}^n$, resp. (cf. Definition A.9)
$C_c^m(\Omega), C_c^m(\Omega, \mathbb{R}^n)$	Space of m -times continuously differentiable functions with compact support in Ω and values in $\mathbb{R}$ and $\mathbb{R}^n$, resp. (cf. Definition A.9)
$L^p(\Omega), L^p(\Omega, \mathbb{R}^n)$	L^p -space or Lebesgue space with $1 \leq p \leq \infty$ on Ω with values in $\mathbb{R}$ and $\mathbb{R}^n$, resp. (cf. Definition A.10)
$H^{m,p}(\Omega)$	Sobolev space with $1 \leq p \leq \infty$ and $m \in \mathbb{N}$ (cf. Definition A.11)
$BV(\Omega)$	Space of functions of bounded variation on Ω (cf. Definition 2.1)

Norms and semi-norms

$ a $	Absolute value of $a \in \mathbb{R}$
$\ \cdot\ $	General notation for a norm
$\ \cdot\ _p$	p -norm for $1 \leq p \leq \infty$ on finite-dimensional spaces
$\ \cdot\ _F$	Frobenius norm of a matrix
$\ \cdot\ _{L^p(\Omega)}, \ \cdot\ _{L^p(\Omega, \mathbb{R}^n)}$	Norm on $L^p(\Omega)$ and $L^p(\Omega, \mathbb{R}^n)$, resp. (cf. Definition A.10)

$\ \cdot\ _{H^{m,p}(\Omega)}$	Norm on $H^{m,p}(\Omega)$ (cf. Definition A.11)
$ \cdot _{\text{BV}(\Omega)}$	Total variation of a function (cf. Definition 2.1)
$ \cdot _{\text{BV}(\mathcal{N})_h}$	Discretized total variation of a function (cf. (3.8))
$\ \cdot\ _{\text{BV}(\Omega)}$	Norm on $\text{BV}(\Omega)$ (cf. Definition 2.1)
$\ \cdot\ _{2 \rightarrow 2}$	Spectral norm of a matrix
$\ \cdot\ _{X \rightarrow Y}$	Operator norm for an operator mapping from X to Y (cf. (3.10))

Other symbols

∂A	Boundary of a set A
$\overline{A}$	Closure of a set A
$A^\perp$	Orthogonal complement of a set A
$\text{span } A$	Span of a set A
$A \triangle B$	Symmetric difference of two sets A and B
$\mathcal{H}^{n-1}$	$(n-1)$ -dimensional Hausdorff measure
$ A $	n -dimensional Lebesgue measure of a measurable set $A \subseteq \mathbb{R}^n$
χ_A	Characteristic function of a set A (cf. (1.1))
δ_A	Indicator functional of a set A (cf. Remark 2.24)
v^T, M^T	Transpose of a vector v or a matrix M
e_i	i -th canonical unit vector
$\text{tr } M$	Trace of a matrix M
I	Identity matrix
$\mathcal{H}_1 \oplus \mathcal{H}_2$	Direct sum of two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$
$\langle \cdot, \cdot \rangle$	General notation for an inner product
$B_r^{\ \cdot\ }(x)$	Open ball with radius $r > 0$ with respect to norm $\ \cdot\ $ around x
$\mathcal{T}_X \mathcal{M}$	Tangent space of a manifold $\mathcal{M}$ at $X \in \mathcal{M}$
$B^\dagger$	Adjoint operator of a linear operator B

$f _A$	Restriction of a function f to a subset A of its domain
$\text{supp } f$	Support of a function f (cf. Definition A.9)
$\text{dom}(f)$	Effective domain of a function f with values in $\mathbb{R} \cup \{\infty\}$ (cf. Definition 3.2)
$\text{epi}(f)$	Epigraph of a function f with values in $\mathbb{R} \cup \{\infty\}$ (cf. Definition 3.2)
prox_f	Proximal operator of a function f (cf. Definition 3.5)
f^*	(In Chapter 3) Fenchel conjugate of a function f (cf. Definition 3.10)
$\Gamma_0(X)$	Set of proper, closed and convex functions on a vector space X (cf. Definition 3.2)
$\rightarrow, \rightharpoonup, \overset{*}{\rightharpoonup}$	Strong, weak and weak* convergence
$\mathbb{K}_k$	k -dimensional unit simplex (cf. (3.1))
L_0	Operator norm of linear operator in PDHG optimization problem (cf. Proposition 3.14)
τ, κ	Step sizes in PDHG method
ϑ	Scaling factor in extrapolation step in PDHG method

List of Figures

1.1	Visualization of spectral bands used by a typical RGB camera, the Sentinel-2 multispectral sensor and the AVIRIS hyperspectral sensor.	3
1.2	Exemplary FT-IR image; data courtesy of Rohit Bhargava (University of Illinois Urbana Champaign).	5
1.3	Distribution of spectra from one tissue type in the FT-IR image; data courtesy of Rohit Bhargava (University of Illinois Urbana Champaign).	6
1.4	An exemplary RGB image showing a red box on a wooden surface with a gray background, a segmentation of it into three segments and the corresponding spectral clusters.	10
2.1	Sampled spectra from an image g with a two-dimensional value range ($L = 2$) and two spectral clusters. The arrows depict the principal components of the spectral clusters, scaled by their standard deviations.	34
2.2	Samples from a spectral distribution corresponding to a segment $\mathcal{O}$ of an image g with a two-dimensional value range ($L = 2$) and the means computed with respect to the squared $\ \cdot\ _2^2$ and non-squared $\ \cdot\ _2$ Euclidean norm.	36
2.3	Sampled spectra from two spectral distributions $g^m _{\mathcal{O}_l}$ of a segment $\mathcal{O}_l \subseteq \Omega$ corresponding to images $g^m \in L^\infty(\Omega, \mathbb{R}^L)$, $m \in \{1, 2\}$, projected onto the i -th and j -th eigenvector $v_{l,i}^m \in \mathbb{R}^L$ and $v_{l,j}^m \in \mathbb{R}^L$ of the covariance matrix $\Sigma_l^m \in \mathbb{R}^{L \times L}$.	70
4.1	The IP dataset.	133
4.2	The KSC dataset.	134
4.3	The PU dataset.	135
4.4	The SA dataset.	136
4.5	Ground truth and segmentations of raw IP data produced by the competing methods.	152
4.6	Ground truth and segmentations of MNF-transformed IP data produced by the competing methods.	153

4.7	Ground truth and segmentations of the learned FL features of the IP data produced by the competing methods.	154
4.8	Ground truth and segmentations of raw KSC data produced by the competing methods.	156
4.9	Ground truth and segmentations of MNF-transformed KSC data produced by the competing methods.	157
4.10	Ground truth and segmentations of raw PU data produced by the competing methods.	158
4.11	Ground truth and segmentations of MNF-transformed PU data produced by the competing methods.	159
4.12	Ground truth and segmentations of the learned FL features of the PU data produced by the competing methods.	160
4.13	Ground truth and segmentations of raw SA data produced by the competing methods.	161
4.14	Ground truth and segmentations of MNF-transformed SA data produced by the competing methods.	162
4.15	Ground truth and segmentations of the learned FL features of the SA data produced by the competing methods.	163
4.16	Henze–Zirkler statistics for each dataset and each segment.	167
4.17	Plots of functional values and stopping criterion against outer iterations of ϵ AMS on raw IP.	168
4.18	Plots of functional values and stopping criterion against outer iterations of ϕ AMS on raw IP.	168
4.19	Plots of functional values and stopping criterion against outer iterations of kMS-DSK on raw IP.	168
4.20	Plots of functional values and stopping criterion against outer iterations of kMS-WSK on raw IP.	169
4.21	Plots of functional values and stopping criterion against outer iterations of kMS-CIK on raw IP.	169
4.22	RGB versions of the two considered Sentinel-2 images and the D1 crop (region of D1 covered by the crop marked in D1 with a red box). . .	172
4.23	Segmentation results of D1 with $\lambda = 5 \cdot 10^{-5}$	173
4.24	Segmentation results of D1 crop with $\lambda = 5 \cdot 10^{-5}$	174
4.25	Segmentation results of D2 with $\lambda = 5 \cdot 10^{-5}$	174
4.26	Comparison of the segmentation result of D1 crop and the corresponding crop of the full D1 segmentation with $\lambda = 5 \cdot 10^{-5}$ and $k = 5$. . .	176

List of Tables

4.1	Image-independent parameters of ϵ AMS.	143
4.2	Image-independent parameters of ϕ AMS.	144
4.3	Image-independent parameters of kMS.	145
4.4	Choices of image-dependent model parameters.	146
4.5	OA of the competing methods on the four datasets.	148
4.6	AA of the competing methods on the four datasets.	149
4.7	KC of the competing methods on the four datasets.	150
4.8	Comparison of estimated sea ice concentrations in D1 and D2.	175
4.9	Agreement scores of the segmentation of the D1 crop and the crop of the full D1 segmentation.	177

Eidesstattliche Erklärung

Ich, Jan-Christopher Cohrs, erkläre hiermit, dass diese Dissertation und die darin dargelegten Inhalte die eigenen sind und selbstständig, als Ergebnis der eigenen originären Forschung, generiert wurden.

Hiermit erkläre ich an Eides statt

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorand dieser Fakultät und Universität angefertigt;
2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;
3. Wenn immer andere eigene oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;
4. Wenn aus anderen eigenen oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Diese Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;
5. Alle wesentlichen Quellen von Unterstützung wurden benannt;
6. Wenn immer ein Teil dieser Dissertation auf der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;
7. Ein Teil oder Teile dieser Arbeit wurden zuvor veröffentlicht und zwar in:

- [1] Jan-Christopher Cohrs, Chandrajit Bajaj und Benjamin Berkels. „A Distribution-Dependent Mumford–Shah Model for Unsupervised Hyperspectral Image Segmentation“. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), S. 1–21. DOI: 10.1109/TGRS.2022.3227061.
- [2] Jan-Christopher Cohrs und Benjamin Berkels. „On the importance of the ε -regularization of the distribution-dependent Mumford–Shah model for hyperspectral image segmentation“. In: *Multiscale, Nonlinear and Adaptive Approximation II*. Hrsg. von Ronald DeVore und Angela Kunoth. Cham: Springer Nature Switzerland, 2024, S. 201–222. DOI: 10.1007/978-3-031-75802-7_10.
- [3] Jan-Christopher Cohrs, Ekaterina Kim und Benjamin Berkels. „Sea-Ice States: Unsupervised Segmentation from Multispectral Data Based on Mumford–Shah Functional“. In: *Proceedings of the 27th IAHR International Symposium on Ice*. Hrsg. von Tomasz Kolerski. IAHR, 2024.
