University of Koblenz-Landau

Faculty 4: Computer Science

Nuclear Potential Clustering of Data

Determination of Image-based Biomarkers for Multiple Sclerosis

by

Dipl.-Ing. (FH) Vyara Tonkova

Approved Dissertation thesis for the partial fulfillment of the requirements for a Doctor of Natural Sciences (Dr. rer. nat.)

Koblenz - 2019

Chair of PhD Board:	Prof. Dr. Maria A. Wimmer
Chair of PhD Commission:	Prof. Dr. Petra Schubert
Examiner and Supervisor:	Prof. Dr. Dietrich Paulus
Further Examiners:	Prof. Dr. Heiko Neeb
Date of the doctoral viva:	31.07.2019

"Anything created by human beings
is already in the great book of nature."

ANTONI GAUDI

Abstract

Initial goal of the current dissertation was the determination of image-based biomarkers sensitive for neurodegenerative processes in the human brain. One such process is the demyelination of neural cells characteristic for Multiple sclerosis (MS) - the most common neurological disease in young adults for which there is no cure yet. Conventional MRI techniques are very effective in localizing areas of brain tissue damage and are thus a reliable tool for the initial MS diagnosis. However, a mismatch between the clinical findings and the visualized areas of damage is observed, which renders the use of the standard MRI difficult for the objective disease monitoring and therapy evaluation. To address this problem, a novel algorithm for the fast mapping of myelin water content using standard multiecho gradient echo acquisitions of the human brain is developed in the current work. The method extents a previously published approach for the simultaneous measurement of brain T_1, T_2^* and total water content. Employing the multiexponential T_2^* decay signal of myelinated tissue, myelin water content is measured based on the quantification of two water pools (myelin water and rest) with different relaxation times. Whole brain *in vivo* myelin water content maps are acquired in 10 healthy controls and one subject with MS. The *in vivo* results obtained are consistent with previous reports.

The acquired quantitative data have a high potential in the context of MS. However, the parameters estimated in a multiparametric acquisition are correlated and constitute therefore an ill-posed, nontrivial data analysis problem. Motivated by this specific problem, a new data clustering approach is developed called Nuclear Potential Clustering, NPC. It is suitable for the explorative analysis of arbitrary dimensional and possibly correlated data without *a priori* assumptions about its structure. The developed algorithm is based on a concept adapted from nuclear physics. To partition the data, the dynamic behavior of electrically even charged nucleons interacting in a d-dimensional feature space is modeled. An adaptive nuclear potential, comprised of a short-range attractive (Strong interaction) and a long-range repulsive term (Coulomb potential), is assigned to each data point. Thus, nucleons that are densely distributed in space fuse to build nuclei (clusters), whereas single point clusters are repelled (noise). The algorithm is optimized and tested in an extensive study with a series of synthetic datasets as well as the Iris data. The results show that it can robustly identify clusters even when complex configurations and noise are present.

Finally, to address the initial goal, quantitative MRI data of 42 patients are analyzed employing NPC. A series of experiments with different sets of image-based features show a consistent grouping tendency: younger patients with low disease grade are recognized as cohesive clusters, while those of higher age and impairment are recognized as outliers. This allows for the definition of a reference region in a feature space associated with phenotypic data. Tracking of the individual's positions therein can disclose patients at risk and be employed for therapy evaluation.

Zusammenfassung

Initiales Ziel der vorliegenden Arbeit ist die Definition von Biomarkern für die sensitive Diagnose neurodegenerativer Prozesse im menschlichen Gehirn. Ein solcher Prozess, bei dem die isolierende Nervenzellenschicht Myelin zerstört wird, ist charakteristisch für Multiple Sklerose, MS. Sie ist die häufigste neurologische Erkrankung der jungen Bevölkerung, die einen chronischen Verlauf hat und derzeit noch nicht heilbar ist. Hier ermöglicht die konventionelle MRT eine sensitive Erstdiagnose, hat dennoch eine eingeschränkte Aussagekraft darüber wie sich die Krankheit weiter entwickelt. Um eine objektive Verlaufskontrolle und Therapiebewertung zu unterstützen, wurde in dieser Arbeit ein neuartiges MRT Verfahren entwickelt. Es erlaubt die Quantifizierung vom Myelin-Wassergehalt mit konventionellen Multiecho-Gradientenecho-Sequenzen. Die Methode basiert auf einem bereits publizierten Verfahren zur schnellen simultanen Quantifizierung von T_1, T_2^* und absolutem Wassergehalt. Der Ansatz nutzt den multiexponentiellen T_2^*-Zerfall im myelinisierten Gewebe aus, um basierend auf einer Zweikompartment-Analyse den myelingebundenen Wasseranteil zu messen. Es wurden Myelinwasserkarten von 10 gesunden Probanden und einem MS Patienten angefertigt. Die Ergebnisse sind konsistent mit Literaturangaben.

Die multimodalen quantitativen Daten haben ein hohes Potential für die MS Diagnostik, sind aber biophysikalisch miteinander korreliert. Dadurch ist der Einsatz von Standardtools der Datenanalyse hier erschwert. Diese Problematik motivierte die Entwicklung eines neuen Verfahrens zur Gruppierung zusammengehöriger Datenpunkte. Ziel dabei ist das Clustering von möglicherweise korrelierten Daten beliebiger Dimensionalität ohne *a priori* Information über die zugrundeliegende Datenstruktur. Das entwickelte Nuclear Potential Clustering, NPC, basiert auf der Simulation der Dynamik gleichgeladener Nukleonen in einem d-dimensionalen Merkmalsraum. Jeder Datenpunkt hat dabei ein Potential, das anziehend für geringe Reichweiten wirkt (Starke Wechselwirkung) und abstoßend für lange Reichweiten (Elektromagnetische Wechselwirkung). Bei der Simulation eines solchen Systems fusionieren naheliegende Nukleonen zu Kernen (Cluster), während einzeln liegende Punkte abgestoßen werden (Ausreißer). Die Methode wurde in einer umfassenden Studie optimiert und mit verschiedenen Testsamples sowie den Iris Datensatz evaluiert. Die Ergebnisse zeigen, dass NPC sehr zuverlässig Ausreißer sowie Cluster beliebiger Form detektiert.

Schlussendlich wurde NPC verwendet, um quantitative MRT-Daten von 42 MS Patienten zu analysieren. Es wurde eine Reihe von Experimenten mit verschiedenen bildbasierten Merkmalen durchgeführt. NPC erkennt dabei konsistent als Cluster jüngere Patienten mit wenigen bzw. keinen neurologischen Ausfällen, während Personen mit einem hohen Schweregrad der Behinderung als Ausreißer erkannt werden. Dies erlaubt die Definition einer Referenz-Region im Merkmalsraum. Durch Tracking der individuellen Bewegung eines Patienten in diesem Raum können Risikopatienten identifiziert und eingesetzte Therapien evaluiert werden.

Acknowledgments

First and foremost I want to thank my advisor Prof. Dr. Heiko Neeb for being the driving force behind this project. Being his first PhD student in Remagen was an honor to me and I truly appreciate all of the inspirational ideas, fruitful discussions and his constant support in any occasion. His enthusiastic way of doing research work as well as his inevitable positive attitude were extremely supportive especially in tough times. In this regard, I truly gained a new vision for live as I learned being generally more positive from his contagious candid optimism.

I would also like to thank Prof. Dr.-Ing. Dietrich Paulus for the opportunity to realize this dissertation at the University of Koblenz Landau. Having his support has been essential for the success of this work.

It was a pleasure to me working at RheiAhrCampus, University of Applied Sciences Koblenz with all of my colleagues. A special thank goes to Marie Sophie Lafontaine for her intensive support im my research and teaching work and also for being a wonderful friend. For the great teamwork I want to thank the members of the Multimodal Imaging Physics Team, Anna Peneva, Markus Rausch, Svenja Grieger, Felix Schyboll and Jasmin Taha, as well as the colleagues from the Department of Mathematics and Engineering, deliberately in random order: Dirk Thomsen, Volker Luy, Tobias Bildhauer, Prof. Dr. Dietrich Holz, Prof. Dr. Uwe Jaekel, Prof. Dr. Sönke Carstens-Behrens, Prof. Dr. Marco Junglas, Prof. Dr. Jens Bongartz, Alexander Jenal, Caspar Kneer and Immanuel Weber. All of the exciting conversations and the wonderful times we were laughing and exchanging sometimes crazy ideas turned the work into a great experience.

To Dagmar Lang and Delia-Lisa Feis I want to thank for providing me invaluable professional and emotional support in assembling this thesis. We have started our PhD journeys together and it was really a pleasure to me having them as friends and colleagues on my side.

To my dearest friends May Klein and Nevena Varbanova I want to thank for motivating me to start working in the scientific field and constantly encouraging me during the last years.

I am also deeply grateful for the support of my parents, my sister and my dearest husband Pencho Stefanov without whom I would have never succeeded in completing this doctoral thesis.

Last but not least, I want to thank Tania Bentley for being such an inspiring person. Tania, I do realize that facing and dealing with MS personally is something completely different from attempting to "fight against" it as a scientist. I admire your strength and ability to look towards the future with so much hope despite the confrontation with the MS uncertainty. And though I am sure you know it, I want to assure you once again that there is a whole army of scientists who are passionately involved in supporting every person encountering MS. Keep on fighting and do never forget that your are not alone!

Contents

Chapter 1

Preface

1.1 Motivation

Contemporary scientific research often starts with a concrete problem and evolves in the course of work to a truly interdisciplinary challenge. And this is exactly how the current doctoral thesis has come about. The main issue of this work is the development of a novel data mining tool called Nuclear Potential Clustering, short NPC. In the beginning however, the focus was on a seemingly pure medical problem: the development and validation of image-derived biomarkers sensitive to multiple sclerosis (MS). Here, the digital nature of medical imaging, which has a high potential of improving the diagnosis and monitoring of this disease, brings along two main aspects that have to be concerned. On the one side, novel clinical examination techniques are constantly being developed, that allow us to gain new insights into the pathophysiologic mechanism of the disease. On the other side, the continually growing amount of novel medical data has ideally to be turned into new objective medical knowledge. Currently, a major problem in the context of multiple sclerosis is the observed mismatch between the standard magnetic resonance imaging (MRI) and the clinical findings of the patients. To work this problem out, two research topics are being increasingly investigated:

- new techniques for brain imaging based on multiparametric quantitative MRI, and
- automated multidimensional image analysis to provide an observer independent analysis.

In the latter, cluster analysis is one of the standard methods to "blindly" group data into different categories (this idea is extensively explained in Chapter 3). In the context of MS, such categories could be patients with different disease grade or, more importantly, patients responding differently to a given therapy. However, most of the MRI parameters in a multiparametric acquisition are correlated and constitute therefore an ill-posed, nontrivial clustering problem. Motivated by this specific problem, an idea for a new clustering approach suitable for the explorative analysis of high-dimensional and possibly correlated data is born. Its basic idea is to imitate a physical system in which clusters appear naturally: the process of nuclear fusion where nearby nucleons fuse to build stable nuclei, hence the name of the algorithm "Nuclear Potential Clustering". The development of this new concept has evolved to a major challenge itself and became ultimately a main subject of this thesis, as well.

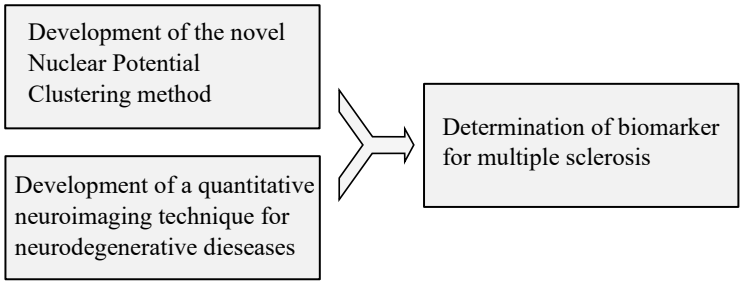


Figure 1.1: Main research topics of the current work.

1.2 Outline of the Work

The current thesis is structured in a way that corresponds to the chronological progress of the presented scientific work.

It begins with an introductory part concerning the motivating medical imaging aspect. Concretely, in Chapter 2, Sec. 2.1, the challenges of the diagnosis and monitoring of multiple sclerosis are introduced, with the stress put on the state of the art neuroimaging techniques of the disease. Here, the conventional clinical MRI examinations are presented, along with a quantitative MRI technique developed by [Neeb 2008]. The latter is a multimodal approach, that allows for the simultaneous calculation of three tissue specific parameters sensitive to pathophysiologic inflammatory processes. This technique constitutes the base of a new full-brain coverage myelin imaging method developed in the course of the current work and presented in the remaining parts of Chapter 2 (Sec. 2.2 to Sec. 2.4).

The following Chapter 3 introduces the topic of cluster analysis. Here, the setting up of different clustering tasks is discussed (Sec. 3.1), followed by a detailed overview of the clustering algorithm family (Sec. 3.2).

Apart from the two introductory parts (Sec. 2.1 and Chapter 3), the remaining part of the document (Chapter 4 to Chapter 7, along with Sec. 2.2 to Sec. 2.4) represent the novelty and the essential contribution of the current work to the scientific topics introduced.

As indicated in Fig. 1.1, three separate research topics can be distinguished here: First, the existing quantitative MRI method presented in Chapter 2 is extended by one parameter, particularly sensitive to neurodegenerative diseases such as MS: the myelin bound water content. With this extension, one brain scan of a subject now results in a multimodal brain map, containing four parameters sensitive to inflammatory and neurodegenerative processes. Results from this study are published in a journal article [Tonkova 2012a].

Chapter 4 to Chapter 6 present the development and evaluation of a new data analysis method, Nuclear Potential Clustering, inspired by the particularly problematic nature of such multiparametric measurements as the quantitative MRI mapping (in its original and its expanded version). Different stages of the development of this tool are presented in [Tonkova 2012b] and [Tonkova 2013]. The final version of

the NPC, as described in this thesis includes a series of refinements. A publication of this results is currently under revision.

It should be explicitly stated that both of these scientific topics are independently invented. On the one side, though motivated by the specific problem of the correlated MRI parameters, the new clustering algorithm is not designed to be applied solely on that particular type of data. It rather aims generally on the correct grouping of high dimensional, possibly correlated, noisy samples. On the other side, the expanded multimodal imaging method is not developed to produce data only to be processed with Nuclear Potential Clustering. Both methods, though part of the same work process, are completely independent of each other and can be employed separately.

The third research topic is presented in Chapter 7 and consists of the application of the newly developed Nuclear Potential Clustering method for the analysis of quantitative multimodal MRI data. The goal here is to search for and determine biomarkers sensitive to multiple sclerosis. The results obtained here are discussed in Sec. 7.5, while the results considering the other two research topics are discussed separately: quantitative myelin assessment in Sec. 2.4 and the developed Nuclear Potential Clustering in Chapter 8.

Chapter 2

Multiple Sclerosis and Assessment of Tissue Myelination

Contents

Main topic of the current chapter is the development of a neuroimaging technique for the quantitative assessment of myelin - a substance which is essential for the proper functioning of the nervous system.

It extends a clinically validated approach for the simultaneous full-brain mapping of three quantitative parameters sensitive to inflammatory processes: the tissue characteristic relaxation time constants T_1 and T_2^*, as well as the total water content H_2O^{total}. In the context of a chronic inflammatory demyelinating disease such as multiple sclerosis, the multivariate explorative analysis based on this multimodal/multiparametric imaging data has a high potential of providing an improved correspondence to the observed patients' clinical disability. It is this very challenging task that has inspired the nuclear potential clustering method presented and discussed in the following chapters. Thus, prior to considering the novel clustering algorithm, the current chapter presents the development of a full-brain coverage myelin mapping approach that extends the quantitative measurement of T_1, T_2^* and H_2O^{total}, thus allowing to simultaneously assess four different disease-sensitive parameters. Section 2.1 begins with a motivation. Next, the developed myelin imaging

technique is demonstrated (Sec. 2.2) followed by the obtained results (Sec. 2.3) and a discussion (Sec. 2.4).

2.1 Motivation

2.1.1 Multiple Sclerosis

Multiple sclerosis is the most common immune-mediated neurological disease in young adults. It is a progressive disorder associated with a damage of the layer that surrounds and insulates the nerve fibers (axons) and assures the rapid electrical impulse transmission between nerve cells (neurons), the myelin sheath. Myelin is produced by oligodendrocytes (a type of glial cell) whose outgrowths are wrapped tightly around the axons in a concentric way (Fig. 2.3). Due to its damage, the communication between neurons is hampered which is the reason why MS can practically cause any neurological symptom and often leads to sensory, motor and cognitive impairment. The damaged myelin forms scar tissue termed white matter plaques or lesions.

Patients can exhibit different set of symptoms at different times depending on the location of the plaques and there is no particular symptom, laboratory examination or physical finding that can by itself definitely confirm if a subject suffers from MS or not. Based on the pattern of symptoms over time, four different main courses of disease development can be recognized as shown in Fig. 2.1. The first type referred to as relapsing-remitting multiple sclerosis (RRMS), is the most common one. It is characterized by attacks causing suddenly an increased level of disability which are separated by periods of no disability increase. An attack itself can last from a few days to months and is also termed relapse, exacerbation, bout or flare. After an attack (involving e.g. sensory or vision loss) the patient recovers partly. However, a residual level of impairment is often observed so that with each relapse, the central nervous system (CNS) becomes increasingly and irreversibly damaged. The second MS type is called secondary progressive multiple sclerosis (SPMS). It is characterized by partial recoveries after the MS bouts in the early stage of the disease, but proceeds with a steady progression of disability. The third possible disease course is the primary progressive multiple sclerosis (PPMS) where a steady progression of the patient's disability over lifetime is observed. The last main type is the progressive relapsing multiple sclerosis (PRMS). Similar to the primary progressive course, in this case a steady disability progression is observed as well, however with superimposed MS attacks.

To assess MS severity and monitor changes in the level of disability over time, an ordinal clinical rating scale termed expanded disability status scale (EDSS) is employed [Kurtzke 1983]. It ranges from 0 (which corresponds to a normal neurological finding) to 10 (which corresponds to death due to MS) in half-point increments. The assessment of the EDSS is based on a standard neurological examination in which each of the following functional systems is scored between 0 (no disability) and 5 or 6 (severe disability): pyramidal (weakness or difficulty moving limbs), cerebellar

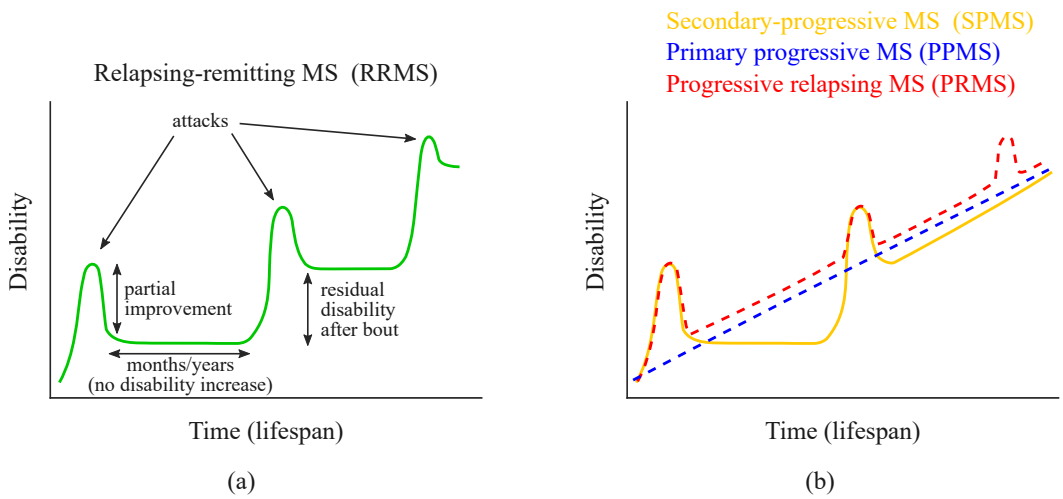


Figure 2.1: Main types of multiple sclerosis disease course. (a) Relapsing-remitting multiple sclerosis (RRMS), characterized by attacks of increased impairment, followed by a partial recovery and separated by periods of no disability increase. (b) Typical course of development of secondary progressive multiple sclerosis (SPMS), primary progressive multiple sclerosis (PPMS) and progressive relapsing multiple sclerosis (PRMS) detailedly described in the text.

(ataxia, loss of coordination or tremor), brainstem (problems with speech, swallowing and nystagmus), sensory (numbness or loss of sensations), bowel and bladder function, visual function, cerebral (or mental) functions, and other. This scoring of the functional systems, along with information concerning the patient's gait and use of assistive devices is employed for the final EDSS estimation. Table 2.1 gives a detailed description of the EDSS scores [Kurtzke 1983, Haber 1985].

Currently, there is no cure for MS. However, different treatments are available which can modify the course of the disease and manage the MS symptoms. Their goal is to lessen the severity of the relapses and their lasting effects as well as to support and improve patient's recovery after an attack and delay or possibly prevent new relapses [Compston 2002].

2.1.2 Conventional Imaging Diagnosis

The Revised McDonald Criteria [Polman 2011] are the state-of-the art guidelines for multiple sclerosis diagnosis based on magnetic resonance imaging (MRI), visual evoked potentials and cerebrospinal fluid analysis. As a determining evidence for central nervous system damage, the demonstration of a temporal and spacial lesions dissemination is employed (i.e plaques are detected at different time points and affect at least two different CNS regions).

An early diagnosis and corresponding onset of treatment is of crucial importance as MS tends to lead to more damage in the first year than in later stages of the disease [Kuhlmann 2002]. Thus, the diagnostic guidelines aim at an earliest possible initial

Table 2.1: Expanded disability status scale (EDSS) [Kurtzke 1983]

Score	Description
0.0	Normal neurological exam (all grade 0 in all Functional System (FS) scores*).
1.0	No disability, minimal signs in one FS* (i.e., grade 1).
1.5	No disability, minimal signs in more than one FS* (more than 1 FS grade 1).
2.0	Minimal disability in one FS (one FS grade 2, others 0 or 1).
2.5	Minimal disability in two FS (two FS grade 2, others 0 or 1).
3.0	Moderate disability in one FS (one FS grade 3, others 0 or 1) or mild disability in three or four FS (three or four FS grade 2, others 0 or 1) though fully ambulatory.
3.5	Fully ambulatory but with moderate disability in one FS (one grade 3) and one or two FS grade 2; or two FS grade 3 (others 0 or 1) or five grade 2 (others 0 or 1).
4.0	Fully ambulatory without aid, self-sufficient, up and about some 12 hours a day despite relatively severe disability consisting of one FS grade 4 (others 0 or 1), or combination of lesser grades exceeding limits of previous steps; able to walk without aid or rest some 500 meters.
4.5	Fully ambulatory without aid, up and about much of the day, able to work a full day, may otherwise have some limitation of full activity or require minimal assistance; characterized by relatively severe disability usually consisting of one FS grade 4 (others or 1) or combinations of lesser grades exceeding limits of previous steps; able to walk without aid or rest some 300 meters.
5.0	Ambulatory without aid or rest for about 200 meters; disability severe enough to impair full daily activities (e.g., to work a full day without special provisions); (Usual FS equivalents are one grade 5 alone, others 0 or 1; or combinations of lesser grades usually exceeding specifications for step 4.0).
5.5	Ambulatory without aid for about 100 meters; disability severe enough to preclude full daily activities; (Usual FS equivalents are one grade 5 alone, others 0 or 1; or combination of lesser grades usually exceeding those for step 4.0).
6.0	Intermittent or unilateral constant assistance (cane, crutch, brace) required to walk about 100 meters with or without resting; (Usual FS equivalents are combinations with more than two FS grade 3+).
6.5	Constant bilateral assistance (canes, crutches, braces) required to walk about 20 meters without resting; (Usual FS equivalents are combinations with more than two FS grade 3+).
7.0	Unable to walk beyond approximately 5 meters even with aid, essentially restricted to wheelchair; wheels self in standard wheelchair and transfers alone; up and about in wheelchair some 12 hours a day; (Usual FS equivalents are combinations with more than one FS grade 4+; very rarely pyramidal grade 5 alone).
7.5	Unable to take more than a few steps; restricted to wheelchair; may need aid in transfer; wheels self but cannot carry on in standard wheelchair a full day; May require motorized wheelchair; (Usual FS equivalents are combinations with more than one FS grade 4+).
8.0	Essentially restricted to bed or chair or perambulated in wheelchair, but may be out of bed itself much of the day; retains many self-care functions; generally has effective use of arms; (Usual FS equivalents are combinations, generally grade 4+ in several systems).
8.5	Essentially restricted to bed much of day; has some effective use of arm(s); retains some self-care functions; (Usual FS equivalents are combinations, generally 4+ in several systems).
9.0	Helpless bed patient; can communicate and eat; (Usual FS equivalents are combinations, mostly grade 4+).
9.5	Totally helpless bed patient; unable to communicate effectively or eat/swallow; (Usual FS equivalents are combinations, almost all grade 4+).
10.0	Death due to MS.

*Excludes cerebral function grade 1.

diagnosis, primarily enabled due to MRI which is very effective in localizing areas of damage. The conventional techniques here include T_1-weighted images, with and without contrast agent, as well as T_2-weighted images [Barkhof 1997, Tintoré 2000, Polman 2005].

However, a mismatch between the lesion load and the clinical findings (actual status/development of the disability rate) is observed, a phenomenon termed the clinico-radiological paradox [Neema 2007, Bakshi 2008]. Figure 2.2 shows two FLAIR scans (fluid attenuated inversion recovery, a standard neuroimaging sequence employed in the MS diagnosis) of patients with different EDSS. One of the patients has an EDSS of zero corresponding to a normal neurological exam. The second patient has EDSS=5 which indicated a disability severe enough to impair full daily activities. However, when comparing the FLAIR images of both subjects, it is evident that the first patient (EDSS=0) has the higher lesion load (Fig. 2.2(a)). This example demonstrates that a higher lesion burden does not necessarily correspond to an increased disability level. Furthermore, the standard MRI techniques employed in MS diagnosis produce qualitative images, meaning that the contrast ratio of different images is the same, but the gray value scale is not standardized and images cannot be directly compared. These facts along with the need to understand deeper the pathophysiologic processes of MS which could possibly enable better individual diagnosis, monitoring and therapy evaluation, have led to the development of MRI techniques assessing quantitative tissue parameters. Such quantitative measurements allow for intra- and inter-subject comparisons (particularly advantageous when an assessment of time and space alternations is needed), estimation of reference tissue values in healthy controls and, most importantly, studying (possibly subtle) changes in the normal appearing neural tissue that are otherwise not accessible in the standard imaging diagnostic process.

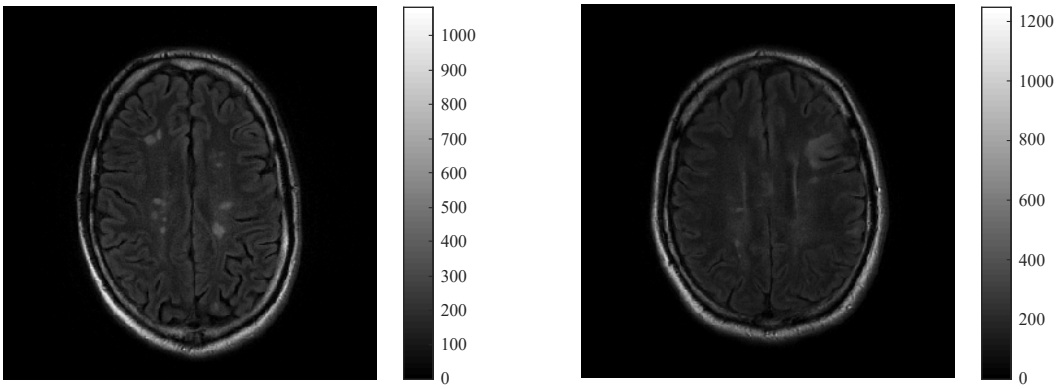


Figure 2.2: FLAIR images of two MS patients demonstrating the mismatch between clinical findings and observed lesion load. (a) Patient with EDSS=5. (b) Patient with EDSS=0.

2.1.3 Quantitative Assessment of Tissue Specific Parameters and Measuring of Myelin Integrity

In the context of a chronic demyelinating disease such as MS, the quantitative MRI of neural tissue myelination has a high potential to support disease diagnosis and monitoring as well as possibly enable prognosis predictions and a deeper understanding of brain plasticity. Therefore, intensive research efforts aiming at the development of myelin mapping techniques have been expended in the recent past [Laule 2004, Du 2007, Oh 2007, Hwang 2010, Lenz 2012, Hwang 2011, Nguyen 2012]. As a surrogate *in vivo* estimate for myelin integrity, they all employ a quantification of the myelin bound water fraction (MWF). In this water pool, the motility of the H_2O molecules trapped between the lipid myelin bilayers is restricted (Fig. 2.3) which results in a characteristic shorter transverse relaxation (qualified either by a T_2 or a T_2^* constant depending on the applied MRI sequence) as compared to other MR detectable water pools [Mathur-De Vré 1984].

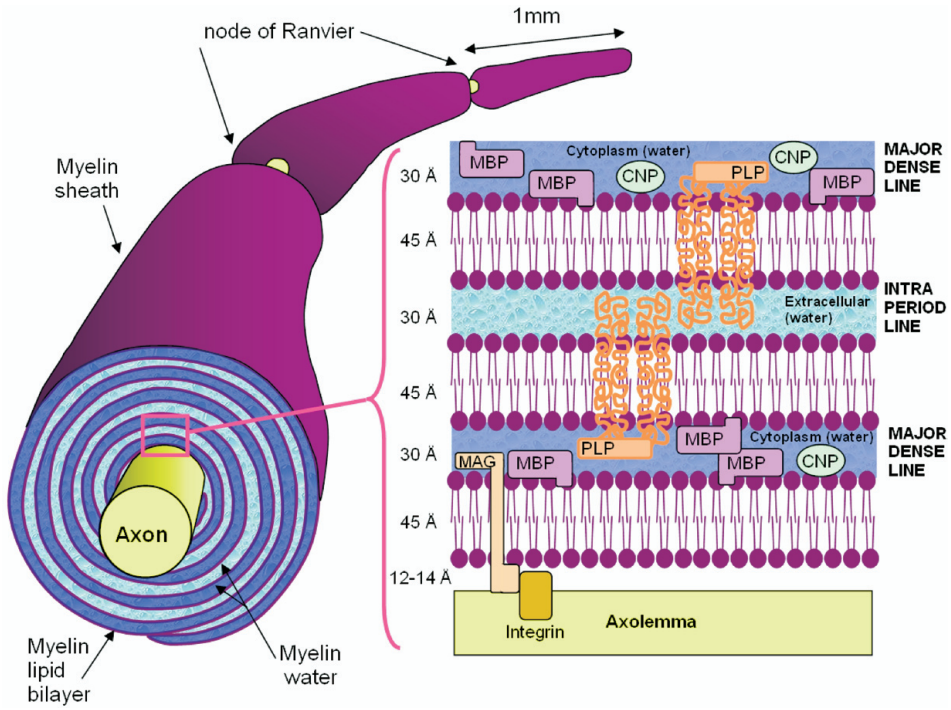


Figure 2.3: Myelin structure (graphic source: [Laule 2004]).

The initial research on this topic has shown, that *in vivo* imaging of CNS myelin based on this phenomenon is not only realizable but the myelin water fraction also correlates with the myelin density estimated by quantitative histopathologic measures [Laule 2004, Laule 2006, Laule 2008, Meyers 2009, Kolind 2009, Minty 2009, Laule 2011]. In the imaging technique employed in these studies, a multi-exponential analysis of the T_2 relaxation data is conducted and the ratio between the fast decaying water pool amplitude and the amplitude of all detected water pools is employed

as an estimate for myelin water content. However, the employed MRI protocol is restricted to a single transverse plane acquisition in approx. 25 minutes, which renders inter-subject registration and comparison needed in clinical studies problematic. Techniques aiming at the acquisition of higher number of slices in less time, are introduced by [Oh 2007, Du 2007, Deoni 2008, Hwang 2010]. The approach presented by [Oh 2007] allows for the acquisition of 16 slices with a voxel volume of $2 \times 2 \times 5\ mm^3$ within 10 minutes. Here, 12 points on the T_2 relaxation curve are sampled using a fast T_2-prep spiral sequence with a maximum echo time of 294 ms. [Du 2007] are the firs to demonstrate the feasibility to map MWF based on the T_2^* relaxation curve sampling (126 points) using a two-dimensional multi-slice multi-echo gradient echo (MEGE) measurement. They employ a 3-pool model to study the myelin distribution in postmortem brains at 3 T. The method enables the acquisition of only 5 slices with a voxel size of $0.78 \times 0.78 \times 5\ mm^3$ in 8.7 minutes. This technique is modified by [Hwang 2010] who increase the number of slices to 8 in an acquisition time of 8.5 minutes and adapt it for *in vivo* measurements of healthy subjects.

However, none of these approaches allows for a full-brain myelin mapping which would provide a more clinically applicable examination. The first such technique is introduced by [Deoni 2008] (and applied in a study on MS by [Kolind 2012]) who employ a rapid method for the voxelwise determination of the longitudinal and transverse relaxation times, T_1 and T_2 respectively: mcDESPOT (multicomponent driven equilibrium single pulse observation of $T1\backslash T2$). Even though 60 slices can be reconstructed in 16 to 30 minutes, this measurement exhibits a systematic shift of the estimated MWF in white and gray matter towards higher values as compared to results reported by other studies [Laule 2004, Kolind 2009, Vavasour 2009, Oh 2007, Du 2007, Hwang 2010].

Another full-brain approach based on a multicomponent T_2^* decay analysis is presented by [Lenz 2012] who demonstrate the feasibility of *in vivo* myelin water imaging using three-dimensional multigradient-echo pulse sequences (88 slices, voxel size $2.2 \times 2 \times 2\ mm^3$, acquisition time 9.7 minutes). However, though able to detect demyelination in multiple sclerosis lesions, the myelin water fraction values obtained with this technique remain systematically in the lower range of myelin water reported in the literature (mean value in white matter amounts to 6.9% compared to e.g. [Du 2007] who report 11% or [Hwang 2010] who obtain 10.2%).

Latest research avenues investigate the potential of modeling myelin more precisely based on an observed T_2^* anisotropy. This effect is suggested to originate from a frequency shift dependency of the axonal orientation relative to the main magnetic field, $\mathbf{B_0}$ [Bender 2010, Lee 2017]. Though no details about the underlying effects are known yet, different models have emerged that aim to incorporate these findings into the MWF estimation. [Van Gelderen 2012] consider a model including a frequency offset term in the myelin water component aiming to reduce systematic errors in early echoes. They employ a gradient-echo acquisition at 3 and 7 T to reconstruct a single slice with 2 mm thickness, an in plane resolution of 256×96 voxels and a field of view of $(240 \times 180)\ mm$. Furthermore, [Sati 2013] and [Nam 2015]

investigate a complex three-component model including frequency shift in each of the pools (myelin, axonal and extracellular water), which is fitted either to magnitude or to complex data. [Nam 2015] publish a comparison of these approaches when employed for the MWF estimation with different numbers of echoes (from 12 to 32 echoes). They conclude that the complex model is advantageous when the frequency shift between components becomes large (that is fibers are perpendicular to B_0), but does not outperform the magnitude model when there is small or absent frequency shift between the signals (i.e. fibers are parallel to B_0). In the publication of [Lee 2017], an extensive review on this topic is presented.

Non of all introduced methods accounts for the simultaneous assessment of multiple quantitative parameters with full brain coverage. However, as simultaneous changes in different quantitative parameters such as diffusion coefficients and characteristic relaxation times have been reported in MS patients [Neema 2007, Bakshi 2008], it is expected that a multimodal imaging approach could potentially make the MRI of MS more pathologically specific.

In the current work, an acquisition protocol optimized for the rapid high-resolution ($1 \times 1 \times 2\ mm^3$ voxel size) assessment of the tissue-specific parameters T_1, T_2^* and total (absolute) water content H_2O^{total} introduced by [Neeb 2008] is employed as a basis to develop a full-brain measurement of the myelin bound water fraction, H_2O^{myelin}. Thus, with the proposed extension, a single scan allows for the reconstruction of four congruent brain maps depicting the spatial distribution of four different disease-sensitive quantitative parameters. This multiparametric data is the basis for the analysis presented in Chapter 7.

The quantitative MRI protocol employed for the reconstruction of T_1, T_2^* and H_2O^{total} includes two types of standard sequences available on any modern MRI scanner: multi-echo gradient echo (MEGE) and echo planar imaging (EPI). In the current work, all measurements are conducted at a 3T TRIO System (Siemens Medical Solutions, Erlangen, Germany) and the original $QUTE-ST_2^*$ sequence ("Quantitative T_2^* Image") as introduced in [Neeb 2008] is replaced by a multi gradient echo sequence termed "gre" provided by the manufacturer. Briefly, in this quantitative MRI protocol two MEGE sequences, MEGE1 and MEGE2, with different acquisition parameters are used. MEGE1 measures the decay signal at ten equidistant time points between $TE = 4.8\ ms$ and $TE = 42\ ms$. A single exponential function is fitted to the signal intensities at the ten times points for each voxel to estimate the transverse relaxation time, T_2^*, and the extrapolated initial signal intensity, $S_{0,SE}$ which is a parameter proportional to the total water content. For the quantification of H_2O^{total}, a series of different influencing factors are considered. For this purpose, the rest of the acquired sequences are used to correct for:

- differences in the signal saturation affected by the longitudinal relaxation time, T_1 (here, the second multi echo gradient echo sequence, MEGE2, which has flip angle and repetition time different from MEGE1 is used and the signal intensities from both measurements are employed to define a function of T_1);
- inhomogeneities of the transmitting $\mathbf{B_1}$ field, resulting in a spatial variation of

the effective excitation flip angle (an effective flip map is calculated here based on the signal intensity ratio of two EPI sequences with different nominal flip angles, the rest of the acquisition parameters are identical);

- imperfections of the receive coil system (corrected by comparing the signal intensities of two identical EPI sequences, where in the first case the signal is acquired with the head coil while in the second case it is acquired with the body coil).

Having considered all these factors, the total water content is estimated as the ratio between the corrected signal intensity (for each voxel) and the signal intensity within voxels depicting solely water. To avoid the time-consuming placement of additional external water reference probe (which would need an additional correction for temperature differences), the average signal intensity within the cerebrospinal fluid (CSF) segment in the images is employed as an internal reference for 100% water content. Further details concerning the exact acquisition procedure as well as the reconstruction of the T_1, T_2^* and H_2O^{total} maps can be found in [Neeb 2008, Tonkova 2010].

The method developed is published in [Tonkova 2012a] and has been successfully employed in a multicenter study for the development of a whole brain atlas and its application to low-grade multiple sclerosis [Neeb 2012]. In this study its independence of a specific measuring setup has been validated. In addition, [Neeb 2014] employ this protocol to design a supervised learning model predicting the presence or absence of MS based solely on imaging information characterizing normal appearing brain tissue. This study demonstrates that neural tissue appearing not to be affected in standard MRI *is* experiencing pathological changes and these changes can be assessed by means of the proposed quantitative MRI protocol.

2.2 Myelin Water Content Mapping Method

The developed myelin water content mapping is based on the signal acquired with the MEGE1 sequence. Multi-exponential T_2^* signal decay of myelinated tissue is employed and H_2O^{myelin} is measured based on the quantification of two water pools characterized by different transverse relaxation times: "myelin water" and "rest". As the MRI protocol used as basis is focused on the fast mapping of quantitative MR parameters with whole brain coverage in clinically relevant measurement times, the sampling density of the T_2^* curve is compromised to the ten echo times measured with MEGE1. Therefore, the pool amplitudes are determined using a quadratic optimization approach. The optimization is constrained by including *a priori* knowledge about the brain water pools. All constraints are optimized in a simulation study to minimize systematic error sources due to the incomplete knowledge about the real pool-specific relaxation properties. Based on the simulation results, whole brain *in vivo* myelin water content maps are acquired in ten healthy controls and one MS patient.

2.2.1 Simulations

For the evaluation of accuracy and precision of the proposed MEGE based myelin water content quantification, artificial T_2^* data is generated. In this, a two-pool model of white matter, extensively studied and validated by [Laule 2006], is employed. Assuming that the pools have relaxation times $T_2^{*,f}$ (characterizing the fast relaxing component) and $T_2^{*,s}$ (characterizing the slow relaxing component) and corresponding pool fractions of 10% and 90%, respectively, the artificial decay signal is generated as follows:

$$y_i^{meas} = \left(0.1\ e^{\frac{-t_i}{T_2^{*,f}}} + 0.9\ e^{\frac{-t_i}{T_2^{*,s}}}\right) + \mathcal{N}\left(0, \frac{1}{SNR}\right), i = 1, \ldots, 10 \tag{2.1}$$

where $\mathcal{N}\left(0, \frac{1}{SNR}\right)$ represents normally distributed random noise with zero mean and standard deviation given by the reciprocal of the signal-to-noise-ratio (SNR), $\frac{1}{SNR}$. To conform the *in vivo* measurements, ten points with the same TE and echo spacing as employed in MEGE1 are simulated.

2.2.2 Quadratic Programming

In the most general case, an ideal MRI measurement can be expressed as data points y_i^{theo} which are linear combinations of n exponential functions, depending on the echo time, t_i, and the compartment specific relaxation time, $T_2^*(j)$.

$$y_i^{theo} = \sum_{j=1}^{n} s_j\ e^{\frac{-t_i}{T_2^*(j)}}, i = 1, \ldots, 10 \tag{2.2}$$

Here, $s_j \equiv s\big(T_2^*(j)\big) > 0$ represents the relative fraction occupied by the j-th compartment with relaxation time $T_2^*(j)$. For a given measured decay curve, y_i^{meas}, Eq. 2.2 has to be inverted to determine the corresponding pool amplitudes, s_j. However, this is not possible given the underdetermined linear system of equations. Therefore, a quadratic programming (QP) approach is employed which determines the most likely amplitude vector $\mathbf{s}^*$ by minimizing the squared difference between the real data and the decay model,

$$\min\left(\sum_{i=1}^{10}\left|y_i^{meas} - \sum_{j=1}^{n} s_j^* e^{\frac{-t_i}{T_2^*(j)}}\right|^2 + \lambda A\right) \tag{2.3}$$

The additional term

$$A = \sum_{j=1}^{n-1}\left|s_{j+1}^* - s_j^*\right|^2 \tag{2.4}$$

is introduced to obtain a smooth distribution of s_j^* as described e.g. in [Laule 2006]. The regularization parameter λ is set to 1 which results in the smallest overall systematic error in the simulation experiments. Omitting the term A in the objective

function, i.e. setting $\lambda = 0$, results in a discrete distribution of the pool amplitudes, s_j^*. In this case, a tendency of s_j^* to cluster at the lower or upper end of both intervals defined by Eq. 2.5 is observed, resulting in an increased systematic error. Furthermore, a significant dependence of the optimization results on SNR is observed in this case. Both effects are largely suppressed by extending the objective function by the proper smoothness constraints for s_j^*.

Quadratic programming offers the possibility to linearly constrain the solution space for the vector $\mathbf{s}^*$, similar to standard linear optimization approaches [Murty 1988]. As the current work is based on a two-pool model of white matter, the amplitudes of the two distinct T_2^* intervals are constrained by

$$s\big(T_2^*(j)\big) \equiv s_j^* < a_f \ \mathrm{rect}\left(\frac{T_2^*(j) - \widehat{T_2^{*,f}}}{\widehat{w_f}}\right) + a_s \ \mathrm{rect}\left(\frac{T_2^*(j) - \widehat{T_2^{*,s}}}{\widehat{w_s}}\right) \tag{2.5}$$

whereas amplitudes outside those two intervals are forced to zero.

Here, $\widehat{T_2^{*,f}}$ and $\widehat{T_2^{*,s}}$ determine the positions, while $\widehat{w_f}$ and $\widehat{w_s}$ the total widths of the intervals where a nonzero amplitude is allowed for the fast and the slow relaxing pool, respectively. Furthermore, $\mathrm{rect}(x) = 1$ for $|x| < 0.5$ and zero elsewhere. The parameters a_f and a_s control the desired maximum amplitudes. Finally, $s_j^* \geq 0$ is required for the whole T_2^* range.

Equation 2.5 contains six parameters, $\widehat{T_2^{*,f}}$, $\widehat{T_2^{*,s}}$, $\widehat{w_f}$, $\widehat{w_s}$, a_f and a_s, which are in general not *a priori* known. However, information extracted from the measured T_2^* decay curve can be incorporated into the definition of the free parameters. Most importantly, it is noted that a single exponential fit of the decay curve will in general yield a result close to $\widehat{T_2^{*,s}}$ as the myelin pool accounts for only approx. 10% of the whole decay signal. In order to estimate $\widehat{T_2^{*,s}}$ from the data, the decay curve of each point is therefore first fitted using a single exponential model. In this way the parameters $T_{2,SE}^*$ and $S_{0,SE}$ are obtained, which denote the corresponding transverse relaxation time and the extrapolated initial signal intensity, respectively (Fig. 2.4, left top). To estimate $\widehat{T_2^{*,s}}$ for a given $T_{2,SE}^*$, biexponential decay curves are modeled according to Eq. 2.1 for different values of $T_2^{*,s}$ in the range between 25 ms and 230 ms with $T_2^{*,f}$ fixed to 10 ms. This provides the required mapping between $T_{2,SE}^*$ and $\widehat{T_2^{*,s}}$ that can be formally expressed by the function $\widehat{T_2^{*,s}} = f(T_{2,SE}^*)$, defining the center position of the slow relaxation pool as shown in Fig. 2.4 (right panel). Furthermore, the maximum amplitudes of the fast and slow relaxing compartments are set to $a_f = 0.5\ S_{0,SE}$ and $a_s = 2\ S_{0,SE}$, respectively. These are weak conditions which do not significantly constrain the estimation of myelin water content which is supposed to be $\leq 20\%$ [Laule 2004, Kolind 2009, Vavasour 2009, Oh 2007, Du 2007, Deoni 2008, Hwang 2010].

The remaining three parameters in Eq. 2.5, $\widehat{T_2^{*,f}}$, $\widehat{w_f}$ and $\widehat{w_s}$, are determined based on simulation results as described in the next Sec. 2.2.3. Once the proper

values for all constraints ($\widehat{T_2^{*,f}}$, $\widehat{T_2^{*,s}}$, $\widehat{w_f}$, $\widehat{w_s}$, a_f, a_s) are estimated, they can be employed for the calculation of the relative myelin water content, W_{My}^{rel}, in the following way. For a given measured decay curve, first the most likely amplitude spectrum, $s(T_2^*(j))$, is determined as described above (Fig. 2.4, left bottom). Based on $s(T_2^*(j))$, the relative myelin water content, W_{My}^{rel}, is then defined by the ratio between myelin amplitude, $S_{My} = \sum_{T_2^*(j)\in I^f} S(T_2^*(j))$, and the total amplitude, $S_{Tot} = \sum_{T_2^*(j)\in(I^f \cup I^s)} S(T_2^*(j))$. Here, I^f and I^s represent the intervals defined by the first and the second term on the right hand side of Eq. 2.5, respectively.

2.2.3 Protocol Optimization and Systematic Errors

To estimate the remaining free parameters in Eq. 2.5, $\widehat{T_2^{*,f}}$, $\widehat{w_f}$ and $\widehat{w_s}$, simulations are performed aiming to minimize the average systematic error. As the systematic shift might depend on relaxation times of the fast and slow pool as well as on SNR, synthetic data are generated according to Eq. 2.1 with $T_2^{*,f} \in [5, 18]$ ms, $T_2^{*,s} \in [35, 130]$ ms and $SNR \in [10, 250]$, respectively. In addition, the three unknown parameters are changed as follows:

$$\begin{aligned} \widehat{T_2^{*,f}} &= 5\ ms, 6\ ms \ldots, 20\ ms \\ \widehat{w_f} &= 5\ ms, 7.5\ ms, 10\ ms \\ \widehat{w_f} &= 2\ ms, 5\ ms, 10\ ms, 15\ ms. \end{aligned} \tag{2.6}$$

For each of the possible combinations, the average relative myelin water content is determined from 10000 independent samplings of the relaxation curve according to Eq. 2.1. The result can formally be expressed as $W_{My}^{rel}\left(T_2^{*,s}, T_2^{*,f}, \widehat{T_2^{*,f}}\widehat{w_f}, \widehat{w_s}\right)$.

As $T_2^{*,s}$ and $T_2^{*,f}$ generally depend on the spatial position in a real data set, the goal here is to find the parameter vector $\left(\widehat{T_2^{*,f}}, \widehat{w_f}, \widehat{w_s}\right)$ which minimizes the deviation of W_{My}^{rel} averaged over all possible $T_2^{*,s}$ and $T_2^{*,f}$. An exhaustive search is performed to define the corresponding optimized parameter set. Furthermore, the result might be biased by the presence of further relaxation pools which have been observed in white matter [Laule 2007]. Based on the results given in [Laule 2007], a third component with amplitude 0.05 and $T_2^{*,3rd-pool} = 200$ ms is added to the model defined by Eq. 2.2. The corresponding $T_2^{*,3rd-pool}$ here is chosen to be 0.5 T_2, as a pool with relative amplitude of 5% and T_2 of ≈ 400 ms has been reported in [Laule 2007]. However, the transverse relaxation time T_2 is reduced by 50% in the current study in order to account for the faster relaxation process employing gradient echo data at 3 T as compared to spin echo acquisitions at 1.5 T.

2.2.4 Magnetic field inhomogeneity correction

To compensate modulations of the signal decay in regions with strong magnetic field inhomogeneities, an approach similar to the one described in [Fernández-Seara 2000,

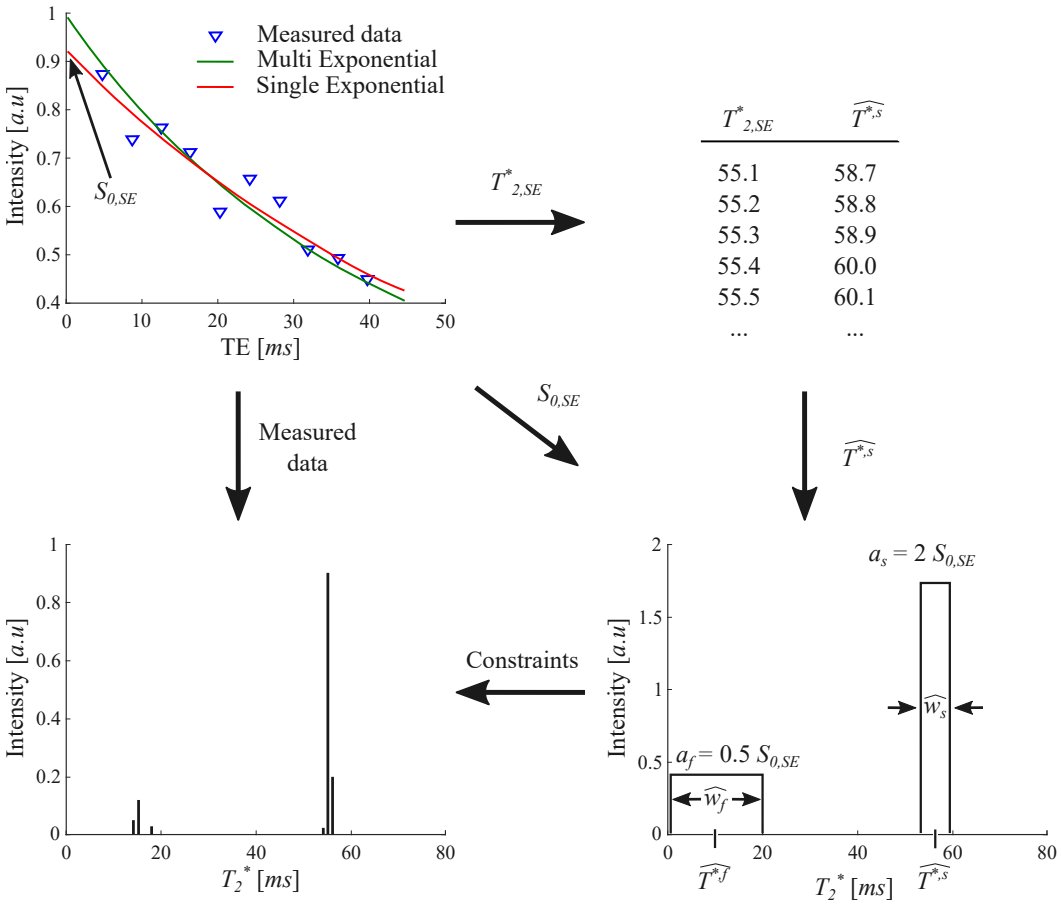


Figure 2.4: Myelin Constraints Worklow.

Deoni 2008] is applied in the current work. Concretely, magnetic field offset maps, $\Delta\mathbf{B}(x, y, z)$, are constructed from a linear fit of the phase data acquired with the MEGE sequence. The phase data are corrected for 2π-fold effects prior fitting. A gradient map, $\Delta\mathbf{G}(x, y, z)$, along the slice-select direction is determined based on $\Delta\mathbf{B}(x, y, z)$ as the voxel size in z-direction, Δz, is a factor two larger than those in x or y. The measured signal intensity at time point, TE_i, is then corrected by multiplication of each measured data point with the factor $c\ \mathrm{sinc}^{-1}(\frac{\gamma G_z TE_i \Delta z}{2})$, where c is a normalization constant and $\mathrm{sinc}\ (x) = \frac{\sin\ (x)}{x}$. As the determination of relaxation pool fractions can be altered by magnetic field imperfections, the average and the 99% quantile of $\Delta\mathbf{B}(x, y, z)$ and $\Delta\mathbf{G}(x, y, z)$ are determined for each subject and are correlated with the average myelin water content.

2.2.5 *In vivo* Measurements

Whole brain myelin water content maps are acquired in a group of ten healthy controls and one MS patient.

The acquisition parameters of the MEGE1 sequence employed for the myelin

water content measurement are as follows: $TE = 4.8\ ms$; echo spacing = 3.74 ms; $BW = 310\ Hz/Pixel$; repetition time $TR = 2100\ ms$; flip angle $\alpha = 40°$; ten points sampled on the T_2^* decay curve; matrix size = 256×192; 50 slices with 2 mm thickness and a gap of 1 mm; in plane resolution = $1 \times 1\ mm^2$. All measurements are performed on a 3T TRIO system (Siemens Medical Solutions, Erlangen, Germany) equipped with a 16 channel head coil for signal reception. The standard automated second order shimming procedure is performed for each patient individually prior to the start of data acquisition. Informed written consent is obtained from each subject.

2.2.6 Image Processing

Quantitative myelin water content maps are reconstructed as described above. To better display relevant structures, the corresponding myelin water maps are smoothed with a Gaussian filter with kernel width of 1.5 pixels. Furthermore, quantitative T_1 maps are reconstructed as described in [Neeb 2008]. A mask for white matter is created using a T_1-based histogram segmentation [Neeb 2006a] with adapted thresholds, $T_1 \in [500\ ms, 899\ ms]$, due to the increased T_1 at 3 T. The average myelin water content in the white matter segment is determined for each subject. No analysis is performed for gray matter, as the T_1 distributions of gray matter and MS lesions are comparable. The corresponding algorithms are implemented as described in [Tonkova 2010].

2.3 Results

Figure 2.5 shows the relative myelin water content as function of SNR for different combinations of fast and slow relaxation time. The corresponding *optimized* parameter vector, $\left(\widehat{T_2^{*,f}}, \widehat{w_f}, \widehat{w_s}\right)$ is determined to $(10, 5, 10)\ ms$. As can be seen from Fig. 2.5, the myelin water content measurement is almost independent of the signal-to-noise ratio for SNR values ≥ 40. Furthermore, a systematic deviation of W_{My}^{rel} is observed when changing the fast relaxation time, $T_2^{*,f}$. Pools with small $T_2^{*,f}$ are overestimated whereas slower relaxing pools are correspondingly underestimated. The error shows a consistent pattern in the whole range of slow relaxation times as can be seen from Fig. 2.5. Moreover, it is obvious that the systematic error decreases with increasing $T_2^{*,s}$ and therefore increased spacing between both peaks. For a SNR of 70, typical for the conducted *in vivo* experiments, the myelin water content ranges between $0.085 - 0.115$ for a true value of 0.1. Consequently, the maximum systematic error amounts to ± 0.015. The average absolute deviation of all simulated points is smaller (± 0.009). In contrast to the dependence on $T_2^{*,f}$, the presence of a third relaxation component does not introduce a significant bias (absolute deviation ± 0.0091).

Results from the healthy control group are given in Table 2.2. The relative myelin water content in white matter ranges between 10.5% and 11.1% with an

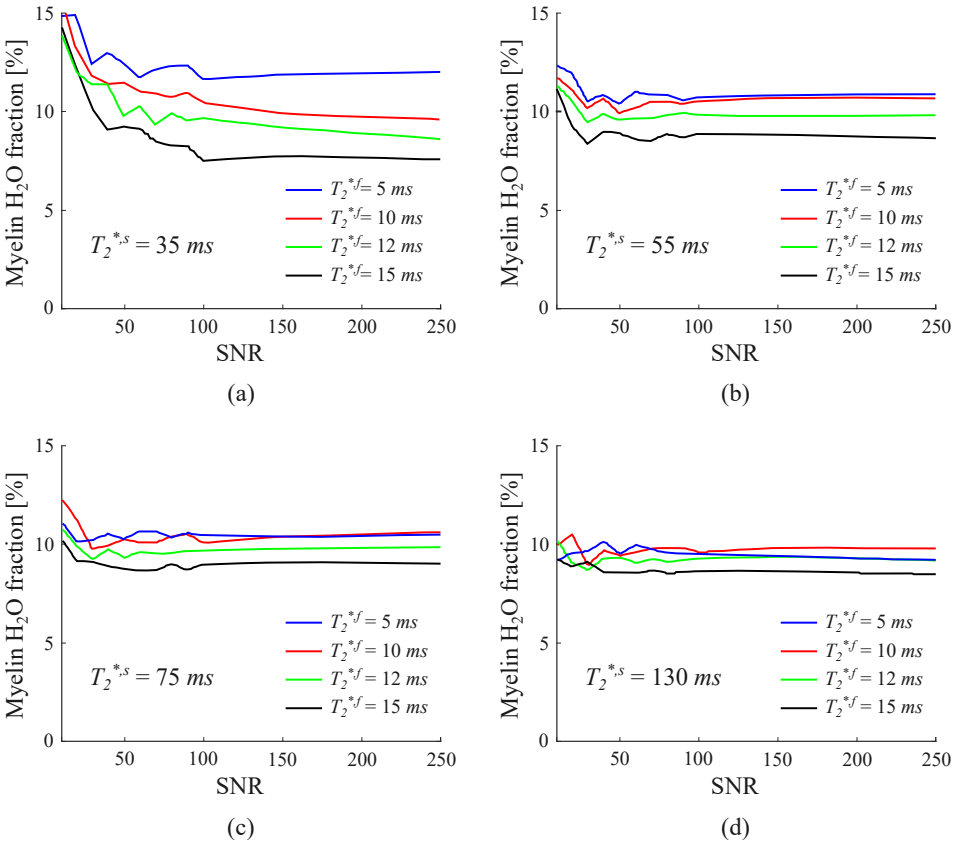


Figure 2.5: Relative myelin water content as a function of SNR for different combinations of fast and slow relaxation times.

average of $(10.7 \pm 0.2)\%$. The corresponding value is slightly reduced in the MS patient (10.3%). Furthermore, no significant correlation is observed between myelin water content and the four variables measuring magnetic field inhomogeneities in white matter ($r_{My-\Delta Bavg} = 0.18$, $p = 0.62$; $r_{My-\Delta B99\%Quant} = 0.21$, $p = 0.56$; $r_{My-\Delta Gavg} = 0.05$, $p = 0.9$; $r_{My-\Delta G99\%Quant} = -0.01$, $p = 0.98$).

Figure 2.6(a) shows a myelin water content map for the MS patient. For a better visualization of the underlying anatomy, Fig. 2.6(b) shows the corresponding quantitative T_2^* map. Myelin water content is reduced in the periventricular lesion (white arrow) to 4.0% in comparison to 11.2% in the normal appearing white matter surrounding the lesion.

Figure 2.7(a) shows the T_2^* signal decay for the lesion visible in Fig. 2.6 (boxes) and for normal appearing white matter (triangles). The average signal intensity is higher for lesion voxels, consistent with an increased total water content. Furthermore, the biexponential decay is more obvious for normal appearing white matter. This is consistent with the quantitative results obtained from the linear optimization shown in Fig. 2.7(b). As can be seen from Fig. 2.7(a), the T_2^* distribution of the lesion is characteristically different from the corresponding distribution of normal

Table 2.2: Summary of the subjects studied. In the first three rows, age, gender and EDSS score of each individual are shown. As the latter is a measure of MS disease grade it is only determined for the MS patient (last column in the table) and is not acquired (n.a.) in the rest of the cases. Furthermore, the average relative myelin water content in white matter (W_{my}) and the corresponding standard deviation, $\sigma(W_{my})$, are shown. Finally, the last four rows contain the average and the 99% quantile of the magnetic field offset (ΔB) and the field gradient (ΔG) distribution, respectively.

Parameter	Healthy control group										MS patient
Age	35	36	47	42	54	54	21	26	26	27	51
Gender	m	f	m	m	f	f	f	m	m	m	f
EDSS	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	4.5
W_{my} (%)	11.1	10.5	10.7	10.8	10.6	10.9	10.8	11.1	10.7	11.1	10.3
$\sigma(W_{my})$ (%)	3.8	3.5	4.0	3.8	3.5	3.2	3.1	3.0	3.1	3.2	3.6
$T_2^{*,f}$ $[ms]$	9.8	10.2	10.0	9.9	10.3	9.7	9.8	10.1	10.0	9.5	9.8
$T_2^{*,s}$ $[ms]$	60.7	59.3	56.8	61.6	57.3	57.4	1.6	57.2	54.7	59.8	57.4
ΔB_{Avg} ($\times 10^{-8}$ T)	2.7	3.1	4.3	3.8	3.3	4.9	3.9	4.9	4.9	4.7	2.5
$\Delta B_{99\%Quant}$ ($\times 10^{-7}$ T)	3.3	3.6	4.3	3.7	3.7	4.9	4.8	4.9	4.9	4.4	3.3
ΔG_{Avg} ($\times 10^{-6}$ T/m)	2.3	2.6	3.4	2.6	2.6	3.1	1.6	3.5	3.5	2.8	2.0
$\Delta G_{99\%Quant}$ ($\times 10^{-5}$ T)	2.6	3.1	3.2	2.4	3.1	2.9	1.7	4.1	4.1	2.3	2.4

appearing white matter. Both fast and slow relaxation times are more homogeneous in the normal appearing white matter than in pathological tissue. Furthermore, the peak spacing is larger for the lesion with a decreased average fast relaxation time and an increased slow relaxation time. The latter can be attributed to the increased free water content which is correlated with the transverse relaxation time whereas the change of the fast relaxation pool might relate to changes in the microstructural tissue composition.

2.4 Discussion

The current chapter is primarily focused on myelin water content mapping using an existing approach for T_1, T_2^* and absolute water content measurement with whole brain coverage [Neeb 2008]. The corresponding MR protocol is optimized for high spatial coverage and short acquisition time, T_{acq}, resulting in a compromised sampling density of the decay curve. Due to the sparse sampling density and the relatively low SNR, the non-negative least squares multiexponential analysis employed previously is not feasible here [Laule 2004, Oh 2007, Du 2007, Hwang 2010]. In contrast, the solution space has to be constrained in order to obtain reliable results. However, any constraint results in systematic errors given the incomplete knowledge about the true values. To minimize the bias, as much information as possible contained in the measured data is incorporated. Most importantly, results from a single exponential fit effectively constraint the position of the slow relaxation interval and the maximum amplitudes. The remaining three parameters are systematically varied to minimize the average systematic error over a wide range of relaxation times. It is important to keep the corresponding parameter values constant when comparing measurements performed on different scanners and/or different time points.

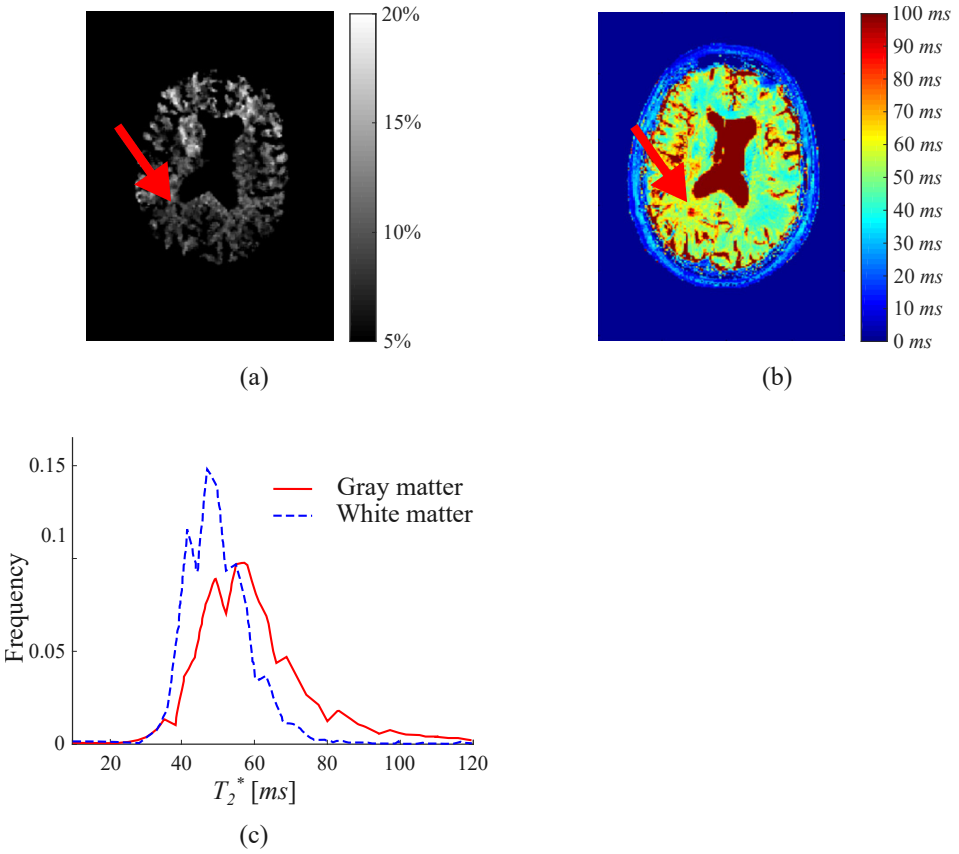


Figure 2.6: (a) Relative myelin water content map for the MS patient. The red arrows point to a visible MS lesion which is analyzed as described in the text. The corresponding T_2^* map is shown in (b) along with the T_2^* distribution in white and gray matter (blue dashed and red solid line in (c), respectively).

The experiments conducted show that only $T_2^{*,f}$ has a significant influence on the systematic error. Unfortunately, not very much is known about the real distribution of the fast relaxation time in cerebral tissue. Thus, a measured myelin water content of 0.1 has to be conservatively estimated with a maximum error bar of ± 0.015 for a single voxel. Nevertheless, the deviation averaged over the whole range of $T_2^{*,f}$ investigated is lower (± 0.009). In contrast, the systematic error is much less influenced by $T_2^{*,s}$ which again stresses the fact that the heuristic approach employed here to define $\widehat{T_2^{*,s}}$ works very well. In general, the systematic error decreases with increasing peak spacing. Furthermore, it is shown that the myelin water content is immune against the presence of a potential third relaxation component with long transverse relaxation. If one is primarily interested in the average myelin water content of white matter, the error estimate needs to be refined. Here, it is important to notice the symmetric deviation from the true myelin water content with respect to $T_2^{*,f}$ changes (Fig. 2.5). Assuming e.g. a flat distribution of $T_2^{*,f}$ in cerebral tissue, the systematic error would be close to zero as negative deviations are almost

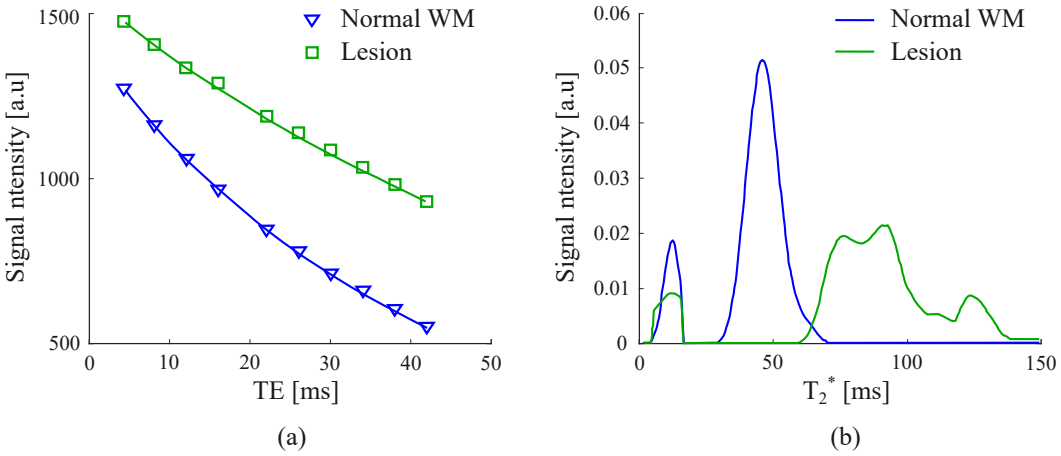


Figure 2.7: (a) T_2^* signal decay for voxels within the lesion shown in Fig. 2.6 (green solid lines and open boxes) and normal appearing white matter (blue solid line and open triangles). (b) Corresponding result from the linear optimization. Each curve represents the average over a ROI comprised of 42 voxels in normal appearing white matter and lesion, respectively.

perfectly compensated by positive deviations.

The approach described in the current work allows for a measurement of myelin water content using low SNR data. No averaging is performed to keep the measurement times short. Apart from the very fast relaxing compartments with $T_2^{*,f} < 35\ ms$, an SNR of 40 is sufficient to provide a reliable measurement within the error bars given above. Therefore, it can be expected that the approach should be equally applicable to 1.5 T data, especially given the increased peak spacing at lower field strength. Moreover, parallel imaging might be an attractive option to further reduce the measurement time as the SNR penalty is not expected to compromise the measurement accuracy.

One has to bear in mind that effects other than demyelination might contribute to an altered estimate of the corresponding pool fractions *in vivo*. Apart from the motility of protons, which forms the basis for the identification of myelin-bound water, the transverse relaxation process is influenced by local magnetic field inhomogeneities or changes in microstructural tissue composition. Potential sources which alter the relaxation times result e.g. from field gradients at tissue borders, differences in blood oxygenation, the deposition of paramagnetic substances or changes in water diffusion. The latter is primarily correlated with the packing density of myelin sheets which is heterogeneously distributed throughout white matter. A higher density subsequently results in a smaller T_2^* due to the restricted motility of water molecules [Mathur-De Vré 1984]. Furthermore, T_2^* depends on the orientation of axonal fiber bundles with respect to the external magnetic field [He 2009, Denk 2011, Lee 2011, Nam 2015, Lee 2017]. A perpendicular orientation of fibers results in a faster decay of the free induction decay (FID) signal than the corresponding parallel orientation [Denk 2011, Lee 2011]. In the current study, the

same tendency for both fast and slow relaxation times is observed. In the optic radiation, where fibers are oriented more perpendicular with respect to the main magnetic field, $T_2^{*,f}$ ranges between $8-9$ ms while $T_2^{*,s}$ is $\approx 43-47$ ms. Both relaxation times are higher in the semioval center, where fibers are oriented more parallel with respect to $\mathbf{B_0}$ ($T_2^{*,f} \approx 10-11$ ms and $T_2^{*,s} \approx 58-62$ ms, respectively). Therefore, the effect of fiber orientation seems to be most prominent for the slow relaxation time. However, no test for significance is performed and a more thorough investigation is necessary to investigate the corresponding regional differences between fast and slow relaxation times.

Although the relaxation process of water protons is altered by the processes described above, the determination of myelin water content is expected to be immune as long as the signal from both pools can be reliably detected and separated. Indirect evidence for the validity of the developed approach is provided by the *in vivo* results ($W_{rel}^{My} \approx 10-11\%$), which are consistent with results obtained by others [Laule 2004, Kolind 2009, Vavasour 2009, Oh 2007, Du 2007, Hwang 2010, Sirrs 2007]. In contrast, if pools would significantly overlap, an overestimation of the myelin water content would have been observed.

As the proposed method is based on multi-slice gradient echo data, two potential error sources might bias the myelin water measurement. First, it is important to notice that gradient echo sequences are sensitive to static magnetic field inhomogeneities. Therefore, the separation between fast and slow component decreases, resulting in a potential overestimation of myelin water content. Moreover, the decay signal deviates from the pure exponential behavior, rendering the estimation of pool-specific amplitudes difficult [Neeb 2006b]. As only $\approx 1\%$ of all voxels in the brain require a correction for off-resonance effects [Neeb 2008], the overall systematic error can be neglected, even though the bias in the regions affected might be considerable. This is consistent with the observed independence between myelin water content and the parameters which quantify the magnetic field distribution in white matter. However, the multi-slice measurement employed might still be biased by magnetization transfer (MT) effects due to off-resonance excitation. [Kalantari 2011] have investigated MT effects between brain water compartments using a 4-pool model including two semisolid proton pools. In this study, myelin water content is measured using the standard multi-echo T_2 based approach [Denk 2011]. However, an MT-pulse precedes the T_2 readout module with variable delay times in the range $18-768$ ms. They conclude that the influence of magnetization transfer on the myelin water fraction measurement is weak as exchange times between myelin and bulk water are much longer (> 1280 ms) than the corresponding pool specific relaxation times [Kalantari 2011].

To sum up, the approach presented in the current chapter basically offers the following three advantages:

- It is independent of specific hardware and/or proprietor sequence software and therefore readily available on every modern MR scanner.
- It allows for a full brain mapping in less than ten minutes.

- It allows for the simultaneous mapping of multiple parameters including total water content.

The latter might be especially attractive for the accurate assessment of myelination in MS lesions as most plaques are associated with increased total water content. Indeed, a considerable spread of total water content increase in MS lesions is observed here, ranging from 0% up to 25%. Therefore, an accurate evaluation of lesions should be ideally based on an absolute measurement of myelin water content, especially for the intra-individual comparison of different lesions or for the longitudinal follow up of a single plaque. However, an absolute measure is readily obtained using the approach described here as both relative myelin water content and total water content are quantified.

The results obtained demonstrates that tissue T_1, T_2^* as well as total and myelin water content can be mapped simultaneously with whole brain coverage in less than ten minutes. Without specific requirements on hardware or measurement sequence, the approach can be applied on almost every clinical scanner. This might help to expedite the further spread of quantitative MRI for both research and diagnosis of demyelinating diseases such as multiple sclerosis.

As this multiparametric quantitative data is particularly advantageous for the inter-subjects studies, it is of great interest to investigate whether cluster analysis could disclose some natural tendency of patients' grouping corresponding to the grade of disease impairment. The search for image-based features sensitive to the patients' disability status could help to cure the diagnostic and monitoring problems emerging from the clinico-radiological paradox observed in MS. An initial study concerning this issue is conducted by [Peneva 2013] who employs standard clustering techniques combined with different feature selection: K-means, which is a representative of the hard clustering algorithms group (the method is tested with different prespecified number of groups) and hierarchical clustering combined with an automatic determination of the clusters number (details concerning the input parameters as well as strengths and weaknesses of the different clustering algorithms are presented in the following Chapter 3). However, non of the applied methods could detect distinct meaningful groups in the investigated cohort of 54 MS patients. A possible reason for these results could be that the features selected are not suitable for the chosen analysis methods. Remarkably, the tested algorithms cannot properly handle correlated data since they employ the Euclidean distance metric as a measurement of similarity between objects. Due to the isotropic nature of this metric, these algorithms tend to detect spherical clusters, even if this does not correspond to the grouping tendency within the sample. However, as the features in the analyzed datasets are based on the multi parametric MRI acquisition presented above, it is expected that the attributes derived are biophysically correlated leading to the formation of non-spherical clusters. Furthermore, none of the tested methods is capable of recognizing outliers in the data, as they perform a so called complete clustering (cf. Sec. 3.2), where each point is assigned to a cluster. In the case of MS such outliers may be for example patients suffering from a severe injury impairment

and their assignment e.g. to the most closely positioned compact group might be unreasonable and result in misleading interpretations. Ideally, for this analysis a clustering approach should be able to detect groupings of irregular form and distinguish them unambiguously from outliers. This particular challenge has led to the development of the novel Nuclear Potential Clustering. Prior to presenting details concerning the algorithmic specifications and the validation of the approach, the next Chapter 3 gives a comprehensive discussion on the formulation of a clustering task and the existing techniques for unsupervised explorative data analysis.

Chapter 3

Introduction to Cluster Analysis

Contents

This chapter introduces the topic of cluster analysis and establishes a conceptual foundation for the development of a new clustering algorithm. First, it introduces the notion of learning and how this idea is translated to the processing of digital data. In this context the clustering task is formulated and different types of clusterings (i.e. algorithms' outcome forms) are discussed, with the focus put on the complexity of setting up an exploratory data analysis problem (Sec. 3.1). Next, an overview of the huge clustering algorithm family is given based on one of the most popular taxonomies of this field. The aim here is to form a concise, yet comprehensive overall frame in which the development of a new algorithm can be justified and properly positioned. In both, Sec. 3.2.1 and Sec. 3.2.2, the characteristics of different methodologies as well as their standard and most prominent representatives are discussed. Whenever possible in this context, the strengths and weaknesses are noted and current methods aiming to cure emerging drawbacks are presented.

3.1 Machine Learning and Formulation of a Clustering Task

The concept of encountering and dealing with data by organizing it in sensible groups is one of the most fundamental ways of understanding and learning. Generally, to understand a new phenomenon, people try to find the features that can distinguish and describe it, and then compare them to other possibly known phenomenon(s). Only by categorizing objects, processes, attributes and basically everything that surrounds and happens to us, we manage to process the constantly growing amount of information at all. This ability of the human thought is the base for optimizing our everyday routines and is actually one of the most primitive abilities of the human beings that has played an essential role in our history [Aderberg 1998]. Therefore, it

is only natural to transfer this notion of dealing with information to tasks concerned with digital data.

Learning processes and algorithms can be broadly divided in two types: supervised and unsupervised [Duda 2001]. The first type pertains to cases in which a new input is assigned to one of a finite number of classes using a function (classifier) designed by exploiting some *a priori* known information [Cherkassky 1998, Duda 2001, Theodoridis 2008]. In the following, a highly illustrative example taken from our everyday life is given: During the last year I very often used to teach my two years old son to tidy up his room at the end of the day. Every time he would hold a small ball or a toy brick, I used to say to him "This is a ball, please put it in the ball pit." or "This is a brick, please put it in the box." In the course of this process he has learned from me which and what the different categories are and after a while he was able to properly assign the objects himself (Fig. 3.1(a)). Data analysis algorithms based on this concept are referred to as supervised learning or (supervised) classification. Formally, such a classification tasks is associated with a set of N training examples $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), ..., (\mathbf{x}^{(N)}, y^{(N)})\}$, where each single input $(\mathbf{x}^{(i)}, y^{(i)})$ consists of the following two terms:

- d measured attributes[1], $x_j^{(i)}$, commonly represented as a column vector in a d-dimensional feature space $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, ..., x_d^{(i)}]^T$ and
- a label $y^{(i)} \in 1, ..., C$ incorporating *a priori* knowledge about the object, where C denotes the total number of class types.

In a typical classification routine, a new input $\mathbf{x} \in \mathbb{R}^d$ is mapped to the finite set of discrete class labels by means of a mathematical function $f = f(\mathbf{x}, \mathbf{w})$, which is modeled by a vector of adjustable parameters $\mathbf{w}$. The values of these parameters are specified by minimizing an empirical cost functional (related to some inductive principle) using the training set [Duda 2001, Theodoridis 2008].

In unsupervised learning, also referred to as clustering or exploratory data analysis or unsupervised pattern recognition, no labeled data is available. This is why clustering is considered to be the more difficult and challenging problem than classification [Jain 2010]. Just as classification, clustering objects is a natural way for humans to process data. However, in contrast to the predictive goal of classification, the goal of clustering is descriptive. Going back to the example of my son, I very often observe how he groups objects in a way meaningful to him by merely observing them and searching similarities and common attributes (Fig. 3.1(b)). But while humans are excellent in discovering clusters in two- and three-dimensional space, automatic algorithms for the processing of high-dimensional data are needed. Clustering algorithms are one of the standard and most widely spread ways of exploring data in probably every scientific fields ranging from engineering, computer

[1] The different types of attributes is a wide topic itself and it is indispensable in the context of data analysis. However, here the focus lies on the concepts and formalism of learning tasks. The interested reader is referred e.g. to [Theodoridis 2008].

(a)

(b)

Figure 3.1: Different types learning routines demonstrated by the acting of a child. (a) Supervised learning: given the labels to the objects, a toddler can learn from them to properly assign the objects himself. (b) Unsupervised learning: spontaneously grouping the toys in a way meaningful for the kid by merely searching common attributes (not exactly what an adult person would do, as he/she has already seen and comprehended that e.g. the roofs are typically on top of the houses, which is a kind of supervised experience).

sciences, life and medical sciences to geography, social sciences and economics [Aderberg 1998, Theodoridis 2008, Xu 2005]. They are employed for the partitioning of N unlabeled points or observations, each represented by a feature vector $\mathbf{x}^{(i)}$ in a d-dimensional feature space, into M groups in such a way that samples within a group are more similar to each other than to observations belonging to other groups. The input data can be represented as a matrix $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$, referred to as design matrix. Although there is no universal agreement about the definition of clustering, most of the proposed formulations are centered around the idea of exploiting similarity or dissimilarity, generalized as proximity [Theodoridis 2008]. These terms are quantified by means of proximity measures and different such measures are defined for the characterization of the different types of existing features. A short summary concerning this topic is given in [Xu 2005].

Generally, a final clustering solution depends on the one hand on the sample being analyzed and on the other hand on the specific method applied. Thus, different combinations of a data representation, proximity measure and clustering schema can lead to different outcomes. Concerning these aspects, it can be stated that the most beneficial results are to be obtained by combining the following three steps:

- Choose features encoding as much information as possible relevant for the task of interest, where a preprocessing routine may be necessary prior to feeding them into the data analysis algorithm.
- Select a proximity measure suitable to capture the similarity between the chosen features. A special attention has to be payed to ensure that no feature dominates the others and all of the chosen attributes have comparable contributions to the computed proximity measure. This aspect hast to be taken into account during the preprocessing step.
- Apply a clustering routine suitable for the combination of formulated task, chosen features and proximity measure.

Given the data to be clustered, a clustering task can be designed in different ways: either to divide the information into groups that are meaningful, or useful, or both. The definition of what exactly these two terms mean for a concrete clustering task is of crucial importance for the choice of a particular routine and the final result, respectively. As stated in [Jain 2010], "a cluster is a subjective entity that is in the eye of the beholder and whose significance and interpretation requires domain knowledge". In market segmentation for instance, a possible goal can be to find a structure in a population subset, which can then be used as a base for the development of a certain marketing strategy. If, for example, a dataset containing the height and the weight of a customer population is used to define four clothing sizes (small, medium, large, very large), then the data might look like the sample presented in Fig. 3.2 (a subset containing the first 200 observations from the SOCR Data Dinov 020108 HeightsWeights that can be found under `http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights`). There is

no natural clustering tendency in this case. In fact, both features are strongly correlated and form a single cohesive structure. But the goal is still to find four groups. In such a context, even in the case of badly conditioned clustering problem, where actually no well distinguished groups exist, the strategy of constructive grouping and artificially creating classes can still be useful [Dolnicar 2010]. However, the more intuitive notion of clustering, which is mostly used in the scientific field, is to divide the sample into classes that reveal some natural tendency of building groups with respect to the defined features. This view on a clustering problem is referred to as natural clustering in the literature [Jain 2010].

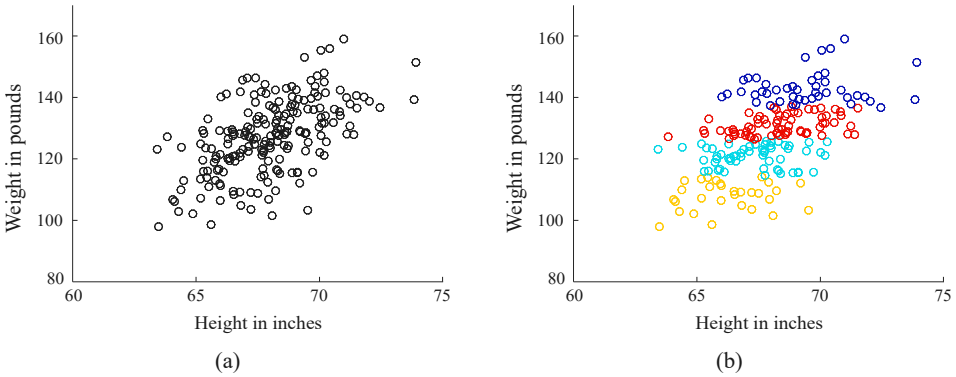


Figure 3.2: Contrived example for demonstrating of different clustering task formulations. (a) A sample of 200 measurements of height and weight. (b) A segmentation into four groups using K-means, useful e.g. for the definition of four clothing sizes or other marketing strategies.

3.2 Clustering Algorithms

The different clustering applications give birth to a vast range of conceptually different clustering paradigms. It is important to emphasize that there is no universal approach of solving any clustering task, but it is rather typical to investigate the characteristics of the problem at hand and choose or design an appropriate analysis strategy. This idea of no best clustering algorithm is formalized by the Impossibility theorem, which states that no single clustering algorithm simultaneously satisfies a set of basic axioms [Kleinberg 2003].

Since clustering algorithms have a truly interdisciplinary history, their development is rather inconsistent and it is not trivial to systematically order all the members of the clustering algorithm family. For this reason, different typologies exist and a summary of the criteria building the basis of the most common ones is

given in the following[2]. Some authors view the problem of systematically organizing the algorithms from a historical point of view, e. g. [Jain 2010, Xu 2005] who starts back in 1956 with the K-means [Steinhaus 1956], one of the most popular algorithms today even more than 60 years after its appearance. There is also an attempt to cluster the clustering algorithms so as to reveal which methods tend to have the same or different performance [Jain 2010]. Further, a differentiation can be made between *serial* and *simultaneous* methods depending on whether the patterns are handled one by one or as an entire set; *monothetic* and *polythetic*, depending on whether the features are used one by one or all at once; *complete* and *partial* referring to whether every object is assigned to a cluster or not; as well as *graph based* and *matrix based*, depending on which mathematic formalism builds the base for expressing the clustering algorithm [Jain 1988]. In the following two subsections, two further taxonomies are more explicitly explained.

3.2.1 Clustering typology based on the resulting structure

This differentiation is based on the form and structure of the results. It is more commonly met in the literature, as it directly concerns the formulation of an analysis task by the researcher. According to this criterion three different clustering types can be defined [Theodoridis 2008, Xu 2005]:

- *Hard* or also called *crisp* clustering is a procedure attempting to find an M-partition of $\mathbf{X}$, $C = \{C_1, ..., C_M\}$ with $M \leq N$ such that
  - $C_i \neq \varnothing$, where $i = 1, ..., M$;
  - $\bigcup_{i=1}^{M} C_i = \mathbf{X}$;
  - $C_i \cap C_j = \varnothing$, where $i, j = 1, ..., M$ and $i \neq j$.

  Here, each pattern belongs to a single cluster.

  The K-means algorithm is the best known representative of this group [Steinhaus 1956, Lloyd 1957, Lloyd 1982]. It can be very easily implemented for solving many practical problems and can work very well for compact hyperspherical clusters. Since it is widely applied, its drawbacks are very well studied as well. First, the number of clusters and an initial partitioning have to be specified, but there is no universal way to do this. Different works address this problem [Peña 1999, Bradley 1998, Likas 2003]. Another three classical initial partition methods are proposed by Forgy [Forgy 1965], Kaufman [Kaufman 1990] and Mac Queen [MacQueen 1967] and later compared with regard

[2] Please note that the aim of this summary is to give an overall conception of the existing clustering types based on their specifications, rather than listing all of the existing methods. The latter is a tremendously effortful task itself (a search for the term "clustering algorithm" e.g. in Google Scholar returns more than 350000 results) and is beyond the scope of this work. A series of valuable survey papers addressing this problem are [Jain 1999], [Baraldi 1999], [Fasulo 1999], [Jain 2000], [Jain 2010] and [Xu 2005].

to effectiveness, robustness and convergence speed by [Peña 1999]. One interesting technique called ISODATA adjusts the number of clusters by merging and splitting the data based on a predefined threshold [Kaufman 1990]. However, in this case the problem of specifying the number of clusters is replaced by the question of specifying the threshold.

Another problem is the sensitivity to outliers and noise. K-means forces every single point into a cluster, even if it is far apart from a cluster centroid and thus distorts the cluster shapes. Algorithms aiming at overcoming this problem are PAM [Kaufman 1990], ISODATA [Ball 1967] and K-medoids [Estivill-Castro 2000]. However, they still need to specify M, either directly or as mentioned above in the case of ISODATA via a threshold.

- *Hierarchical* clustering construct a tree-like structure partition of $\mathbf{X}$ called dendrogram, $H = \{H_1, ..., H_Q\}$ with $Q \leq N$, such that
  - $C_i \in H_m$, $C_j \in H_l$ and $m > l$ imply $C_i \in C_j$ or $C_i \cap C_j = \varnothing$ for all i, $j \neq i$ and $m, l = 1, ..., Q$.

  Here, each data point can belong to multiple subgroups nested into each other.

  Hierarchical clustering algorithms is subdivided into agglomerative and divisive methods, whereby in practice the first type is preferred, since it is computationally less expensive [Everitt 2011]. The different agglomerative methods employ different definitions of intra-cluster distance. The two most popular are the single linkage [Sneath 1957] and the complete linkage [Sørensen 1948] techniques. The common criticism for classical hierarchical clustering algorithms is that they are sensitive to noise, tend to form spherical shapes and are computationally expensive, thus having limited application to large-scale and high-dimensional data sets [Xu 2005]. These drawbacks are addressed by more sophisticated hierarchical techniques such as CURE which is designed to handle more complex cluster shapes [Guha 1998], ROCK addressing large data sets [Guha 1999], Chameleon aiming at better handling highly variable sizes and shapes [Karypis 1999], and BIRCH focusing on outliers, large data and dealing with noise [Zhang 1996b].

- *Fuzzy* or *soft* clustering algorithm of $\mathbf{X}$, into M partitions, such that
  - $u_j : \mathbf{X} \rightarrow [0, 1]$, $j = 1, ..., M$
  - $\sum_{j=1}^{M} u_j(\mathbf{x}^{(i)}) = 1$, $i = 1, ...N$
  - $0 < \sum_{i=1}^{N} u_j(\mathbf{x}^{(i)}) < N$, $j = 1, ..., M$.

  In this case a pattern is allowed to belong to all clusters with a certain degree of membership, specified by the so called membership function u_j [Zadeh 1965]. This type of clustering is especially useful when the boundaries between the

different classes are ambiguous. One of the most popular representatives of such algorithms are the fuzzy relative of the ISODATA [Dunn 1973] and its generalization fuzzy C-means (FCM) [Bezdek 1981, Höppner 1999]. Similar to its hard counterpart, K-means, FCM is prone to having problems with outliers and noise. An improvement addressing this issue is suggested by [Kersten 1997] and generalized by [Hathaway 2000] (both of them suggesting the employment of different distance measures), as well as by [Krishnapuram 1993], [Zhang 2004] and [Davé 1997]. A further problem of fuzzy clustering techniques is their computational burden for large scale data. This issue is addressed by [Kolen 2002] and [Hung 2001]. Last but not least, for the detection of different types of cluster shapes, a series of specifically developed methods exists including FSC (hyperspherical shells)[Bezdek 1992], FCSS (fuzzy c-spherical shells) [Krishnapuram 1992], FCR (fuzzy c-rings) [Man 1994], FCQS (fuzzy c-quadratic shells)[Krishnapuram 1995], and FCRS (fuzzy c-rectangular shells) [Hoeppner 1997].

In addition, an algorithm combining fuzzy and hierarchical clustering is proposed by [Geva 1999]. It addresses one of the major disadvantages of hierarchical clustering, namely the inability to reassign a point once it is ascribed to a cluster.

3.2.2 Clustering algorithms' typology based on the parametric nature

This cluster algorithm taxonomy is probably the most common in the machine learning field. It pertains to both, supervised an unsupervised algorithms and characterizes the algorithm's machinery itself. According to whether an approach has a fixed or a flexible set of parameters it is said to be *parametric* or *non-parametric*, respectively. The study of the theoretical fundamentals of these both types leads to some general recommendations as to which (type) clustering routine is suitable for a given data exploration task.

The use of parametric approaches requires some *a priori* knowledge about the data distribution, such as the cluster structure or the number of clusters. Such techniques can be viewed as cost based and/or model based depending on weather an empirical cost function is minimized (e.g. K-means [Forgy 1965, MacQueen 1967], K-medians [Bradley 1996], single linkage [Sneath 1957], minimum diameter clustering [Hochbaum 1985]) or they model the data in terms of a mixture of parametric densities and the objective is to find the model parameters typically by optimizing a cost derived from likelihood (e.g. the algorithms [Vempala 2004], [Dasgupta 2000], [Sanjeev 2001] and [Achlioptas 2005]). Two main issues concern the parametric unsupervised learning. The first one is the choice of the number of clusters M. Typically, the computed cost decreases when increasing M, so it cannot be directly employed to select M [Tan 2006]. A series of methods addressing this problem exist, e. g. the BIC criterion for mixture models [Nugent 2010], the method of [Achlioptas 2005], as well as X-means [Pelleg 2000], the gap statistic [Tibshirani 2000] or stability

methods [Ben-Hur 2002]. Another main problem is the handling of outliers. The general questions here are which points are outliers and whether the chosen number of clusters M is meaningful in the presence of noise. A series of "remedy" approaches are proposed, e.g. introducing a more robust cluster prototype such as in the case of K-medians [Bradley 1996], replacing Gaussian with a heavier-tailed distribution (e.g. Laplace) when concerning a mixture model [Nugent 2010], or in single-linkage not counting clusters with a number points less than some threshold [Sneath 1957]. However, each of these operations imposes some particular assumption on the data: form, size and number of the clusters. Generally, when applying a parametric clustering routine one has to pay special attention to the fact that wrong assumptions bias the final outcome and eventually result in misleading interpretations.

If no *a priory* knowledge about the data is available, non-parametric methods are applied, where some local criterion is used to characterize the clusters. Common representatives are density-based algorithms such as DBSCAN, a special case of DENCLUE, and grid-based clustering such as STING, GRIDCLUST, WaveCluster, Bang-Clustering, MAFIA. They all detect clusters as dense regions of objects surrounded by low-density areas [Aggarwal 2013, Duda 2001, Everitt 2011, Xu 2005]. The main advantage here is the ability to handle clusters of arbitrary shape and size while being relatively resistant to noise and outliers. One has however to bear in mind that the final clustering outcome can be strongly influenced by the accuracy of the density estimation and that density-based algorithms have limitations concerning high-dimensional data and data containing clusters of widely different densities. Both limitations are rooted in the key notion of locally assessed density, namely to consider the number of points per volume unit at a certain position in feature space. In DBSCAN for example for a particular point the density is estimated by counting its neighbors within a specified radius *Eps*. Since this parameter is fixed, regardless whether data-driven or user-specified, clusters of widely different densities cannot be detected. A further difficulty is that when the data is high dimensional the feature space is usually sparse. As the volume is exponential in the number of dimensions d, the density tends to zero, unless the number of points grows exponentially with d, as well. Thus, the traditional notion of density in such cases becomes useless. One possible resort in this case is to turn to dimensionality reduction techniques. Alternatively, methods that "refine" the density notion can be used such as e.g. DENCLUE where kernel functions are employed to model the density as the sum of the influences of individual data objects [Tan 2006]. Another techniques addressing this limitation are subspace clustering algorithms, which aim at finding clusters embedded in low-dimensional subspaces of the given high-dimensional data set (e.g. the grid-based method CLIQUE [Agrawal 1998]).

3.2.3 Nature-inspired Clustring Algorithms

One fascinating trend in the clustering algorithm development is to design methods based on imitating natural processes. This idea is not new in the scientific world. In fact, it builds the base for an entire discipline - bionics resorts to evolutionary

highly optimized phenomena to design modern technologies. Not surprisingly, a similar movement emerges also in the domain of data clustering.

Bio-inspired clustering methods As the name suggests, Evolutionary algorithms (EA) grasp a basic idea similar to that of bionics, however with the conceptual difference of imitating the evolutionary process itself, rather than directly resembling its highly optimized outcome. They are employed in a vast range of applications including clustering tasks which in this context are addressed as NP-hard grouping problems [Falkenauer 1998][3]. The underlying concept of EA is to address an optimization problem by treating candidate solutions (typically randomly initialized) as members in a population. An artificial evolution of this population takes place through repeatedly applying of the operators reproduction, mutation, recombination, and selection over a series of so called generations. This process is based on the optimization of some objective function, called fitness function, which guides the evolutionary search. Though such an implementation is a strong simplification of its natural paragon, the presumption is that forced by the environmental pressure an artificial Darwinian selection will take place leading to an increase of the population fitness. This process can be conducted until a candidate with sufficient quality is found or a previously set computational limit is reached [De Jong 2006]. Clustering algorithms based on EA are mostly addressing crisp partitional tasks with a fixed number of clusters M [Bandyopadhyay 2002a, Estivill-Castro 1997, Fränti 1997, Kivijärvi 2003, Krishna 1999, Krovi 1992, Bezdek 1994a, Kuncheva 1997, Lu 2004b, Lu 2004a, Lucasius 1993, Maulik 2000, Merz 2002, Murthy 1996, Scheunders 1997, Sheng 2004], hard clustering with no *a priori* given number of clusters where the goal is to optimize M and the corresponding partitioning [Cole 1998, Cowgill 1999, Bandyopadhyay 2001, Bandyopadhyay 2002b, Hruschka 2003, Casillas 2003, Hruschka 2004a, Hruschka 2004b, Hruschka 2006, Ma 2006, Alves 2006, Tseng 2001, Naldi 2007, Handl 2007, Pan 2007], as well as fuzzy-clustering approaches with fixed M [Hall 1994, Hall 1995, Klawonn 1998, Bezdek 1994b, Yuan 1995, Van Le 1995, Egan 1998, Hall 1999, Liu 1995] or methods introducing an optimization of M and the corresponding fuzzy-partitioning [Park 2005, Liu 2003, Maulik 2003, Pakhira 2005, Hruschka 2004c, Alves 2007, Campello 2009, Falkenauer 1998]. By comparison, there are considerably less EA hierarchical clustering methods: to my knowledge only [Chiş 2008] and [Lozano 1999] have published researches pursuing this problem. As stated in [Hruschka 2009] the reason for that lies most likely in the difficulty to define a fitness function able of guiding the evolution of a dendrogram. Main motivation for the employing of EA for clustering tasks is their effectiveness in providing near-optimal solutions for NP-hard global optimization problems. Still, in praxis, the specific design of an evolutionary algorithm (concrete choice and combination of operators, encoding schemes, parameters etc.) is of crucial importance to its success in solving effectively or not a given

[3] As shown in [Brucker 1978] a clustering problem with number of clusters more than three is NP-hard.

problem.

Swarm Intelligence (SI) is a relatively young interdisciplinary research field inspired from the collective intelligence of a group of social animals. When acting collectively bees and termites for example can accomplish very complex tasks, despite the strongly limited capabilities of every single individual. Since the initial introduction of SI in the context of cellular robotics [Beni 1989], two main SI research directions emerge: the Ant Colony Systems (ACS) [Dorigo 1996, Dorigo 1997] and the Particle Swarm Optimization (PSO) [Kennedy 1995]. Standard ACS clustering methods are [Lumer 1994] and [Handl 2003]. A series of modifications is also proposed aiming mostly at an improvement of the convergence rate as well as the automatic determination of the number of clusters M [Tsang 2006, Handl 2002, Ramos 2002]. A PSO-based clustering method is first introduced by [Omran 2002, Omran 2005]. Main advantage of PSO is the ability to recombine and compare several potential solutions simultaneously. However, PSO methods include user-defined constants and much work is devoted to the parameter selection problem [Taherkhani 2016, Eberhart 2000, Jiang 2007, Li-Ping 2005, Carlisle 2001, Shi 1998, Van Den Bergh 2006, Trelea 2003, Pedersen 2010]. A series of modifications, mostly hybrids with another data analysis methods are published, often being designed to serve a particular application. A hybrid ACS mit K-means is introduced by [Monmarché 1999], while [Kanade 2003] introduce a hybridization of ACS with FCM aiming to automatically determine M. More modifications of ACS are provided in [Tsang 2006]. Concerning PSO, a hybrid with K-means is also introduced by [Van der Merwe 2003], while [Xiao 2003] present a synergy of PSO with Self Organizing Maps for clustering of gene expression data. Furthermore, [Cui 2005] propose a PSO based hybrid algorithm for clustering of text documents and [Das 2006] develop a fuzzy PSO-based approach which aims at improving the choice of M, as well.

Somewhat not quite related to biology, but still considered to be swarm intelligence methods are the River Formation Dynamics [Rabanal 2007] and the Gravitational Search algorithm ([Rashedi 2009, Rashedi 2010], [Hassanzadeh 2010]). Though both of these research topics are not devoted to clustering directly, but are rather optimization methods, they reveal a further interesting research tendency, namely to design physics-inspired approaches.

Physics-inspired clustering methods Here, one of the most exciting topics is quantum clustering which relies on the "implication of quantum mechanics for information processing purposes" [Nielsen 2011]. Being conceptually completely different from its traditional counterpart, quantum information cannot be measured unfailingly and is disturbed by observation, it exists yet as a superposition of classical states. It builds the basis for the fascinating idea of constructing methods for computationally efficiently processing of huge data by a quantum computer, if we had one. Most of the work in this field is dedicated towards improving classical algorithms and introducing so called "quantized" counterparts which aim at

being significantly faster than the original versions. A series of so called quantum subroutines are introduced such as e.g. finding the minimum of a function or finding the two farthest points in a data set. These are used to translate a classical routine to a faster quantized version and this is already introduced for a series of popular methods, including K-means [Aïmeur 2007]. This has the advantage of using methods for which the strengths and the weaknesses are well studied. However, fact is that even being computationally more efficient, the quantized methods will still have the same demerits as the original ones. Furthermore, quantum computer history has virtually just started, the world's first fully operational machine is introduced in June 2015 by D-Wave (the corresponding press release can be found under `https://www.dwavesys.com/press-releases/d-wave-systems-breaks-1000-qubit-quantum-computing-barrier`). This technique has surely an incredible feature ahead of it but it is still a long way off, as one of the main reasons why a certain - not only clustering - approach gains popularity in the scientific community is simplicity and accessibility (if one thinks e.g. of the example of the K-means method and refers e.g. to [Jain 2010]).

In this context, another very interesting research avenue is the developing of an entirely new clustering approach based on physical intuition and tools derived from quantum mechanics, rather than speeding-up existing methods. This idea is first grasped by Horn and Gottlieb [Horn 2001b, Horn 2001a] who employ a quantum potential as a density substitute. Here, a Gaussian wave function is associated to each of the data points and the sum of all these terms, the total N-particle wave function ψ, is constructed. Conventionally, this resulting function can be viewed as a probability distribution that could have generated the observed points and its maxima can be regarded as determining locations of cluster centers. Quantum clustering, however, views it in a different way requiring ψ to be the ground state of a Hamiltonian (an operator in Hilbert space) represented by the Schrödinger equation of which the probability function is a solution. This equation contains a potential function that can be derived analytically from the probability function and is eventually used to associate its minima with cluster centers. While quantum clustering employs gradient descent to assign the data points to the nearest potential minimum, an extension called Dynamic Quantum Clustering (DQC), introduces a different final step. Based on the quantum evolution of the system a proxy for each datum is moved to the nearest potential local minimum. This algorithm is designed as a highly visual tool in which the clustering is terminated by the user. The authors put the stress on the ability of the user to see how and why structures form and to determine when the clustering process is complete. In this, they anticipate to reveal hidden structures that may otherwise not be recognized as a cluster. Applications of DQC on a series of real world problems show that such a strategy is a very good way of finding natural clusterings of the data [Weinstein 2009, Weinstein 2013].

Interestingly, this wish of finding out the natural tendency of building groups inside a sample gives raise to two further non-parametric synergy approaches resembling directly physical systems in which clusters appear spontaneously. Superparamagnetic Clustering [Blatt 1996] and Percolation Clustering [Šášik 2001] both

take the way of using statistical physics methodology for solving hard optimization problems. The first approach detects domains of aligned spins which are interpreted as clusters through the investigation of spin-spin correlations, measured in a super-paramagnetic regime. On the other hand, percolation clustering is based on the mutual connectivity of the points that defines the way fluids move through porous materials and thus detect pores which are connected.

In a similar spirit, the current work presents an approach for pattern detection by simulating a system of nucleons in which clusters appear naturally. The basic idea of the so called Nuclear Potential Clustering (NPC) is to treat the points to be classified as nucleons in an N-dimensional space and to model their dynamic behavior. A nuclear potential is associated to every "particle", so that points that are densely distributed in space fuse to form nuclei (clusters) when dynamically modeled. The method inherently allows for the detection of noise and for the proper grouping of high-dimensional data in correlated observations, which are typical for multiparametric life and engineering science dataset.

Chapter 4

Nuclear Potential Clustering Method

Contents

At the very beginning of this project, different ideas of employing physical knowledge for exploratory data analysis emerged. Here, the focus was invariably put on the handling of high dimensional skewed data[1] in the presence of noise, as this constitutes the target type of problems of the current work (see Chapter 1). Natural physical systems characterized by a dependence of the objects' interaction on their mutual distances might be suitable for this purpose, so that the collective behavior of the objects (here and in the following also referred to as particles) allows for the identification and separation of regions of high and low density. Among others, for example, modeling the phase transition of a thermodynamic system was considered as a configuration expected to allow for the detection of noise and outliers in multivariate data containing irregularly formed groups. However, the ultimately pursued idea of designing a tool inspired from nuclear fusion has the key advantage of not merely detecting anomalies, but also distinguishing the dense cohesive regions in a sample. In the following Sec. 4.1 the background physics concerning this topic is briefly introduced. It addresses mainly the nucleon-nucleon interaction and the composition of nuclei as this constitutes the basis of the developed data clustering

[1] In the context of cluster analysis this term indicates data containing irregularly shaped clusters. This has not to be confused with skewed classes as uses in the context of classification problems where one class is overrepresented in the training set.

method. Next, in Sec. 4.2, the Nuclear Potential Clustering approach is explained in detail. In the last Sec. 4.3 of this chapter, the specific implementation of the algorithm is discussed.

4.1 Physical Fundamentals

4.1.1 Nuclear constituents and their interaction

An atomic nucleus is the small, positively charged central part of an atom. It consists of A nucleons: Z positively charged protons and N uncharged neutrons, where Z and A are denoted as atomic and mass number, respectively. Nucleons, in turn, are composite particles themselves. They are built up of three quarks (a basic type of elementary particles) bound by one of the fundamental forces of nature, the strong interaction. The nucleon-nucleon interaction is a spillover effect of the strong force. Except for a very short repulsive core, it is of attractive nature, saturates and is electric charge-independent. It is this short-ranged strong interaction that governs the process of nuclear fusion. In addition, a Coulomb interaction is effective between the likely charged protons. Though it is much weaker than the strong force at short distances (which is the reason why nuclei can be bound), it is of much longer range. The overall nucleon-nucleon interaction is a combination of these two effects. Its strength varies with the mutual distance of the particles.[2] Figure 4.1 shows a schematic representation of the nuclear potential energy and indicates the dominance of the strong attraction for ranges smaller than a few Fermi (the natural length scale in nuclear physics, see Table 4.1).

Table 4.1: Some basic properties of nucleons.

	charge	mass (u)	radius (fm)	spin ($\hbar$)
proton	e	1.007276	0.85	$\frac{1}{2}$
neutron	0	1.008665	0.85	$\frac{1}{2}$

$1\text{u} = 1.6605 \times 10^{-27}\text{kg}$
$1\text{fm} = 10^{-15}\text{m}$
$e = 1.6022 \times 10^{-19}\text{C}$
$\hbar = 1.0546 \times 10^{-34}\text{Js}$

The fact that two forces of different strength, range and opposite sign are responsible for the build-up of nuclei, is of crucial importance for the nuclear stability. Unstable species, called radionuclides, undergo disintegration by emitting a particle and can transform thereby to a different nuclide. As it can be seen in Fig. 4.2(a), in light nuclei the average binding energy per nucleon increases with mass number A. However, the size of the nucleus increases thereby, as well. Since the strong

[2] It furthermore can depend on other variables such as the spin and momentum of the nucleons. However, since these effects are not utilized in the current work, no further details about them are discussed in the following.

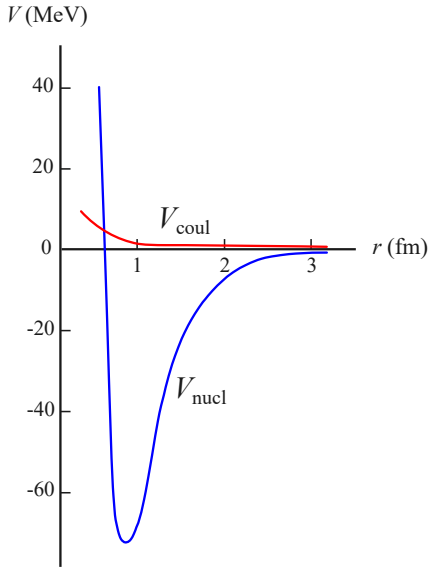


Figure 4.1: Potential energy for a system of two nucleons.

force is short-ranged, nucleons interact directly only with their nearby neighbors. Thus, the binding energy of a nucleon due to the total nuclear force tends to approach a constant value at large mass numbers A. On the other hand, the repulsive Coulomb force acts over much longer distances which leads to a constant increase of the potential energy *per proton* as the atomic number Z increases. Due to these two phenomena, the average binding energy per nucleon in natural nuclei increases up to $A \sim 60$ and eventually decreases as shown in Fig. 4.2(a).

The presence of neutrons in a nucleus has generally a stabilizing effect. They experience and are the source solely of the strong binding force, thus counteracting the repelling electrostatic interaction that acts between protons. This is particularly necessary in the case of heavier elements, where the repelling character of the cumulative Coulomb force becomes increasingly dominant. To maintain stability in this case, the proton-neutron ratio Z/N is shifted to lower values, so the most energetically stable configuration for a heavy element typically exhibits a neutron excess. In Fig. 4.2(b), a so called nuclidic chart is presented in which the stability of the known nuclei is depicted as a function of the number of neutrons N and protons Z. It can be seen that the curve of stability for light nuclei sticks to the $Z = N$ line and bends away from it as Z increases. However, this compensatory mechanism has an ultimate limit. For reasons rooted in quantum physics, getting increasingly neutron-rich makes nuclei unstable, as well. According to the nuclear shell model, nucleons are allowed to occupy discrete quantum energy levels. Protons and neutrons are treated in this regard separately, each particle type has its individual set of quantum states. Being half-integer particles (refer to Table 4.1), nucleons obey the Pauli exclusion principle that states that no identical fermions can occupy the same quantum state. Thus, in cases of a too low proton-neutron ratio Z/N, part of

the neutrons are forced to be in states with high energies, making the system less stable.

For these reasons, no stable element with Z greater than 83 occurs naturally.

4.1.2 Natural Paradigm of NPC

Nuclear Potential Clustering is inspired by a process that is assumed to have taken place about 100 seconds after the Big Bang. According to Georges Lemaitre's theory, the universe expanded explosively from an extremely dense and hot state, and continues to expand ever since. A graphical illustration of the main events taking place in the evolution of the universe is depicted in Fig. 4.3. In the first moments of its existence, the universe started to cool down and the first building blocks of matter emerged - quarks and electrons. A few millionths of a second later, the quarks aggregated to produce protons and neutrons. Within the next three minutes these nucleons fused to form the first ever existing nuclei. Exactly this nucleosynthesis process embodies the prototype of the NPC method. Data points represent herein nucleons. Their position in space, more specifically in feature space, is specified by their domain specific attributes. In a dynamic simulation, the physical behavior of such a system is then modeled. Its goal is that data points distributed densely in space fuse to build nuclei, while outliers and noise are repelled. Such an artificial fusion is likewise governed primarily by the strong interaction. Due to its short-range nature, it is exerted primarily on directly neighboring data points. Thus, an unsupervised machine learning approach based on this logic will tend to detect natural clusters in the sense of the formulation of a clustering task as discussed in Chapter 3 on page 31.

In contrast to its natural paradigm, the proposed nuclear fusion simulation has to take the following aspects under consideration:

- There should be no size limitation for a fused nucleus. No cluster size has to be preferred. During the early nucleosynthesis, temperatures were extremely high (10^{10}°C) and after the Big Bang only light elements are formed (deuterium, helium, lithium). In contrast, in a data clustering approach, any kind of cluster size preference not rooted in the data sample itself, is unreasonable and should be avoided.

- All particles must interact in the same way. In the particular realization of the NPC method, all data points represent protons as the Coulomb term of the potential energy is employed for the detection of outliers. Therefore, neutrons which do not feel the electromagnetic interaction are not modeled by NPC.

- All formed nuclei must be stable. This has to hold true even in the case of a big nucleus consisting solely of protons (see previous item).

- The range of the strong interaction should be data-specific and sensible to the assessed features.

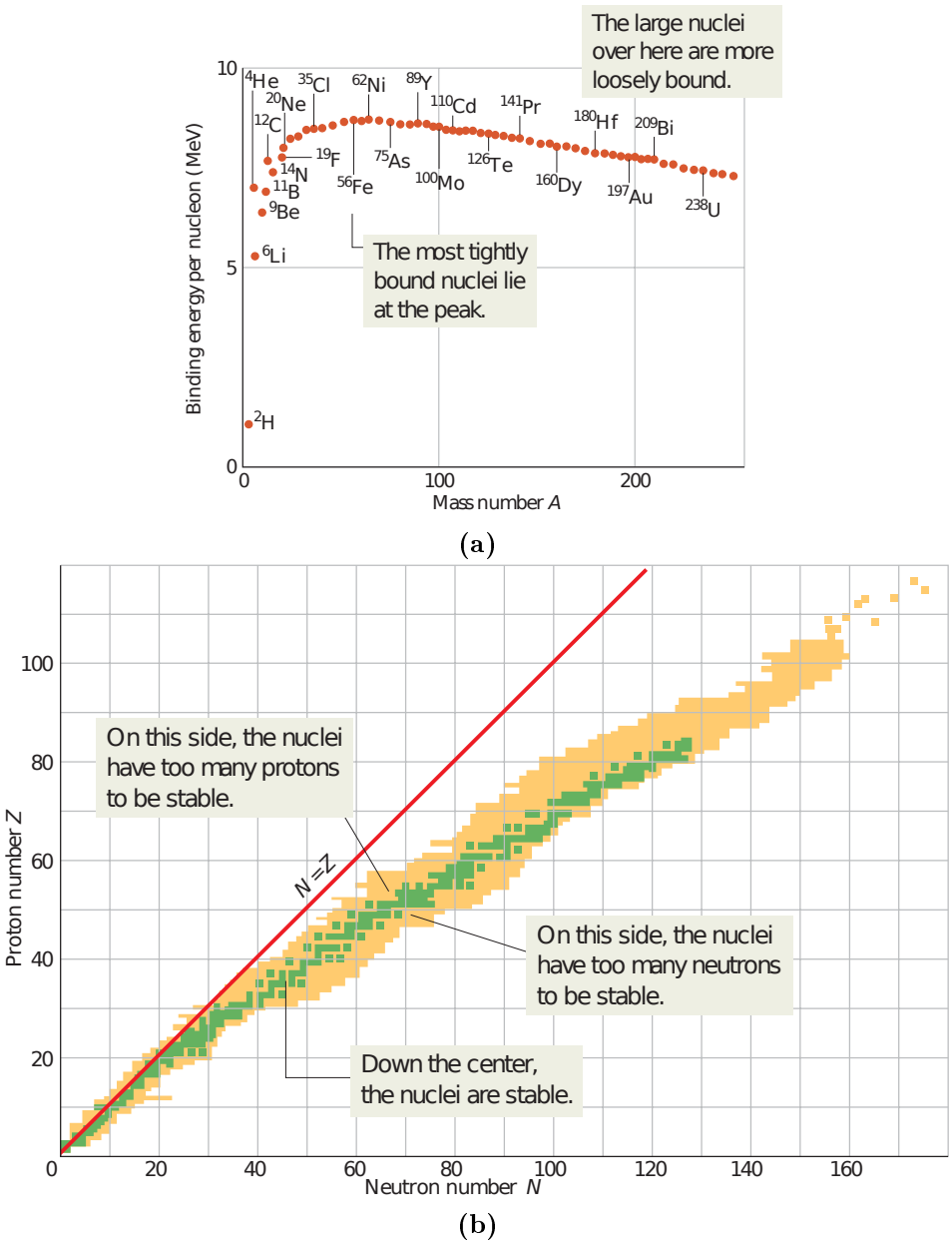


Figure 4.2: Stability of natural nuclei (image source: [Halliday 2014]): (a) Binding energy per nucleon as a function of mass number A. (b) A plot of the known nuclides. The green shading identifies the band of stable nuclides, the beige shading the radionuclides. Low-mass, stable nuclides have essentially equal numbers of neutrons and protons, but more massive nuclides have an increasing excess of neutrons. The figure shows that there are no stable nuclides with $Z > 83$.

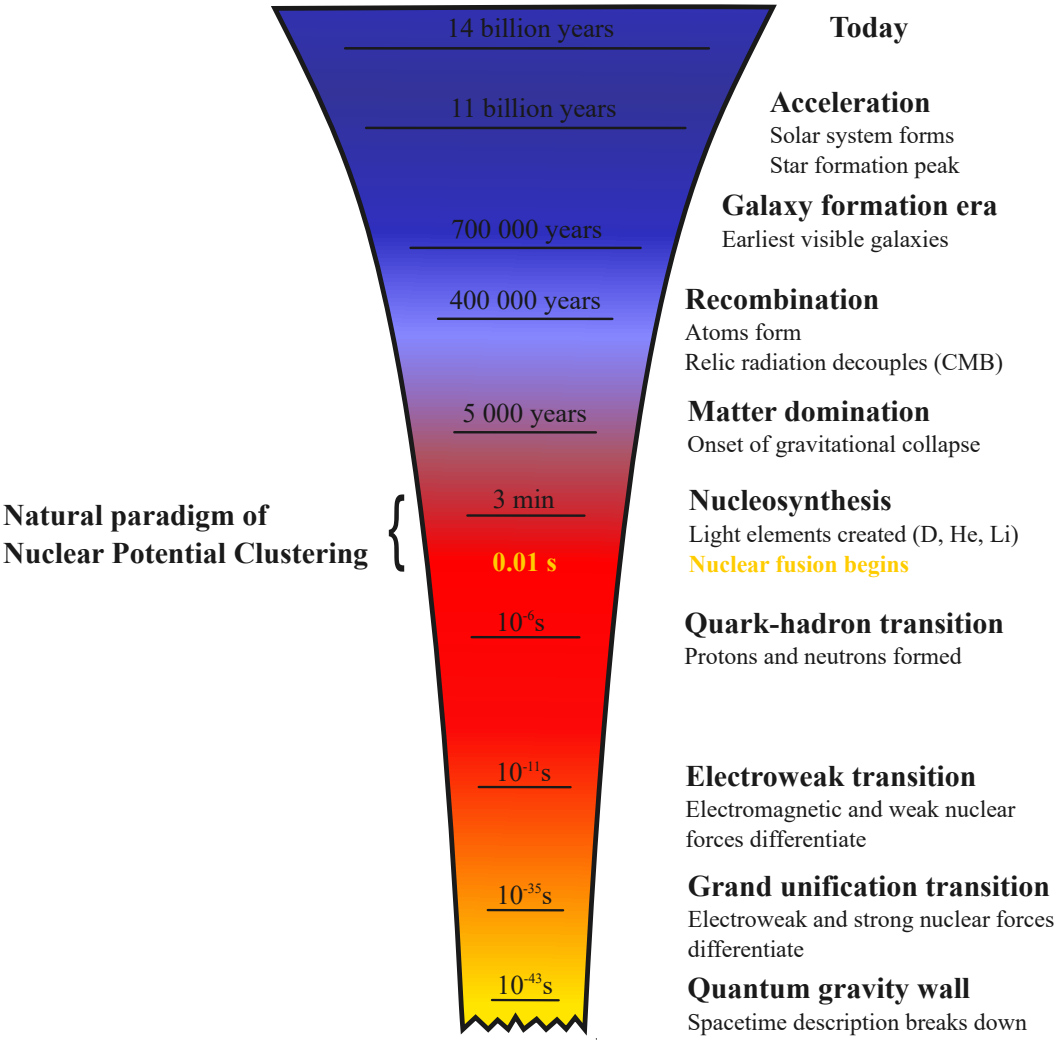


Figure 4.3: Graphical illustration of the main events which have taken place in the history of the universe. The Nuclear Potential Clustering approach is inspired from the process of nucleosynthesis (image inspired by `http://www.ctc.cam.ac.uk/images/contentpics/outreach/cp_universe_chronology_large.jpg`).

- Last but not least, for the purposes of computational efficiency, the developed method should represent a reasonable simplification of its natural paradigm. Thus, no shell structure is emulated. Instead, for the purposes of cluster identification, nucleons experiencing mutual attraction are allowed to fuse into a single point in feature space subsequently occupying a single energy level.

Details about the concrete realization of the method are given in the following section.

4.2 Nuclear Potential Clustering Approach

4.2.1 Nuclear Potential

Nuclear Potential Clustering is based on the dynamic modeling of a physical system in which each data point of the d-dimensional dataset represents a single proton. Each point is the source of a central potential $\phi(r)$ comprised of two terms. First, a short range strong interaction defined by a Woods-Saxon ansatz

$$\phi^s(r) = \frac{-V_0}{1 + e^{\frac{r-R}{a}}} \tag{4.1}$$

is responsible for the mutual attraction of particles in dense regions. Here, the interaction strength is defined by the magnitude of V_0, r defines the distance to the observation point in feature space, R is the range at which the strong interaction is effective and a defines the slope of the potential at the border where it starts to increase from $-V_0$ to 0 [Woods 1954].

At separations larger than the width of the above defined potential well, $r > R$, the nucleons feel merely their mutually repulsive electromagnetic interaction

$$\phi^C(r > R) = \frac{V_C}{r} \tag{4.2}$$

where V_C specifies the normalized electromagnetic coupling constant[3]. Figure 4.4 displays a schematic representation of the effective potential of a point as function of the distance, $\phi(r)$.

As previously mentioned, no neutrons are involved in the simulation. Thus, in contrast to the previous Sec. 4.1, where N denotes the number of neutrons in a nucleus as typically done in the context of nuclear physics, here and in the following this parameter will denote *solely the size of the sample to be analyzed* as defined in Sec. 3.1, i.e. the number of data points.

Given the initial position of the protons, as defined by their feature vectors

$$\mathbf{r_i}(t_0), i = 1, \ldots, N \tag{4.3}$$

[3] In physics, a $V_C = \frac{1}{4\pi\varepsilon_0} Q_1 Q_2$, with charges Q_1 and Q_2 of two particles and $\varepsilon_0 = 0.854\ldots 10^{-12} \frac{\mathrm{As}}{\mathrm{Vm}}$ being the vacuum permittivity.

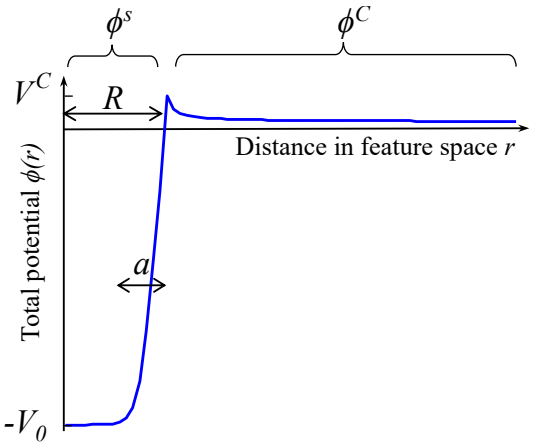


Figure 4.4: Schematic representation of the potential of each data point in a sample as a function of the distance r. It is viewed as a proton whose total potential $\phi(r)$ is comprised of a Woods-Saxon and a Coulomb term denoted by ϕ^s and ϕ^C, respectively. The parameters V_0 and R characterize the strength and the range of the strong interaction, respectively. a defines the thickness of the transition zone between attractive and repulsive term, and V_C specifies the electrostatic potential energy barrier, also called Coulomb wall.

the effective potential at the location of each particle is given by the superposition of the potential of the remaining $N-1$ particles from which the Newtonian force at spatial location $\mathbf{r_i}$ can be calculated from

$$\mathbf{F_i}(t) = -\nabla\phi(\mathbf{r_i}, t) \tag{4.4}$$

As a result, the force between nearby protons is dominated by the strong interaction, resulting in a mutual attraction. On the other hand, particles at larger distances experience solely a cumulative repelling electromagnetic force. Thus, when dynamically modeling such a system, nearby points in space fuse to larger nuclei, representing the clusters, whereas single nucleons, i.e. outliers, move apart from the dense regions of feature space. Figure 4.5 shows a two-dimensional data set and the corresponding potential energy of the system at the initial time point of the simulation, t_0. As it can be seen, nucleons in dense regions form an effective negative potential sink. In contrast, noise points are located on a positive potential plateau, thus not being accelerated towards the local potential minima in feature space. In addition, a Coulomb barrier between adjacent groups builds up, which effectively prevents neighboring clusters from merging.

The default values characterizing the potential function (Eq. 4.1 and Eq. 4.2) are empirically chosen by testing the performance of the algorithm on a series of test datasets with different number, shape and density of cluster as well as different dimensions of the feature space: $V_0 = 3500$, $V_C = 10$, $a = 0.05$ and $R = 10\%$ quantile of the initial distance distribution between all points. The range of the strong interaction is thus adapted to the dataset to be clustered. Since the NPC

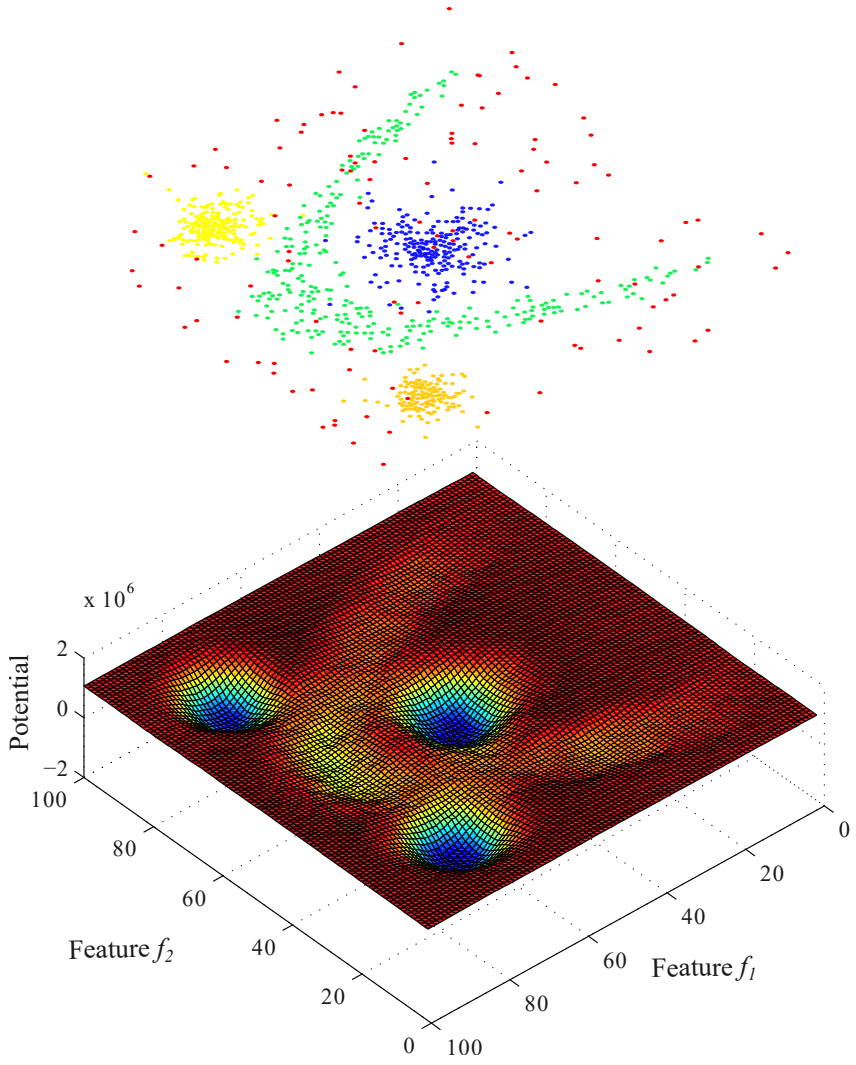


Figure 4.5: Formation of the total potential energy field in feature space (lower graph) as determined by the spatial distribution of the data points in a sample (upper scatter plot). Particles located in dense regions form an effective potential sink, whereas outliers "lie" on a potential plateau. Each dense group of particles is delimited by a Coulomb barrier.

parameters take on relative values, here and in the following, all parameters are assumed to be unitless.

4.2.2 Dynamic Modeling

To explore the dynamical behavior of the given system, the Newtonian equation of motion

$$\frac{d^2}{dt^2}\mathbf{r_i}(t) + \eta\frac{d}{dt}\mathbf{r_i}(t) = -\nabla\phi(\mathbf{r_i}, t) \tag{4.5}$$

is solved. Here, the damping term $\eta\frac{d}{dt}\mathbf{r_i}(t)$ with $\eta < 0$ is introduced in order to suppresses scattering of protons with kinetic energies exceeding the maximum potential energy of a cluster of points. Furthermore, it leads to a quicker transition to equilibrium state, making the algorithm significantly faster.

Typically, points do not directly form stable nuclei but rather oscillate towards an equilibrium state. The precise system development is revealed by its temporal temperature evolution, which typically exhibits a behavior similar to that shown exemplarily in Fig. 4.6(b). To terminate the dynamic modeling of such oscillating systems, an objective stopping condition is needed. For this purpose, the system temperature as a function of time, $T(t)$, is employed. $T(t)$ is determined by the average kinetic energy of all particles:

$$T(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{d}{dt}\mathbf{r_i(t)}\right)^2 \tag{4.6}$$

The simulation is stopped when T is less than 0.1% of the absolute temperature maximum, T_{max}, reached during dynamical modeling. Furthermore, an additional breaking condition is implemented that stops the NPC algorithm when the first derivative of the temperature remains $\frac{d}{dt}T(t) = 0$ over more than ten consecutive time steps.

The Newtonian equation of motion Eq. 4.5 is solved by default with a fourth order Runge-Kutta method with time step size calculated as follows:

$$\Delta_t = 4\sqrt{\frac{bd_{min}a}{NV_0}} \tag{4.7}$$

Here, the value of d_{min} is data specific and is given by the minimum mutual particles' distance in the sample. b is a local scale factor whose value is set to 0.001. Its estimation, as well as different ODE solvers that are implemented and can optionally be used, are extensively discussed in Sec. 5.4.

4.2.3 Potential Modulation

In regions with local density fluctuations, the cumulative repelling Coulomb force can result in a splitting of larger clusters into subgroups. To prevent such an artificial partitioning, the range of the strong interaction R in Eq. (4.1) is modulated during

dynamic modeling as follows :

$$R(t_i) \leftarrow s_R R(t_0) \tag{4.8}$$

The scaling factor is chosen to be $s_R = 1.5$, as larger values do no longer represent the intra-cluster distances properly and therefore may lead to the fusion of distinct adjacent clusters into one larger group. This issue is explicitly discussed in the next Chapter 5.

Furthermore, the parameter a is increased to $a = a_{max}$, where $a_{max} = 0.06$ is an empirically chosen value, discussed in the following chapter.

The modulation of R and a takes place after the first inflection point of the temperature curve is reached, as this is exactly the moment before points inside of the cluster(s) start to repel. Omitting this step "sticks" some of the protons in local minima of the hyper-dimensional potential field, leading to the formation of fragmented clusters.

4.2.4 Cluster Assignment

At equilibrium, the nucleons form almost dot-formed nuclei. Thus, protons belonging to the same cluster have approximately equal final potential energies. Therefore, the number of clusters in a data sample, M, is given by the number of maxima in the final potential energy distribution, resulting in a set of M discrete labels:

$$\mathcal{L} = \{l_1, l_2, ..., l_M\} \tag{4.9}$$

Points with a positive potential energy are classified as outliers as they experience solely the repulsive Coulomb term. The rest of the particles are assigned a label according to the final potential energy minimum they are attracted to.

As the simulation starts, particles belonging to a cluster are accelerated towards the potential minimum corresponding to that particular group. With increasing time, they might repeatedly by-pass the desired potential minimum. If the simulation is left to reach the absolute equilibrium sate, all of the nucleons belonging to a cluster would strive and concentrate into a single point in feature space, which is representative for this group. As an objective termination condition is needed to increase the computational speed, however, the process may be terminated at a moment when some single points still slowly oscillate through that minimum. They typically have a potential higher than the value of the potential minimum it is accelerated to. For this reason, the minimum potential value ever reached by the nucleons during simulation is employed for cluster assignment. This is assumed to be the particle equilibrium state. The number of clusters is therefore given by the number of peaks in the final potential energy distribution and each data point is given a label according to the potential minimum it falls into.

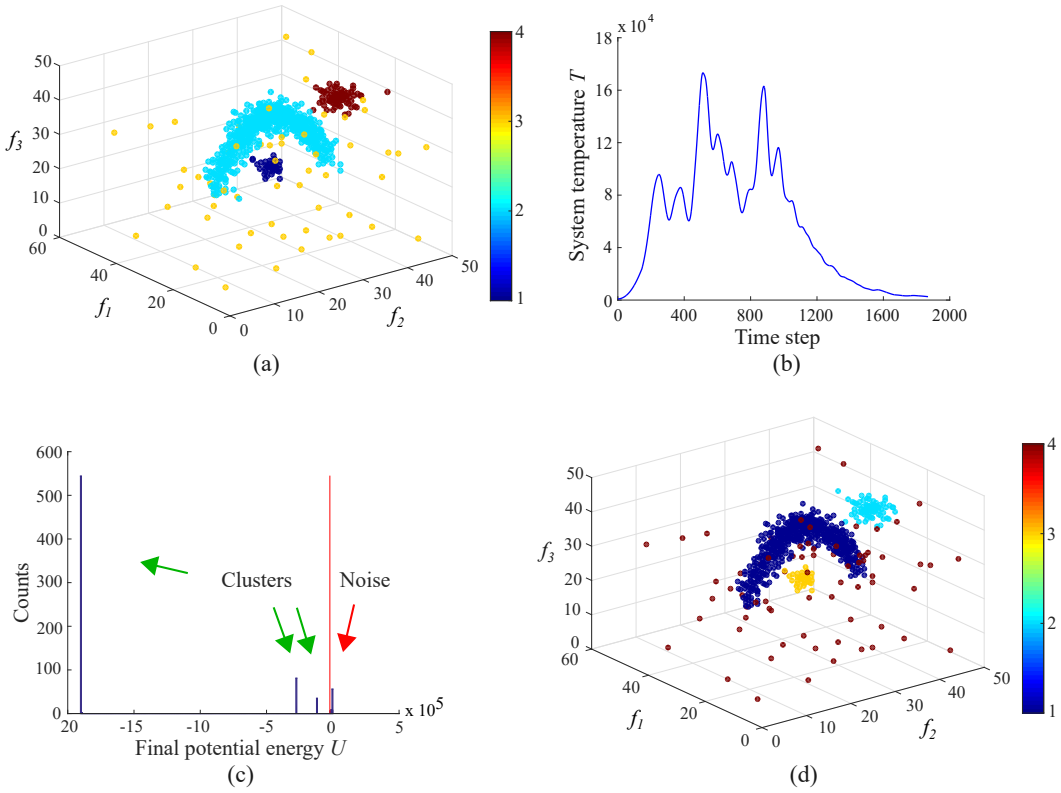


Figure 4.6: NPC cluster assignment demonstration. (a) Original data including two speroids and an elongated formation overlaid by random noise. (b) Temperature evolution of the system assessed during simulation. (c) Final potential energy distribution employed for cluster assignment. (d) NPC results where the data points are grouped according to their final potential energy.

4.2.5 Minimizing Misclassification of Outliers

In order to reduce the incidental assignment of noise points to one of the clusters formed, the NPC procedure can be repeated multiple times with different initial particle velocities. During the first run, particles start at rest. For the following repetitions, the initial velocities are chosen to have random direction. The corresponding absolute values are sampled from a normal distribution with zero mean and standard deviation given by the average velocity calculated from the previous step. The number of iterations, K, is variable and can be increased in cases when noisy data is expected.

Running the algorithm K times results in a set of K final potential energy values for the i-th particle:

$$\{U_i^{(1)}, U_i^{(2)}, ..., U_i^{(K)}\}. \tag{4.10}$$

To obtain the partitioning of the input data, the averaged final potential energy is determined:

$$U_i = \overline{U_i^{(k)}} = \frac{\sum U_i^{(k)}}{K}, k = 1, ...K \tag{4.11}$$

This value is employed for cluster assignment as described in Sec. 4.2.4

4.2.6 Nested Clusters

After defining all clusters using either a single or multiple runs of the NPC routine, the whole procedure can be repeated for each of the groups found. In this way one can explore if sub-structures within a cluster are present in a manner similar to the hierarchical clustering, however having the advantage of being capable to handle correlated attributes and sub-formations of irregular shape. When taking advantage of this additional action, one has to bear in mind that the NPC algorithm intrinsically detects noise. Thus, if the analyzed sample contains formations, where the density falls off e.g. at the rim of the clusters, points located in those regions are categorized as outliers. If one is interested in finding solely subgroups without a "zoomed-in" search for noise in the sub-clusters, an optional additional second last step can be run. Here, the centroids of the found sub-groups are calculated based on the final *positions* of the nucleons after the system has reached state of equilibrium. The points classified as noise are then assigned to the cluster with the nearest center of mass.

4.3 Algorithm Implementation

All of the steps presented in Sec. 4.2 can be assembled as displayed by Algorithm 1.

In the given framework, the Euclidean distance metric is applied. Thus, it is important to properly preprocess the data to be clustered. Otherwise, due to the isotropic nature of this measure, features of vastly different scales would lead

Algorithm 1 Nuclear Potential Clustering algorithm

1: **read** design matrix $\mathbf{X}^{N\times d}$
(optional: mean normalization of features)
2: calculate distance matrix $D = (d_{ij})$
3: estimate minimal nonzero particles' distance in the sample d_{min}
4: determine initial range of strong interaction $R(d_{min})$
5: initialize V_0, V_C, a, η, ρ, Δ_t
6: **for** $k = 1$ **to** K **do**
7: start procedure with different initial particles' velocities
8: **repeat** dynamically modeling the system
9: **for** the i-th particle, $i = 1, ..., N$ **do** identify interaction with remaining nucleons
10: **for** the j-th particle, $i = 1, ..., N-1$, $j \neq i$ **do**
11: estimate potential and force at position $\mathbf{r_i}$ resulting from j-th neighbor
12: accumulate $\phi(\mathbf{r_i}) \leftarrow \phi(\mathbf{r_i}) + \phi(d_{ij})$
13: accumulate $\mathbf{F}(\mathbf{r_i}) \leftarrow \mathbf{F}(\mathbf{r_i}) + \mathbf{F}(d_{ij})$
14: **end for**
15: **end for**
16: solve $\frac{d^2}{dt^2}\mathbf{r_i}(t) + \eta\frac{d}{dt}\mathbf{r_i}(t) = -\nabla\phi(\mathbf{r_i}, t)$ within time interval $[t, t+\Delta_t]$
17: calculate system temperature $T(t)$
18: **if** first flex point of temperature curve $T(t)$ reached **then**
19: $R \leftarrow s_R R(t_0)$
20: $a \leftarrow a_{max}$
21: **end if**
22: detect lowest potential energy $U_i^{(k)}$ of a nucleon in k-th iteration
23: $t \leftarrow t + \Delta_t$
24: **until** breaking condition, $T(t) \leq 0.001 \max T(t)$
or $\frac{d}{dt}T(t) = \frac{d}{dt}T(t - 10\Delta_t) = 0$ is reached
25: **end for** (iteration number K reached)
26: calculate average final potential energies $U_i = \overline{U_i^{(k)}}$ of each particle
27: detected peaks of the averaged potential energy distribution of all N particles
28: assign cluster labels to observations according to their final potential energy

to a shift in the definition of the parameter R and hence the potential function, allowing attributes with larger values to dominate the outcome of the algorithm. Thus, an optional zero-mean and unit-variance standardization of the input data is implemented. In this case, each feature is normalized as follows:

$$\mathbf{x_i} \leftarrow \frac{\mathbf{x_i} - \mu_i}{\sigma_i}, i = 1, ..., d \tag{4.12}$$

where μ_i and σ_i are the distribution mean and the standard deviation of the i-th attribute.

Concerning the computational cost, the most time-consuming step is presented by lines 9 to 15 in Algorithm 1, as the force and potential update of all particles (for each time point) has a computational complexity of $\mathcal{O}(N^2 d)$. This is due to the fact that the resulting interaction on each of the N nucleons is a cumulative pair-wise interaction with all of the rest $N-1$ particles and the effective force is a vector sum in a d-dimensional feature space. Furthermore, the runtime increases linearly with

the number iterations K, as well as with the number of groups identified, when an additional search for subgroups is conducted as described in Sec. 4.2.6. However, no exact determination of the runtime is possible, since the stopping condition for the dynamic modeling depends on the individual behavior of the system and the time it needs to cool down and reach an equilibrium state.

Chapter 5

NPC Parameter Optimization

Contents

The current chapter concerns about the description of parameters which influence the performance of the Nuclear Potential Clustering method. Their determination involves much experimental work and is not a straightforward procedure, but rather a recursive one. Most of the decisions considering the NPC subroutines are introduced after some particular effect is observed, so as to achieve optimal clustering results (e.g. potential modulation, damping). In the conducted tests, both the dynamic simulation and the final clustering outcome have to be studied. For this purpose, test samples with different characteristics are generated as described in Sec. 5.1. The procedure of visualizing and assessing the final clustering outcome are described in Sec. 5.2 and Sec. 5.3, respectively. The rest of the chapter addresses the process of estimating the optimal set of default values for all NPC parameters.

5.1 Generating Test Data

Two types of test data are generated for the performance evaluation. The first type (also referred to further in the text as Type 1 test data) consists of three-dimensional samples containing one bent horseshoe-formed cluster and two spheroids whose size and number of points can be varied. An example of this type test data can be seen in Fig. 4.6(a). The elongated formation (depicted in magenta in Fig. 4.6(a)) consists of n_h points positioned along a parable. Their radial distances to the parable are sampled from a normal distribution and scaled by a variable radius r_h. The spheroids are constructed by points sampled from a normal distribution, as well, with the radial distances to their centers being scaled by the variables r_{is} and r_{os}. The number of points in each spheroid is specified by n_{si} and n_{so}, respectively. In all of the conducted tests, one of the spheroids is positioned inside of the parable-formed cluster and the second one outside of it. In this way toy problems are constructed which represent nontrivial clustering tasks since the intra-cluster distances can be considerably higher than the inter-cluster ones.

Table 5.1: First type of test data employed for NPC parameter estimation (test data of Type 1).

Cluster	Adjustable parameter	
	Radial distance	number points
Horseshoe	towards the parable r_h	n_h
Spheroid outside	towards the center r_{so}	n_{so}
Spheroid inside	towards the center r_{si}	n_{si}
Noise	$\varnothing$	n_{noise}
Note: Generated test data are three dimensional.		

The second type of test samples (also referred to further in the text as Type 2 test data) is constructed of two elongated cylindrical clusters. They consist of points positioned along two skewed lines, g_1 and g_2, with variable radial distances to the lines as sampled from a normal distribution with standard deviations r_{g_1} and r_{g_2}, respectively. This configuration can be expanded to higher dimensions and allows to examine the performance of NPC as a function of dataset dimensionality. In addition, by varying the distance between the lines, the radii of the clusters, r_{g_1} and r_{g_2}, and the number of points, n_{g_1} and n_{g_2}, the performance in the presence of touching or overlapping clusters can be elaborated.

In both types of test data, the cohesive groups are overlaid by variable random noise specified by the parameter n_{noise}. The variables that control the specific appearance of a generated sample are summarized in Table 5.1 for the first type of samples and in Table 5.2 for the second one.

In the same manner as the training data in a supervised learning task, each test sample consists of a design matrix $\mathbf{X}$ and a corresponding label vector $\mathbf{y}$ in which each element $y^{(i)}$ indicates the class affiliation of a data point and takes one of C

Table 5.2: Second type of test data employed for NPC parameter estimation (test data of Type 2).

Cluster	Adjustable parameter		
	Radial distance	Length	Number points
Cluster 1	towards g_1, r_{g_1}	l_{g_1}	n_{g_1}
Cluster 2	towards g_2, r_{g_2}	l_{g_2}	n_{g_2}
Noise	$\varnothing$	$\varnothing$	n_{noise}
Note: Sample dimensionality d is variable.			

discrete values (cf. Sec. 3.1). Here, C equals four for the first set (three clusters and noise), while $C = 3$ different classes (two clusters and noise group) are present in the second test data sample.

5.2 Visualization during Simulation

To evaluate the effect of a certain NPC-parameter value change, the system's behavior (acceleration, oscillation and fusion of the nucleons, changes in kinetic and potential energy), along with the clustering outcome, is studied. For low-dimensional data containing up to three features, it is convenient to observe the particles' motion in feature space as shown in Fig. 5.1. However, to study the dynamics of the system, it is more revealing to observe the particles' energy evolution during simulation. The temporal kinetic energy change of the whole system is revealed by its temperature change (see Fig. 4.6(b)). A more detailed insight into the process can be obtained by visualizing both, the potential and the kinetic energy of each nucleon as shown in Fig. 5.2. The main advantage of this representation is that it can reveal how a sample behaves, regardless of its dimensionality. In this way it can easily be observed how each particle is accelerated towards a potential minimum.

5.3 Parameters of Performance Evaluation

Different clustering evaluation metrics/indices/approaches exist. In the context of the vast number of existing machine learning methods, some of them aim to generally assess the quality of the built groups or the whole clustering outcome, while others focus on some particular task (an extensive comparative overview study on this topic can be found in [Arbelaitz 2013]). Generally, one can distinguish between two types of cluster validation criteria: internal and external. The first type is based on information intrinsic to the data, while the second one relies on available knowledge about the data [Manning 2008]. In the current work, where the quality of a clustering result is evaluated using synthetically generated test data, the label vector $\mathbf{y}$ is used as an evaluation benchmark (or gold standard) and is employed for external validation. The task here is to compare the gold standard partitioning

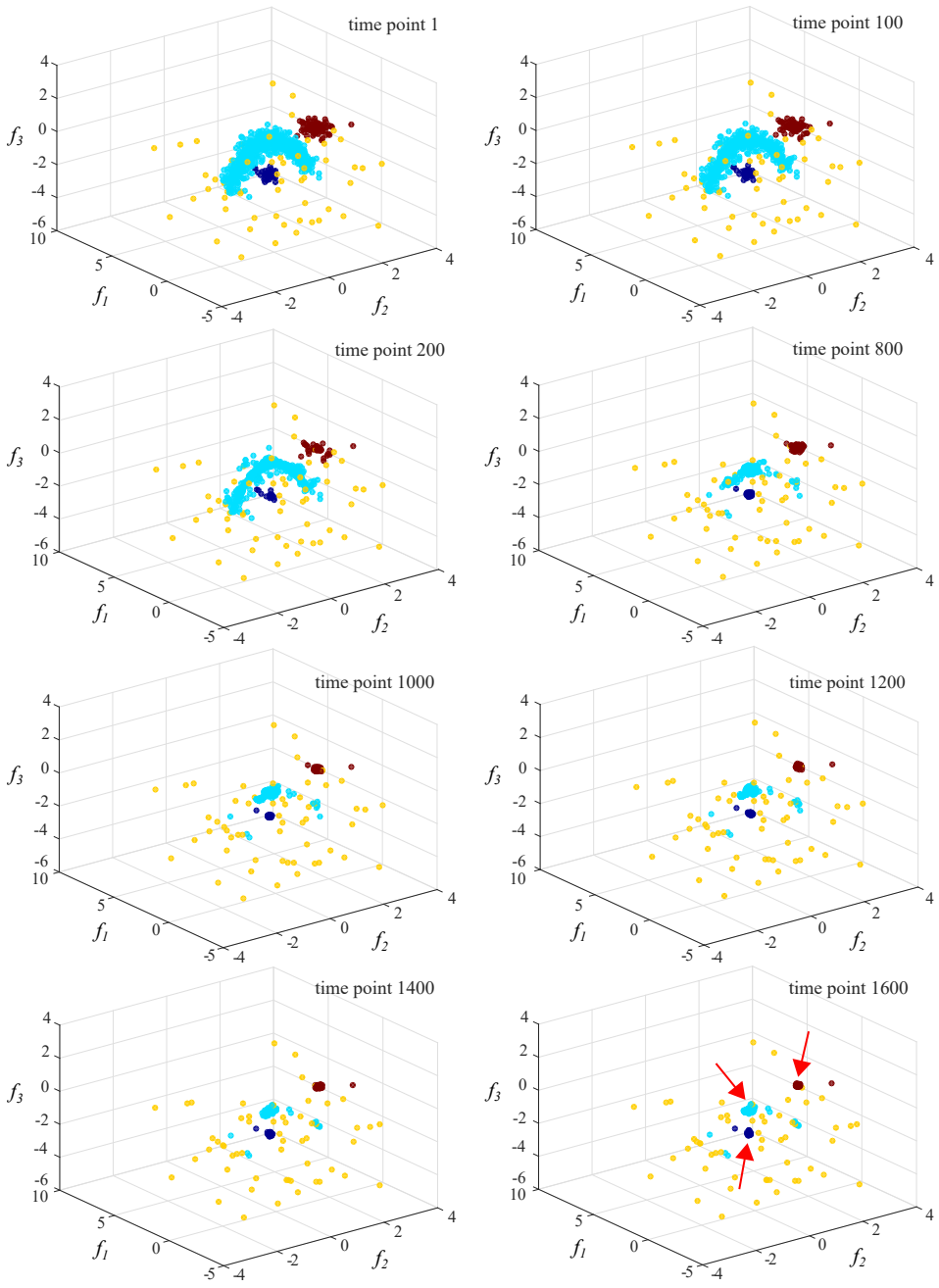


Figure 5.1: Demonstration of the temporal evolution of particles in feature space (spanned by f_1, f_2 and f_3) during simulation. Each panel represents a scatter plot of all particles from the test set, labeled in color using the known ground truth. In each case, the corresponding time step is displayed in the upper right of the panel. The last plot exhibits the final spacial distribution. The red arrows mark the three nuclei that are formed and recognized as coherent cluster.

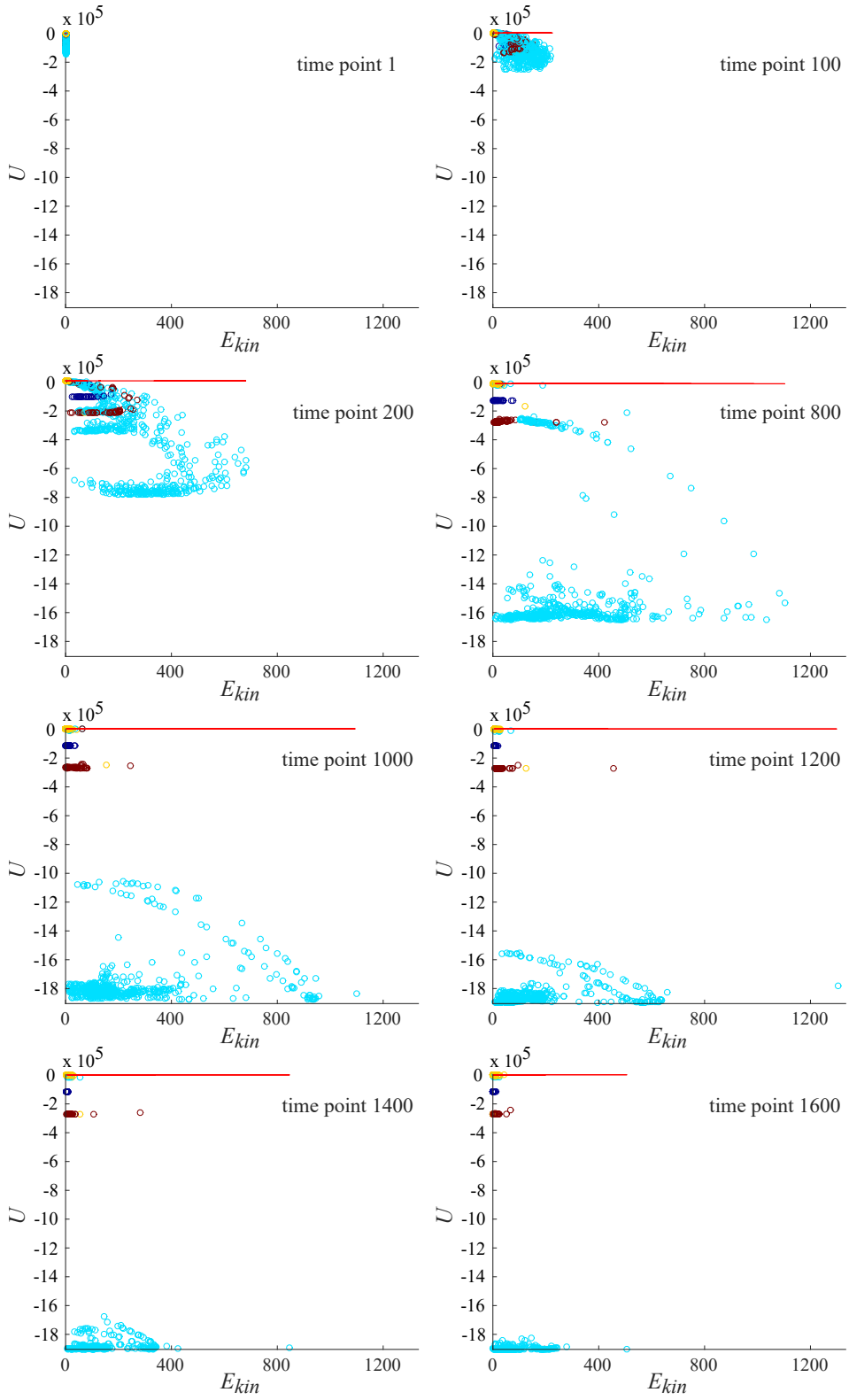


Figure 5.2: Temporal evolution of potential energy, U, and kinetic energy, E_{kin}, during simulation. Each panel represents a $U - E_{kin}$ scatter plot of all particles from the test set, labeled in color using the known ground truth. In each case, the corresponding time step is displayed in the upper right of the panel. The last plot exhibits the final energy distribution. The red line marks zero potential energy level, to which typically outliers stick during the whole procedure.

$\mathcal{C} = \{c_1, c_2, \cdots, c_C\}$ with the clustering $\mathcal{C}' = \{c'_1, c'_2, \cdots, c'_M\}$.

This can be achieved by constructing a contingency table (also confusion matrix, error matrix or matching matrix) with two dimensions: "reference" and "clustering outcome" [Powers 2011][1]. Such a table constitutes a $C \times M$ matrix with C rows corresponding to the groups in the gold standard partitioning and M columns corresponding to the detected clusters. The f_{ij} element of the matrix indicates the number of objects common to the class c_i and the cluster c'_j:

$$f_{ij} = |c_i \cap c'_j|, 1 \leq i \leq C, 1 \leq j \leq M \tag{5.1}$$

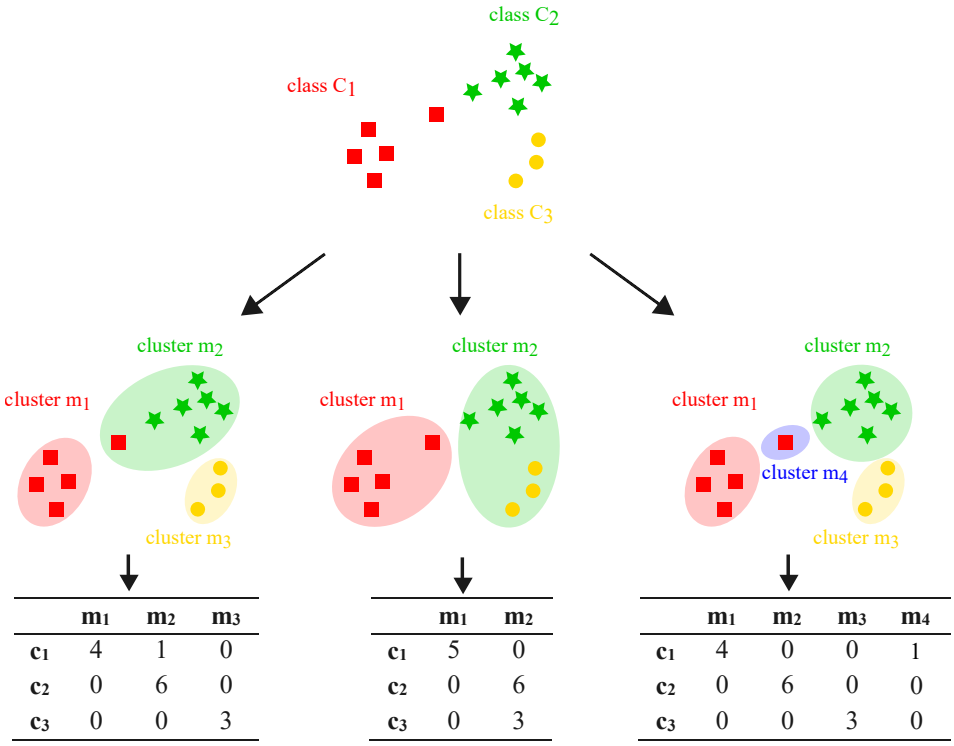
The sum over the elements of row c_i yields the total number of elements within this class. Analogously, the sum over the elements of column c'_j gives the number of elements of the j-th cluster. Table 5.3 represents the general form of such a contingency table, while Fig. 5.3 demonstrates three different clusterings of a toy problem matched to the benchmark solution.

Table 5.3: Contingency table for comparing two partitions.

		$\mathcal{C}'$				
	Class\Cluster	c'_1	c'_2	$\cdots$	c'_M	Sum
$\mathcal{C}$	c_1	f_{11}	f_{12}	$\cdots$	f_{1M}	$f_{1.}$
	c_2	f_{21}	f_{22}	$\cdots$	f_{2M}	$f_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	c_C	f_{C1}	f_{C2}	$\cdots$	f_{CM}	$f_{C.}$
	Sum	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.M}$	$f_{..} = N$

While a contingency table summarizes and visualizes the information on the overlap between two partitionings, a series of methods for external quantitative evaluation exist that measure the global quality of a solution in the form of a single real number. For the NPC parameters evaluation study where a large number of experiments is conducted, a simple and intuitive evaluation index is indeed required. Here, the Rand index, RI, is employed. It is based on counting the pairs of points assigned in the same way (or not) in both partitionings, $\mathcal{C}$ and $\mathcal{C}'$. The key idea is that a pair of objects should be associated with the same group if and only if they are similar with respect to the assessed features. The set of all $N(N-1)$ pairs of

[1] Since in [Powers 2011] a series of different cases and applications is discussed, the author refers generically to both dimensions as to "the model" and "the world". In the current thesis, these notations are changed to better fit in the remaining context.



	m1	m2	m3
c1	4	1	0
c2	0	6	0
c3	0	0	3

	m1	m2
c1	5	0
c2	0	6
c3	0	3

	m1	m2	m3	m4
c1	4	0	0	1
c2	0	6	0	0
c3	0	0	3	0

Figure 5.3: A toy problem demonstrating different contingency tables resulting from matching the ground truth solution (presented in the upper panel) with different clustering solutions (depicted in the second row).

observations is given by the disjoint union of the following sets:

$$\begin{aligned} S_{11} &= \{\text{pairs belonging to the same group within both } \mathcal{C} \text{ and } \mathcal{C}'\} \\ S_{00} &= \{\text{pairs belonging to different groups within both } \mathcal{C} \text{ and } \mathcal{C}'\} \\ S_{01} &= \{\text{pairs belonging to different groups within } \mathcal{C} \text{ but are in the same within } \mathcal{C}'\} \\ S_{10} &= \{\text{pairs belonging to the same group within } \mathcal{C} \text{ but in different ones within } \mathcal{C}'\} \end{aligned} \tag{5.2}$$

The corresponding sizes can be interpreted as true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$\begin{aligned} TP &= |S_{11}| \\ TN &= |S_{00}| \\ FP &= |S_{10}| \\ FN &= |S_{10}| \end{aligned} \tag{5.3}$$

The Rand index, RI, is then given by:

$$RI = \frac{TP + TN}{TP + TN + FN + TN} = \frac{TP + TN}{N(N-1)} \tag{5.4}$$

Its value lies in the range $[0, 1]$ and amounts to the fraction of correct decisions. It thus represents the accuracy of the clustering results. In the following, the index is also referred to as Rand accuracy and is expressed as a percentage.

5.4 Numerical Solution of the Newtonian Equation of Motion

The classical equation of motion Eq. 4.5 governs the dynamics of a given sample when analyzed using NPC. For its numerical integration different approaches can be applied. Here, four commonly used methods are considered: Euler, Runge-Kutta of second order, Runge-Kutta of fourth order and Runge-Kutta-Fehlberg. Their computational schemes are presented in detail in Appendix C. The goal of the conducted experiments is to explore the typical dynamic behavior and determine the most efficient procedure to solve the NPC problem for an arbitrary explored sample. The first three of the approaches evaluated employ a fixed time interval, Δ_t, at which the numerical approximation is calculated. Generally, in the limit $\Delta_t \to \infty$, all of the tested solvers lead to the same results. The main advantage of the higher order methods is that they allow for larger time steps without increasing the discretization error, thus being significantly faster. However, this is possible only at the expense of the space complexity of the algorithm, which in the specific case of NPC is $\mathcal{O}(pqNd)$, with q denoting the order of the ODE to be solved ($q = 2$ for Eq. 4.5), p the order of the ODE solver, N and d the number of points and the dimensionality of the data set to be processed, respectively. As it is revealed by the computational schemes in Appendix C, this is caused by the fact that higher order methods require additional intermediate function evaluations of the integrands within each time step.

In the following, first a discussion and a series of experiments considering the definition of a data specific time step are presented. Its estimation is based on the Euler method and is evaluated with test data of different size and dimensionality. Details about the test samples are given in Sec. 5.4.1. Next, in Sec. 5.4.2 the defined estimate for Δ_t is employed and evaluated for two further solvers of higher orders: Runge-Kutta of second order and Runge-Kutta of fourth order. The experiments with these two solvers are carried out further and a modified version of them is developed, which allows to solve the NPC problem more efficiently with respect to its time consumption.

5.4.1 Problem Specific Time Step Δ_t with Euler Method

Since it is clear that Δ_t should be problem specific as no *a priori* information about the points' spatial distribution and hence their resulting motion is available, the most intuitive choice may seem to be a solver employing an adaptive time step. Such methods dynamically adjust Δ_t by comparing different solutions and keeping the local truncation error of a single integration step within a desired level ε_{tol}.

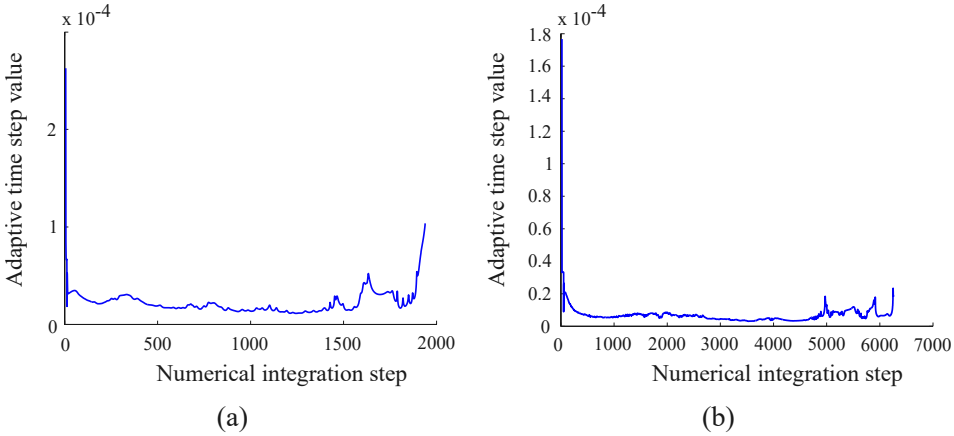


Figure 5.4: Adaptive time step values calculated during simulation using Runge-Kutta-Fehlberg for two test samples. (a) Test data of Type 1 generated with $n_{si} = 30$, $n_{so} = 40$, $n_h = 750$, $n_{noise} = 30$, $r_{si} = 5$, $n_{so} = 5$ and $r_h = 7$; error tolerance is set to $\varepsilon_{tol} = 10^{-4}$. (b) Five-dimensional skewed dataset, consisting of two clusters with $n_{g_1} = 40$, $n_{g_2} = 200$, $n_{noise} = 20$, $l_{g_1} = 5$, $l_{g_2} = 20$, $r_{g_1} = 5$ and $r_{g_2} = 10$; error tolerance is set to $\varepsilon_{tol} = 10^{-3}$.

There are different strategies of realizing this, e.g. performing a single integration step with Δ_t and then repeat the integration within the same interval but taking two single steps with $0.5\Delta_t$. Other adaptive step size algorithms are based on the outcome comparison of two solvers of different orders. Common example here is the Runge-Kutta-Fehlberg approach which is based on the combination of Runge-Kutta of fourth and fifth order. In either case, the calculation of the Δ_t correction term results in increased memory demands. In the concrete case of NPC, the space complexity of the above mentioned Runge-Kutta-Fehlberg method is $16qNd$. This is due to the fact that for the integration within a single time step, six intermediate evaluations of both integrands, velocity and position, are required (Eq. C.10), as well as two final steps in which their fourth and fifth order Runge-Kutta approximations are calculated as shown by Eq. C.11 and Eq. C.12, respectively.

Furthermore, each of the first six steps requires an update of the particles' forces and potentials. However, as previously stated, this update is the most time consuming part of the NPC algorithm, so that employing Runge-Kutta-Fehlberg method does not lead to an efficient solution of the NPC problem, neither with regard to memory consumption nor with regard to speed.

Still, taking a closer look at the values of Δ_t calculated typically during simulation using the Runge-Kutta-Fehlberg approach reveals a useful clue: it is possible to choose larger steps only in the beginning and at the end of the simulation, the rest of the time steps are in the range of some constant value. Two exemplary plots of this demeanor are displayed in Fig. 5.4 for two different test data sets. Obviously, applying an adaptive step method in the case of NPC does not lead to en efficient solution, since Δ_t cannot be largely varied and an adaptive time step strategy results solely in high memory consumption and computational cost.

Therefore, further investigations are conducted aiming at the determination of some simple heuristic to derive a fixed Δ_t sufficient to solve the equation of motion without adapting the time step. Based on what is known about a given sample and the acting forces, the following consideration is employed.

Given a system of N particles, the maximal possible acceleration $\mathbf{a}_{max}$ for a nucleon occurs in the case, when the remaining $N-1$ particles are located exactly at distance $r = R$. As the force is a vector sum, it is maximized when the $N-1$ nucleons are located at the same point in space, i.e. have already fused to a larger nucleus. Given that all particles' masses are normalized to unity, $\mathbf{a}_{max}$ at time point $t = t_0$ is given by:

$$\begin{aligned} \mathbf{a}_{max}(t_0) &= -\nabla\phi(\mathbf{r}, t) \\ &= \frac{V_0(N-1)e^{\frac{r-R}{a}}}{a(1+e^{\frac{r-R}{a}})} \end{aligned} \tag{5.5}$$

The absolute acceleration at $r = R$ is therefore given by:

$$|\mathbf{a}_{max}| \approx \frac{NV_0}{a} \tag{5.6}$$

Within a single Euler step, the acceleration is constant. Hence, the maximal particle's displacement during the first step of numerical integration is given by

$$\Delta r = \frac{1}{2}|\mathbf{a}_{max}|\Delta_t^2 \tag{5.7}$$

Note that no initial velocity term is included in Eq. 5.7, as the particles start at rest.

Next, the minimum mutual particle distance in the sample, d_{min}, with $d_{min} \neq 0$ is determined and employed to define a local scale specific for the analyzed sample, ζ. Restricting the maximal particle displacement to one unit of this local scale $\Delta r = \zeta$ fixes the time step to

$$\Delta_t = \sqrt{\frac{2\zeta}{|\mathbf{a}_{max}|}} \tag{5.8}$$

At time points $t > t_0$ the distances between intra-cluster particles start to decrease. Thus, the local scale ζ is chosen to be a fractional part of d_{min}, $\zeta = bd_{min}$ with $0 < b < 1$. The scale factor b is empirically determined as described in the following.

Estimate local scale factor b based on a fixed sample of Type 1 For a fixed dataset generated with $n_{si} = 36$, $r_{si} = 5$ and $n_{so} = 57$, $r_{so} = 8$, $n_h = 561$, $r_h = 8$, $n_{noise} = 79$, first, a reference solution is calculated with the Euler method by choosing $b = 0.0001$ which results in a sufficiently small time step. Next, the procedure is repeated for:

$$b = 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005. \tag{5.9}$$

All of them lead to identical results, when considering the final clustering outcome. However, taking a closer look at the system temperature evolutions presented in Fig. 5.5, reveals that only time steps calculated with $b \leq 0.001$ lead to results identical with the reference solution, in the sense of identical system behavior. In the rest of the cases, an additional oscillation of the system is observed. The fact that this effect becomes more evident as b and therefore Δ_t increase, indicates that this behavior is solely an artifact resulting from increased numerical error. It is partly compensated by the damping term introduced in the Newtonian equation of motion, which is the reason, why all of the tested values lead to the same clustering results.

Nevertheless, for the further investigations the value of $b = 0.001$ is chosen and tested in the next step on different, now randomly sampled datasets. This choice is motivated by the fact that the temperature evolution is employed as the basis for the NPC stopping condition (cf. Sec. 4.2.2). An additional oscillation may lead on the one side to a delay of reaching equilibrium to later time points thus worsening the computational efficiency of NPC even though a larger time step is employed. In the worst case, the algorithm would not reach the stopping condition at all. On the other hand, an artificial oscillation may cause the algorithm to break before reaching equilibrium if the temperature rapidly falls below $0.01T_{max}$ prematurely. Furthermore, taking undue large steps may cause some strongly accelerated intra-cluster nucleons to "jump out" of the cluster and be eventually falsely recognized as outlier. This effect can lead in the worst case to dissociation of formed nuclei. As in general no *a priori* knowledge about the systems studied is available, there is no possibility to foresee whether one of these effects would occur or not. Thus, a value for the local scale factor b is chosen which guarantees that the solution of the Newtonian equation of motion does not implicate oscillations based on a numerical error, but represents solely the motion resulting from the forces interaction between the particles.

Verifying the estimated local scale factor b for increased particles accelerations taking place in dense cluster(s) In a first test series, samples of Type 1 are generated, where the number of both clusters and the noise points is fixed to $n_{so} = 30$, $n_{si} = 40$ and $n_{noise} = 20$. The number of points within the "horseshoe" formation, n_h, is varied to examine whether the chosen b value is sufficient in the cases when the strong attractive force (and acceleration, respectively) on the nucleons inside of the group increases. Datasets with total number of points, N, set to

$$N = 250, 350, \ldots, 950 \tag{5.10}$$

are processed now only with $b = 0.001$ and $b = 0.0001$ and the system behaviors in both cases are found to be identical.

Verifying the estimated local scale factor b for samples of different dimensionality To explore effect of changing the dimensionality, in a second series, test

data of Type 2 are generated and processed in the same way (with $b = 0.001$ and $b = 0.0001$). Here, all samples are generated with $n_{g_1} = 80$, $n_{g_2} = 100$, $n_{noise} = 20$, $l_{g_1} = 5$, $l_{g_2} = 10$, $r_{g_1} = 3$ and $r_{g_2} = 7$. The dimensionality is varied as follows:

$$d = 5, 10, 20, 30, 50, 80, 100. \tag{5.11}$$

As in the previous test series, the processing with $b = 0.001$ and $b = 0.0001$ leads to identical system behavior.

5.4.2 Higher Order ODE Solvers with Fixed Time Step: Runge-Kutta of Second and Runge-Kutta of Fourth Order

Once a data specific time step for the Euler method, Δ_t, is defined, the algorithm is run with Runge-Kutta of second order with doubled time step $\Delta_t := 2\Delta_t$ and Runge-Kutta of forth order with $\Delta_t := 4\Delta_t$ for the same two series of test datasets employed for the estimation of the local scale factor b (Type 1 samples with varied number of points as stated by Eq. 5.9 and Type 2 data with varied dimensionality as stated by Eq. 5.11). Here, the results are found to be identical with the references calculated with $b = 0.0001$ and the Euler solver. As expected, compared to the reference Euler solution the processing time is reduced on average by 40% and 65%, respectively. However, the space consumption in the case of second-order Runge-Kutta amounts to $4qNd$, whereas the faster fourth-order Runge-Kutta method demands $8qNd$.

Since the force update is the most time-consuming part of the algorithm, some further tests are conducted. In this, a standard fourth-order Runge-Kutta solver is employed, where the nuclear force term is assumed to be constant within a single time step. The resulting computational scheme Eq. C.9 is shown in Appendix C.3, along with the standard fourth-order Runge-Kutta scheme. The background idea motivating this modification is the following: For an intra-cluster nucleon the force is invariably directed towards the nearest potential minimum, whereas for an outliers it has exactly the opposite direction. Choosing a data-specific time step as given by Eq. 5.8, however, allows to consider this minimum to be fixed in feature space and the resulting force to be constant (within the time interval Δ_t).

The experiments with this modified approach showed that it reproduces the same results as the rest of the considered solvers in less time. Here, the same two test data series employed previously in this section are processed. Apart from some small discrepancies, the systems behave in the same manner, regardless of which solver is employed. An exemplarily demonstration of this trend is displayed in Fig. 5.6(a) where the temperature curves of a sample (Type 2, $n_{g_1} = 80$, $n_{g_2} = 100$, $n_{noise} = 20$, $l_{g_1} = 5$, $l_{g_2} = 10$, $r_{g_1} = 3$ and $r_{g_2} = 7$, $d = 5$) processed with the different solvers are plotted. Furthermore, Fig. 5.6(b) shows the run times assessed for datasets of different sizes processed with all five discussed solvers. Here, again the second test data series, employed previously in the current section, is processed where N is varied as stated by Eq. 5.10. The results show that the modified fourth-order Runge-Kutta solver is especially efficient for larger datasets with high N, since the

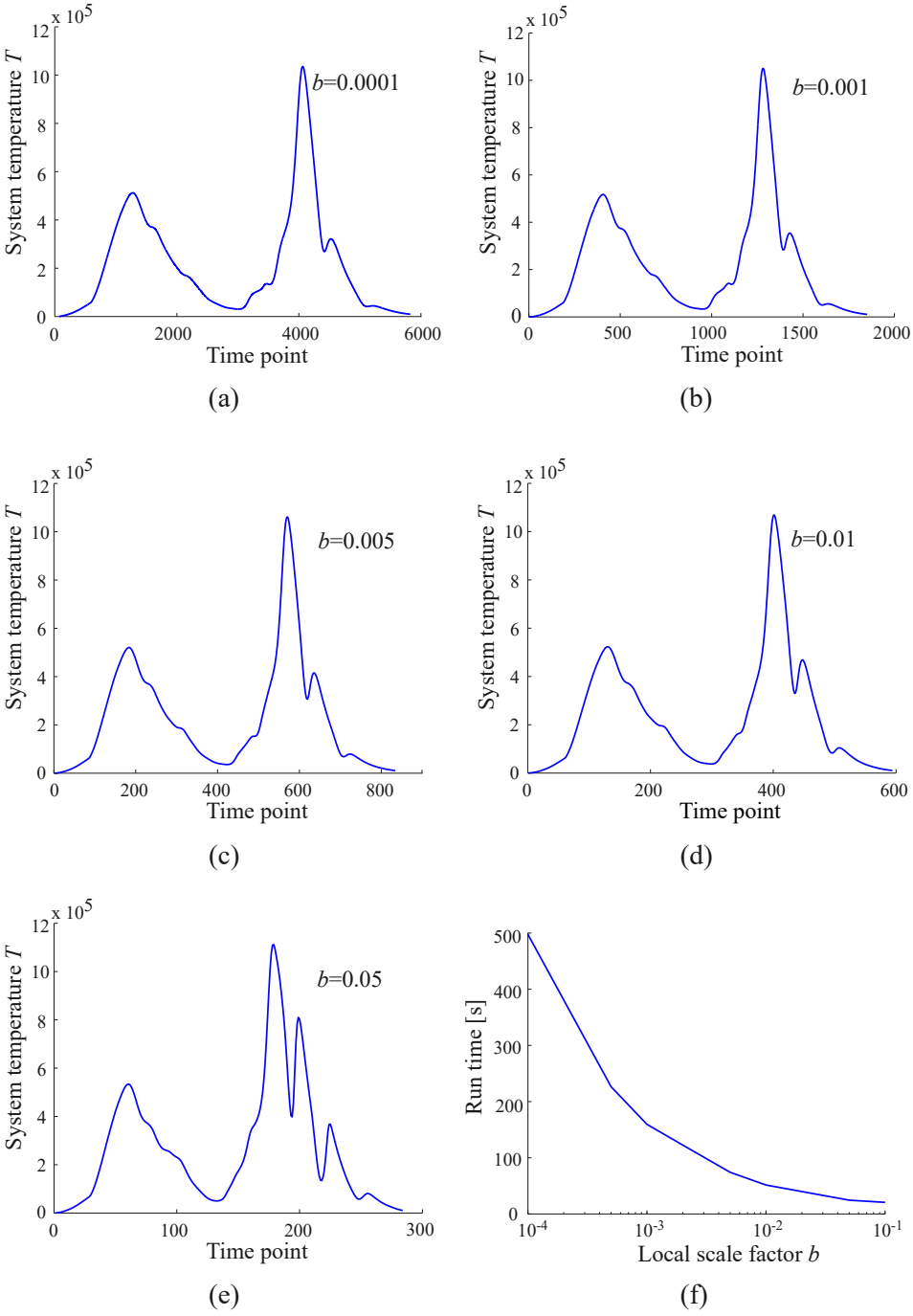
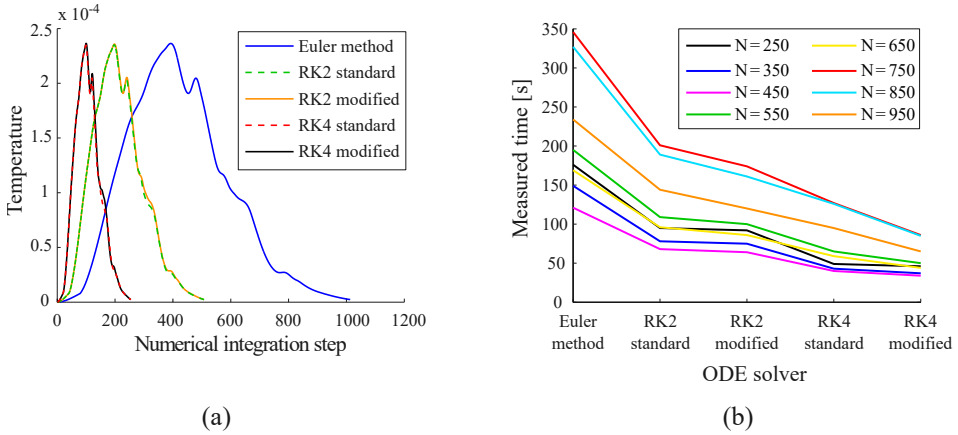


Figure 5.5: (a)-(e) Temperature evolution curves demonstrating the dynamic behavior of a system ($n_{si} = 36$, $r_{si} = 5$ and $n_{so} = 57$, $r_{so} = 8$, $n_h = 561$, $r_h = 8$, $n_{noise} = 79$) when the local scale factor b is varied. In each case the used b-value is displayed in the upper right of the panel. The reference temperature curve calculated with $b = 0.0001$ is displayed in (a). (f) Dependence of the measured runtime on the local scale factor.



(a) (b)

Figure 5.6: Comparison of the different ODE solvers tested. (a) Temperature curves revealing the temperature evolution of a test sample when processed with different solvers. (b) Run time for a series of test data sets of different size, where the number of data points N is varied.

calculation of the mutual particles' forces has time complexity of $\mathcal{O}(N^2)$. Thus, it is chosen to be the default NPC solver.

5.5 NPC Parameter Bounds

The initial choice of the parameters specifying the NPC performance is not based on a straightforward procedure, but is rather an integral part of the whole method development process. First of all, the parameters specifying the potential function, V_0, R, a and V_C in Eq. 4.1 and Eq. 4.2 are considered. All the additional components of the algorithm as described in the previous chapter, have then emerged based on the observed effects during the dynamic modeling of test data. For example, the idea of introducing a damping term η was born after observing that all of the modeled systems oscillate. By testing a range of different values for this parameter, it became evident, that $\eta > 0$ not only leads to a quicker transition to equilibrium, but also allows for the development of an objective stopping condition for the dynamic simulation. This, in turn, has led to the idea of using the temperature development of the system to define when the algorithm has to stop. By the time NPC was established in its final form, many different ideas were tested. The first decisions concerning reasonable parameter ranges are an inquiry resulting from this process of experimenting and deciding whether to keep working with an idea or discard it. The following subsections describe the experiments and observations that have led to the choice of proper ranges which are the basis for the further parameter estimation (Sec. 5.6).

5.5.1 Strong Interaction Potential Well Depth and Coulomb Barrier

The assignment of cluster members, as well as the decision whether a fluctuation in the particles density in feature space should be considered as a border between two touching clusters or not, is controlled by the parameters V_0, V_C and a. As stated in 4.2, V_0 is responsible for the mutual attraction of cluster members, V_C controls the strength of repelling nuclei and single protons, and a specifies the thickness of the transition zone delimiting a formed cluster. They are thus responsible for the detection of noise in marginal regions and for the splitting or the combination of particles located in local potential minima.

To achieve an optimal clustering result, a balance between the attractive strong interaction and the repulsive Coulomb term must be found. Clearly, the height of the Coulomb barrier V_C should be chosen to be greater than zero, to assure that non-cluster structures are repelled. If its value is set to zero, no repulsion would occur and adjacent clusters may easily be combined, especially in the presence of noise. On the other hand, its value must be considerably smaller than V_0, in order to allow for the formation of any nuclei. Thus, a V_0/V_C ratio suitable for the purposes of NPC has to be estimated. As the NPC parameters are relative values and presumably many different combinations of V_C and V_0 exist, here the experiments are started with a fixed value for the strong interaction magnitude, $V_0 = 3500$, and V_C is first varied within the boundaries $V_C \in [0, V_0]$. However, preliminary tests on a samples of Type 1, values $V_C > 100$ result in splitting of the elongated horseshoe-formed structure in the majority of the cases into two ot three separate groups. Thus the further investigations are restricted to $V_C \in [0, 100]$.

5.5.2 Transition Zone between Strong and Coulomb Interaction

The slope of the potential function at the border between strong and Coulomb interaction is controlled by the parameter a which is increased to a maximal value a_{max} during NPC simulation. As shown in Fig. 5.7, a value of $a = 0$ corresponds to a Woods-Saxon potential, ϕ^S, having the form of a box potential. Increasing the value leads to a decrease of the slope and a simultaneous decrease of the effective height of the Coulomb wall. Thus, the range of attractive interaction is increased and the strength of electrostatic repulsion is lowered. These effects result in a smoother transition zone between attractive and repulsive term.

Considering the clustering results, a narrow transition zone leads to a higher sensitivity to local density fluctuations. This is expected to be beneficial in the case of closely positioned clusters and detecting noise points on the rim of a cluster. However, only values $a > 0$ are considered, to avoid the discontinuity of the potential function as the Newtonian force is the spatial derivative of $\phi(\mathbf{r_i}, t)$ (see Eq. 4.4). Furthermore, an unsuitably low value may lead to the splitting of cohesive structures due to small intra-cluster density fluctuations. On the other hand, increasing a too much can result in the inability to detect any cluster as all points fuse to one single

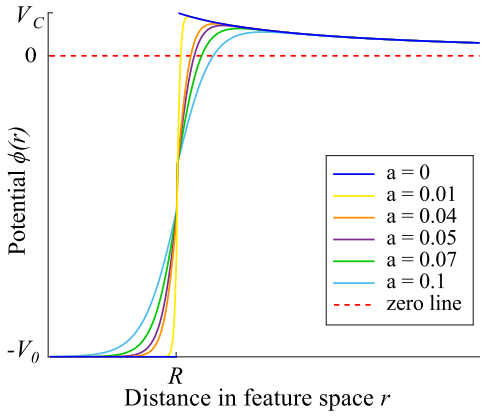


Figure 5.7: One-dimensional potential curves for different values of the parameter a as discussed in the current chapter. In the case of $a = 0$, the strong interaction term of the potential function, ϕ^s, has the form of a box potential (blue curve). Increasing the value of a lowers the slope at the border between the strong attractive and the weak repulsive term (ϕ^s and ϕ^C, respectively), as well as the effective height of the Coulomb barrier specified by the parameter V_C.

nucleus. As can be seen in Fig. 5.7, when compared to $a = 0$, a value of $a = 0.1$ leads to an increase of the attractive range by approx. 50% and a drop of the Coulomb barrier by almost the same factor. For these reasons, the further estimation of the values a and a_{max} is restricted to the interval $[0.01, 0.1]$, $a < a_{max}$.

5.5.3 Strong Interaction Range

Along with V_0, V_C and a, the range of strong interaction, R, is a parameter that governs the mutual attraction of points and thus controls the formation of nuclei. Clearly, it has to be a data specific measure, representing the mutual distances of adjacent particles inside of the cluster(s). To determine a suitable estimate of R, the distribution of intra-cluster distances in three series of test data sets with different characteristics are studied.

As described in detail in the following, the first set is used to study effects based on changes of the geometrical form of a cluster, especially when it differs from the case of a compact spheroid. Here, elongated cylindrical formations are generated. The second experimental series explores effects based on changes of the particles' density within a cluster. In the last series, the impact of the samples' dimensionality is considered.

Effects related to changes in the form of the cluster(s) In this series, three dimensional test data sets of Type 2 are generated with $d = 3$. In all samples, the noise level is set to $n_{noise} = 50$ and one of the clusters is chosen to be a compact group with $r_{g_1} = 5$, $l_{g_1} = 5$ and $n_{g_1} = 50$. The second group has the same width, $r_{g_2} = 5$, but is gradually elongated by varying its length l_{g_2} between 10 and 510

with

$$l_{g_2} = 10, 35, 60 \ldots, 510 \tag{5.12}$$

To explore solely the effects of varying cluster forms, the average intra-cluster particles density (clusters do not have sharp borders) is kept approximately constant by increasing the number of points n_{g_2} proportional to the length, l_{g_2}. In this, it is assumed that the volume occupied by the elongated cluster in feature space can be approximated by the volume of a cylinder:

$$V_{g_2} = \pi r_{g_2}^2 l_{g_2} \tag{5.13}$$

and the density is given by:

$$\rho_{g_2} = \frac{n_{g_2}}{V_{g_2}} \tag{5.14}$$

Three such samples are exemplary demonstrated in Fig. 5.8 along with the normalized histograms of the mutual particles' distances within each group of the sample. As it can be seen, the distance distribution of the compact unchanged group shows a single narrow peak. In comparison, increasing the length of the second cluster, l_{g_2}, leads to a continuous shift of the mutual distances towards higher values. The effect increases with the length of the cluster. In all generated samples, however, the maxima of the intra-cluster distances distributions of both clusters lie closely together and represent the mutual distances of the close neighbors. When compared to the cumulative distribution functions of the mutual distances within the whole test sample (red curves in Fig. 5.8), it becomes evident that the two intra-cluster distances peaks correspond to values near the 0.1-quantile (marked with a red circle in Fig. 5.8).

A similar behavior is observed in the case of clusters whose form differs from that of a straight line, such as the one presented in the last row of Fig. 5.8. Here, a test sample of Type 1 is shown that contains only one spheroid (green) enclosed by the horseshoe-formed group. It is generated with $r_h = 5$, $n_h = 350$, $r_{si} = 5$, $n_{si} = 50$, $r_{so} = 0$, $n_{so} = 0$ and $n_{noise} = 50$. The length of the (blue) bent structure is approximately 60 and is thus comparable to l_{g_2} of the third sample of Type 2 presented in Fig. 5.8. As it can be seen, due to the bent form of the elongated cluster in this case, the shift towards higher distances is enhanced. Thus, the maxima of the mutual distances of particles belonging to both clusters (blue and green) lie in the interval between the 0.1-quantile and the 0.2-quantile (marked with a red circle and a red square in the lower right plot of Fig. 5.8, respectively).

Effects related to changes in the density of the cluster(s) To explore effects related to density changes of a cluster, three types of experiments are conducted. First, the same setting as in the previous experimental series is employed (test samples of Type 2), however, with keeping the number of points in the elongated cluster constant to $n_{g_2} = 50$ while varying its length as stated by Eq. 5.12. As the particles' density is inversely proportional to l_{g_2}, it drops as l_{g_2} is increased.

In the second case, test data of Type 1 is generated with the same parameters

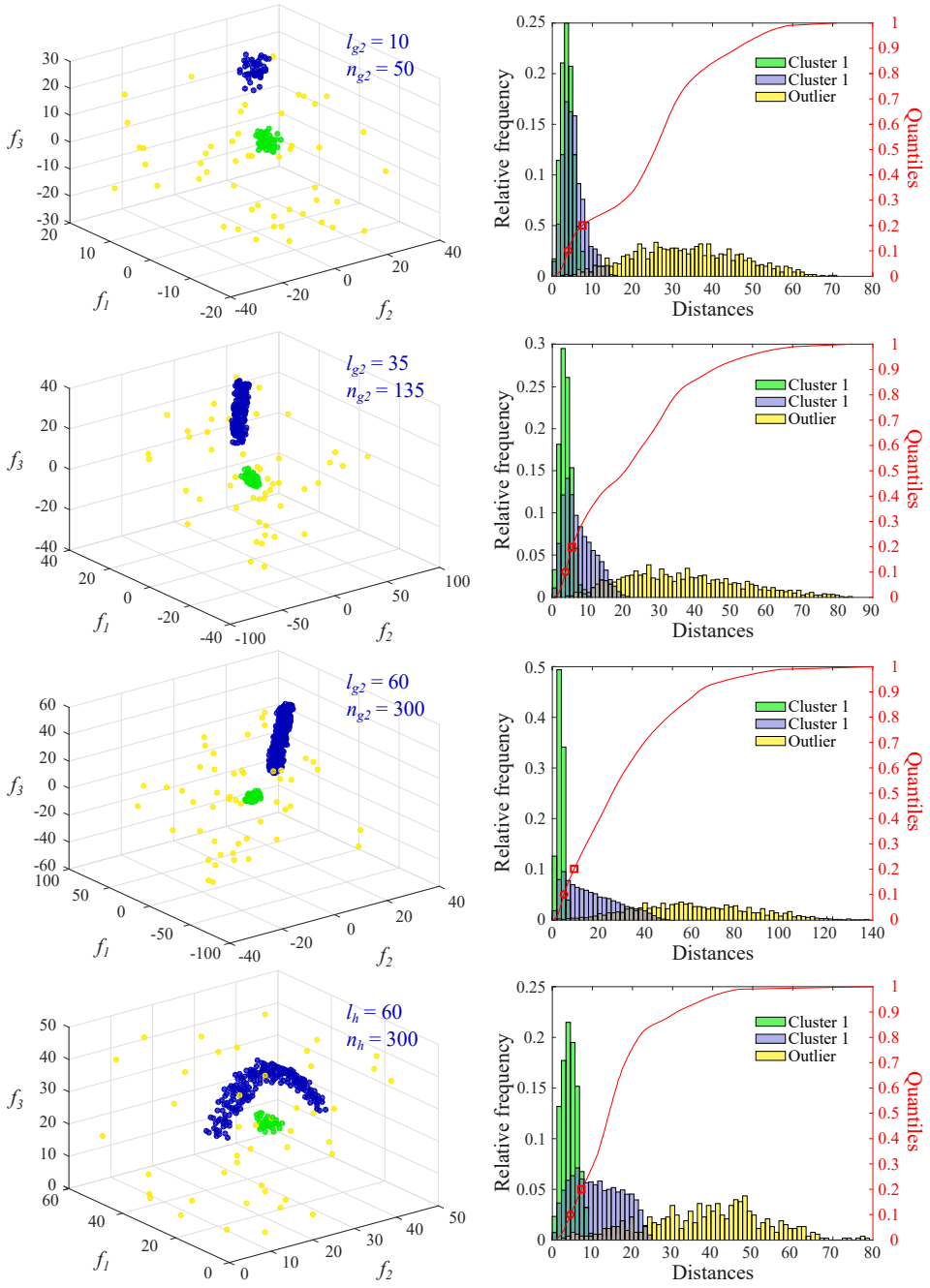


Figure 5.8: Test data employed to explore the effects of changing the clusters' form while keeping the density approximately constant as described in Sec. 5.5.3. The left panels show scatter plots of the samples. The first three present data of Type 2 where the length of one of the clusters (blue) is varied. In each case, the corresponding length and number of points are displayed in the upper right of the panel. For a comparison, the density and form of the compact cluster (green) are kept unchanged. The fourth sample is of Type 1. The elongated cluster here has the same length and density as the third sample of Type 2 presented. The right panels show the normalized histograms of the mutual particles' distances within each group of the samples, where noise points (yellow) are regarded as one group. In each case, the histograms are overlaid by the cumulative distribution function of the mutual points' distances within the whole test data (red curves). The red circles and squares denote the 0.1-quantile and the 0.2-quantiles in each case.

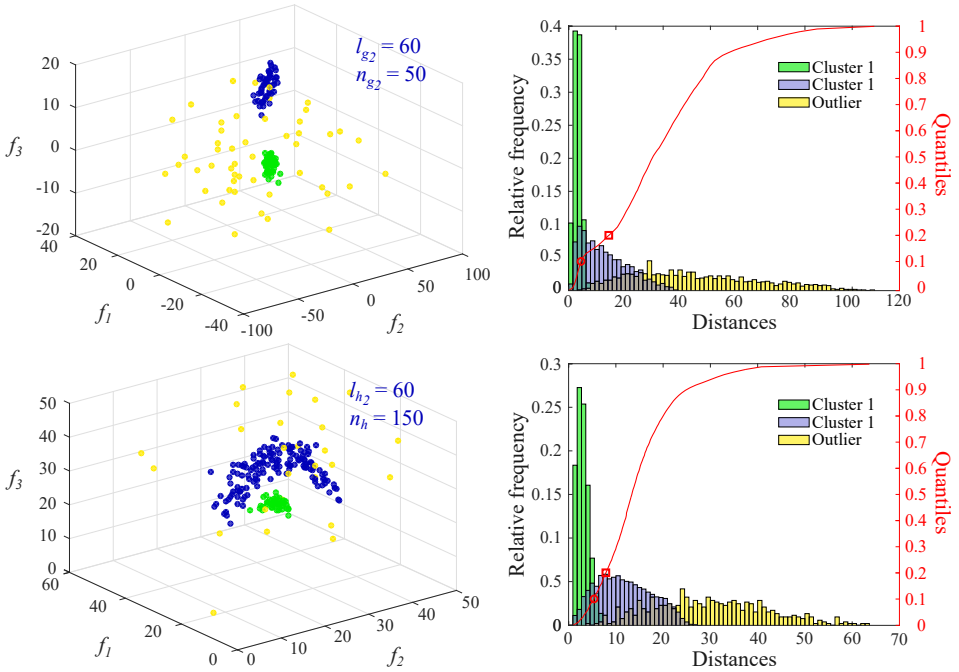


Figure 5.9: Two test samples, generated with the same parameters, as the third and the fourth test data presented in Fig. 5.8, except for the number of points in the elongated blue clusters. The left panels show scatter plots of the samples. In each case, the corresponding values for l_{g_2}, l_h, n_{g_2} and n_h are displayed in the upper right of the panel. The right panels show the normalized histograms of the mutual particles' distances within each group of the sample, where noise points are regarded as one group. The histograms are overlaid by the cumulative distribution function of the mutual distances within the whole test sample, which is represented by the red curves in each graph. The red circles and squares denote the 0.1-quantiles and the 0.2-quantile in each case.

as the fourth dataset displayed in Fig. 5.8. However, the number of points within the elongated horseshoe-formed cluster is gradually reduced as follows:

$$n_h = 600, 300, 150, 75 \tag{5.15}$$

Figure 5.9 shows two samples generated with the same specifications as the last two test data depicted in Fig. 5.8, except for the number of points in the elongated clusters: $n_{g_2} = 50$ in the first case and $n_h = 150$ in the second one. When considering test data from Type 2, the effects are similar to those observed in the first series. Increasing l_{g_2} leads to shift of the intra-cluster distances to higher values, but the maxima of the distributions of both dense groups lie near values corresponding to the 0.1-quantile. In the case of samples of Type 1, however, a more prominent shift of the "horseshoe" distances towards higher values occurs and the distribution maximum reaches values near the 0.2-quantile.

In a third experiment series, three dimensional test data sets of Type 2 are generated (dimensionality set to $d = 3$). In all samples, the noise level is set to

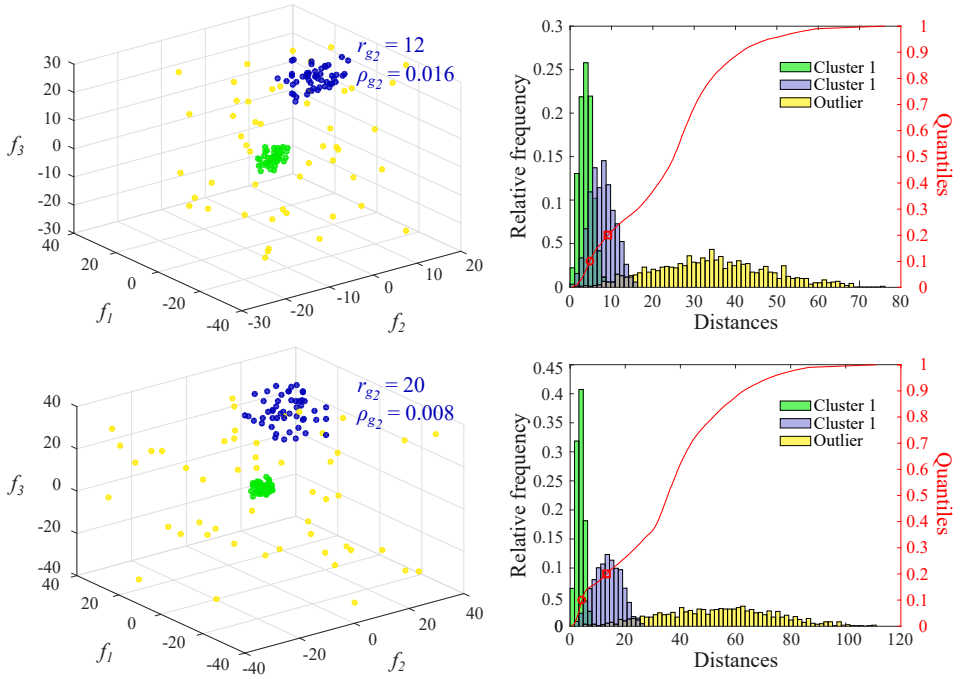


Figure 5.10: Three-dimensional test samples of Type 2 demonstrating the effects of a variable radius, r_{g_2}. The left panels show scatter plots of the samples. In each case, the corresponding values for r_{g_2} and the resulting particles density ρ_{g_2} are displayed in the upper right of the panel. The right panels show the normalized histograms of the mutual particles' distances within each group of the sample, where noise points are regarded as one group. The histograms are overlaid by the cumulative distribution function of the mutual distances within the whole test sample, which is represented by the red curves in each graph. The red circles and squares denote the 0.1-quantiles and the 0.2-quantile in each case.

$n_{noise} = 50$ and the first cluster is chosen to be a compact spheroid with $r_{g_1} = 5$, $l_{g_1} = 5$ and $n_{g_1} = 50$. The second cluster is generated with the same parameters as the first one, whereby the radius r_{g_2} is varied as follows:

$$r_{g_2} = 10, 12, \ldots, 100 \tag{5.16}$$

As previously stated, the volume of the formations are approximated by the volume of a cylinder (Eq. 5.13). Thus, the average particles density ρ_{g_2} within the changed cluster (Eq. 5.14) is inversely proportional to $r_{g_2}^2$.

Increasing r_{g_2} leads to shift of the intra-cluster distances within the cluster changed. In Fig. 5.10, two of the samples are exemplarily shown in which the resulting average particle density differs by a factor of 50%. As it can be seen, the compact groups (displayed in green) have maximum intra-cluster distances around the 0.1-quantile. In comparison, the maximum of the dilated groups is shifted and corresponds to a value near the 0.2-quantile in all cases examined.

Note that if the intra-cluster densities of both cohesive groups are similar (such as in the test data displayed in Fig. 5.8), the cumulative distribution function in the

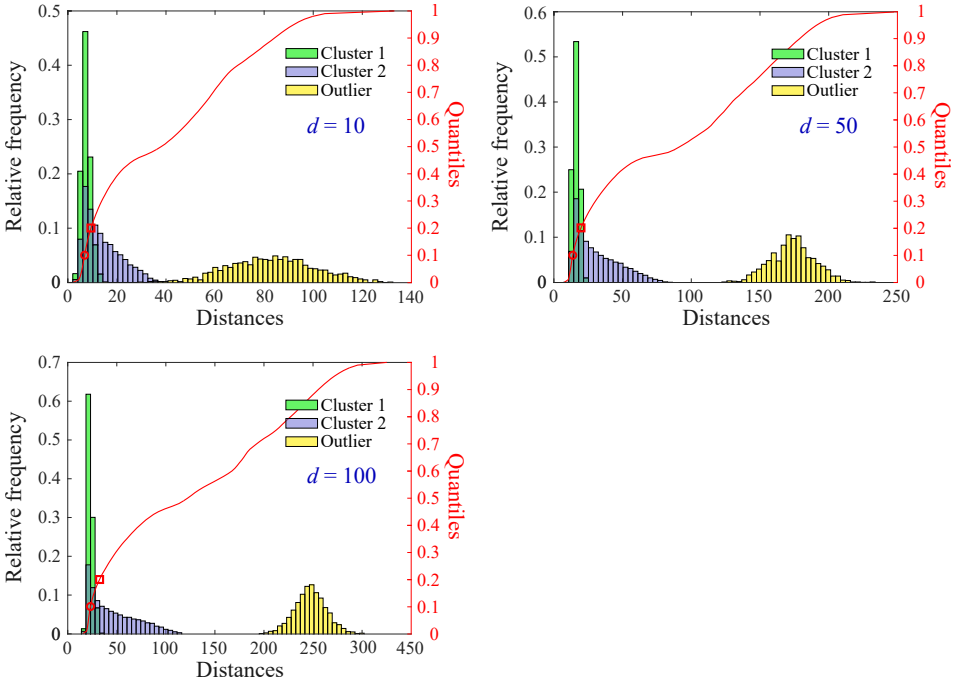


Figure 5.11: Three test samples of Type 2 generated with different dimensionality d. Each panel shows the normalized histograms of the mutual particles' distances within each group of the sample, where noise points are regarded as one group. These histograms are overlaid by the cumulative distribution function of the mutual distances within the whole test sample, which is represented by the red curves in each graph. The red circles and squares denote the 0.1-quantiles and the 0.2-quantile in each case. In each case the corresponding sample dimensionality d is displayed in blue color under the legend of the plot.

interval between the 0.1-quantile and the 0.2-quantile has a steep increase. Thus, the distance values corresponding to the 0.1-quantile and the 0.2-quantile, respectively, are similar, too. Increasing the difference between the clusters' densities, leads to a drift of the distance values corresponding to the 0.1-quantile and the 0.2-quantile.

Effects related to changes in the sample dimensionality To explore effects related to changes in the sample dimensionality d, a series of test data of Type 2 is generated with $r_{g_1} = 5$, $l_{g_1} = 5$, $n_{g_1} = 50$, $r_{g_2} = 5$, $l_{g_2} = 20$, $n_{g_1} = 200$ and $n_{noise} = 50$. The dimensionality d is varied as follows:

$$d = 10, 50, 100 \tag{5.17}$$

Figure 5.11 displays the resulting normalized histograms of each group within the test sets again, overlaid by the cumulative distribution functions of the mutual distances within the whole samples. As expected when employing the Euclidean

distance, the inclusion of additional dimensions leads to a shift of all distances towards higher values. This effect is most prominent when considering the noise points. Nevertheless, the distributions of the intra-cluster distances in all examples exhibit the same form as in the previous tests with data of Type 2: the distances of the compact cluster ($r_{g_1} = 5, l_{g_1} = 5$) build a single narrow peak, while the elongated cluster (r_{g_2}, l_{g_2}) has an asymmetric distribution with a shift towards higher distance values. However, both distribution maxima lie again near values corresponding to the 0.1-quantile of the cumulative distance function. Note that this holds true, even though the density drops rapidly when increasing d since the volume of a d-cylinder, V_d, is proportional to the d-th power of its radius, $V_d \propto r^d$.

Based on the observed effects, R is chosen to be the q_R-quantile of the distance distribution of a sample. The search for the optimal values of q_R and for the modulation scaling factor, s_R, is carried out in the intervals $q_R \in [0.08, 0.2]$ and $s_R \in [1.1, 2]$, respectively.

5.5.4 Damping Constant

The initial experiments showed that the above discussed values and ranges of the remaining NPC parameters cannot be combined with values for the damping constant less than 70. In cases of $\eta < 70$, all systems studied showed an ongoing oscillation (best revealed by the temporal evolution of the potential and kinetic energies of the particles as demonstrated in Fig. 5.2), so that no equilibrium is reached at all and the NPC breaking condition is therefore not reached. On the other hand, an undue high η value prevents the particles from being sufficiently accelerated at all and forces them towards local minima, thus leading to the fragmentation of cohesive groups. In the cases of $\eta > 500$ the points move extremely slow even towards the local minima. Therefore the bounds for the damping turn constant are chosen to be $\eta \in [70, 500]$.

5.5.5 Parameter Bounds Summary

In the following table, all considerations concerning the parameter bounds employed in the further optimization study are summarized:

Table 5.4: NPC parameter bounds.

	Parameter						
	V_0	V_C	a	a_{max}	q_R	s_R	η
Lower bound	3500	1	0.01	0.01	0.08	1.1	70
Upper bound	3500	100	0.1	0.1	0.2	2	500

5.6 Estimation of Default Parameter Values

5.6.1 Optimization Procedure Setup

For the purposes of parameter optimization, a cost function, J, is defined as the error rate (fraction of incorrectly clustered points), $1 - RI$, averaged over P independent test data sets:

$$J = \frac{1}{P} \sum_{i=1}^{P} (1 - RI_i) \tag{5.18}$$

The parameter optimization is based on two different test data series:

- A series of $P = 25$ independent samples of Type 1, with $r_h = 5$, $n_h = 350$, $r_{si} = 5$, $n_{si} = 50$, $r_{so} = 7$, $n_{so} = 40$ and $n_{noise} = 50$. These data are employed to evaluate effects on clusters of irregular and intertwined shape as one cluster is partly enclosed by an elongated curvilinear group. For the calculation of a single cost with a certain set of parameters, 25 new samples are randomly generated. The average error rate, denoted in this case as J_1, is accordingly calculated as $J_1 = \frac{1}{25} \sum_{i=1}^{P} (1 - RI_i)$. An example of a test sample and the corresponding cost, J_1, calculated with varied q_R, V_C, s_R and fixed $a_{max} = 0.055$, $\eta = 200$ and $K = 1$, are shown in Fig. 5.12(a) and Fig. 5.12(b), respectively.

- A set of $P = 5$ samples of Type 2, with $d = 3$, $r_{g_1} = 5$, $l_{g_1} = 30$, $n_{g_1} = 100$, $r_{g_2} = 7$, $l_{g_2} = 35$, $n_{g_1} = 150$ and $n_{noise} = 70$. Here, special attention is paid to the distance between the clusters. In particular, only test data are chosen in which the two cohesive groups are positioned very closely in order to evaluate effects on touching clusters. In contrast to the fist test series of 25 samples, the cost here is calculated based on the same set of five samples, $J_2 = \frac{1}{5} \sum_{i=1}^{P} (1 - RI_i)$ [2]. An example of a test sample and the corresponding cost, J_2, calculated with varied q_R, V_C, s_R and fixed $a_{max} = 0.055$, $\eta = 200$ and $K = 1$, are shown in Fig. 5.12(c) and Fig. 5.12(d), respectively.

For a given parameter set, the error rates, J_1 and J_2, are first calculated. This allows to estimate if there are regions in parameter space which are particularly beneficial (or not) for different cluster specifications which are characterized by both types of test data. To equally weight these effects, the total cost is then estimated as the average of J_1 and J_2:

$$J = \frac{1}{2}(J_1 + J_2). \tag{5.19}$$

For the experiments performed here, only three-dimensional test data are employed, since they allow a detailed visual inspection of the results in feature space, in addition to the information revealed by the contingency tables. In this way, details

[2] Note, that no averaging $P = 5$ is needed for the actual minima search. However, to keep the calculated cost comparable, i.e. $J \in [0 \;\; 1]$, the J_2 calculation is kept consistent with Eq. 5.18

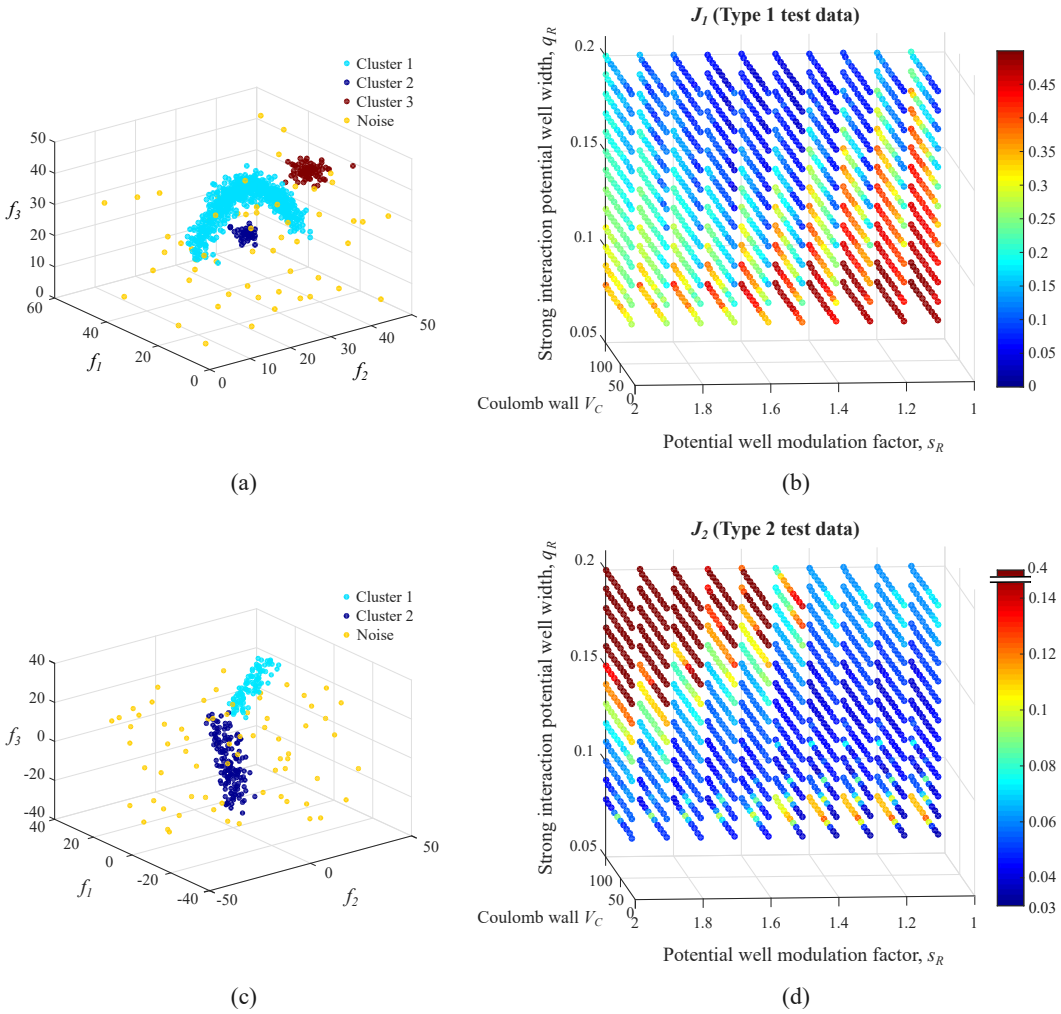


Figure 5.12: Demonstration of the intermediate steps when calculating the cost J as the average of the terms J_1 and J_2. (a) An example of a test sample of Type 1 employed for the calculation of J_1. (b) J_1 as a function of s_R, V_C and q_R. (c) An example of a test sample of Type 2 employed for the calculation of J_2. (b) J_2 as a function of s_R, V_C and q_R. In the demonstrated example the remaining NPC parameters are fixed to $a = 0.05$, $\eta = 200$, $K = 1$ and $a_{max} = 0.055$.

about the effects occurring can easily be evaluated, especially if misclassifications of a certain data fraction occur.

The first attempts to find a parameter value combination suitable for the clustering task are conducted with the interior point algorithm [Waltz 2006, Byrd 2000, Byrd 1999]. These experiments, however, have led to no satisfactory results since the observed minima were extremely sensitive to the chosen initial starting points. Table 5.5 shows exemplarily two results obtained with different initial points.

Table 5.5: Demonstration of two attempts to optimize the NPC parameters by means of the interior point algorithm where starting the optimization with different initial points in parameter space leads to different results.

	Parameter						
	V_0	V_C	a	a_{max}	q_R	s_R	η
Initial point	3500	1	0.04	0.07	0.08	1.5	200
Estimated optimum	3500	4.599	0.051	0.055	0.125	1.541	200.021
Initial point	3500	100	0.04	0.07	0.08	1.5	200
Estimated optimum	3500	40.010	0.042	0.070	0.185	1.500	200.142

These preliminary results suggest that the cost has either many local minima or a wide potential minimum valley in parameter space, or both. In the given optimization setup, both cases may exist. Thus, an extensive search is conducted as described in the following. It allows to not only estimate an optimal parameter set, but also study closely the areas of increased cost and the particular effects that changing the parameters' values have on the clustering results.

In the course of estimation of suitable parameters' bounds, it became evident that V_C, q_R, s_R, and a_{max} have the most significant influence on the NPC performance. In contrast, a affects the initial "sharpness" of cluster borders and thus influences solely the assignment of particles located near the edge of a cluster (this, however holds true solely if a is kept in the estimated bounds, otherwise it can significantly impair the NPC accuracy). Another parameter that controls the correct clustering of marginal nucleons is the number of iterations, K. The damping constant η influences primarily the run time of the algorithm (here, too, if kept within the estimated bounds, otherwise it can significantly reduce the NPC accuracy, as well).

Thus, the further investigation is divided into three steps. First, the parameters V_C, q_R, s_R, and a_{max} controlling the global NPC performance are studied, as described in the following Sec. 5.6.2. Next, a and η are considered in Sec. 5.6.3, followed by a discussion on the number of iterations, K, in Sec. 5.6.4.

5.6.2 Estimation of Optimal V_C, q_R, s_R, and a_{max}

In this first step, the parameters V_C, s_R, q_R and a_{max} are varied within the estimated bounds, while keeping the rest fixed to:

- $a = 0.05$, a value in the middle of the investigated range (see Table 5.4);
- $\eta = 200$, a value that is known to lead to a sufficient damping allowing the breaking condition to be met, but not suppressing the nucleons' motion excessively so as to stick them to local minima;
- $K = 1$, to optimize the main NPC routine.

To allow for a visual examination, q_R, s_R and V_C are varied and the resulting cost, J, is displayed as a scatter plot. This procedure is repeated for different a_{max} values. Since the cost calculation is a highly time consuming task, first the following discrete values are employed:

$$\begin{aligned} q_R &= 0.08, 0.09, \ldots, 0.2 \\ s_R &= 1.1, 1.2, \ldots, 2 \\ V_C &= 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 \\ a_{max} &= 0.02, 0.04, 0.06, 0.08 \end{aligned} \tag{5.20}$$

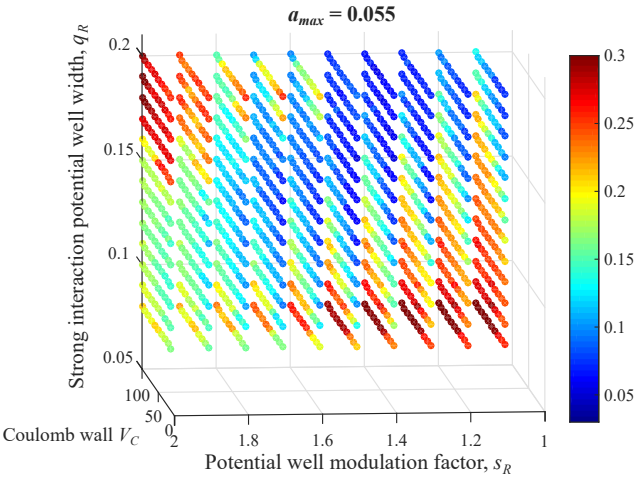
The minimal cost within the entire ranges of q_R, s_R and V_C tested, is observed for $a_{max} = 0.06$. In a next step, the cost is calculated for $a_{max} = 0.055$ and $a_{max} = 0.065$, in addition to $a_{max} = 0.06$. The resulting s_R-V_C-q_R scatter plots are shown in Fig. 5.13. Comparing the results obtained for $a_{max} = 0.065$ (Fig. 5.13(c)) and $a_{max} = 0.06$ (Fig. 5.13(b)), it becomes obvious that values $a_{max} > 0.06$ lead to an increase of the error rate for all values of q_R, s_R and V_C investigated.

Note that due to the random nature of the test data generated, in all samples some noise points are located inside a cluster so that the calculated average minimum error rate is systematically slightly greater than zero with $J \approx 0.03$.

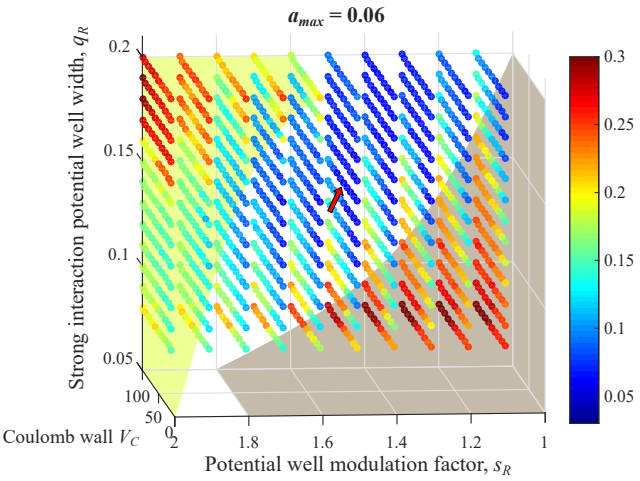
The results obtained with $a_{max} = 0.055$ and $a_{max} = 0.06$ exhibit similar values. To compare them, the difference $J(a_{max} = 0.055) - J(a_{max} = 0.06)$ is calculated for all examined q_R, s_R and V_C, as displayed in Fig. 5.14(a). In addition, the distribution of the difference values is examined (Fig. 5.14(b)). It shows clearly that the cost in the case of $J(a_{max} = 0.055)$ is higher on average than in the case of $J(a_{max} = 0.06)$. Thus, the default value for this parameter is chosen to be $a_{max} = 0.06$.

When considering the scaling factor s_R, there is a distinct minimum valley of the cost at $s_R = 1.5$. In this region of s_R-V_C-q_R space, different combinations of V_C and q_R lead to satisfactory results. In the range $q_R \in [0.15, 0.18]$ the average error rate is $J \approx 0.03$ for all V_C values tested, except for $V_C \geq 90$ where the cost increases to $J \approx 0.10$. As previously discussed, however, a lower strong interaction range value is expected to increase the NPC sensitivity in the cases of closely positioned clusters. Thus, $q_R = 0.15$ is chosen which is the lowest value for which all V_C values tested lead to minimal cost, $J \approx 0.03$.

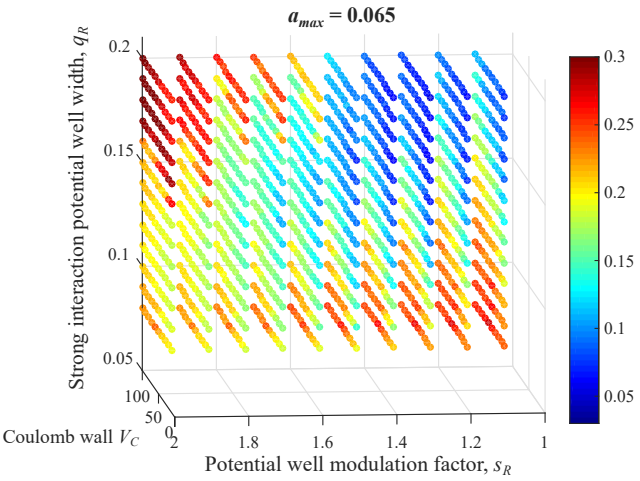
Next, the default height of the Coulomb wall is set to a value in the middle of the range tested, $V_C = 50$. This aims at an optimal balance between the ability to detect correctly cluster boundaries and prevent the splitting of large and/or elongated structures due to the cumulative repelling Coulomb force (see Sec. 5.5.1).



(a)



(b)



(c)

Figure 5.13: Cost J calculated as a function of s_R, V_C and q_R for three different values of a_{max}. (a) Results for $a_{max} = 0.055$. (b) Results for $a_{max} = 0.06$. The two regions of increased cost marked in light green and gray result from different effects, as explained in the current section. The chosen parameter optimum is marked with a red arrow. (c) Results for $a_{max} = 0.065$.

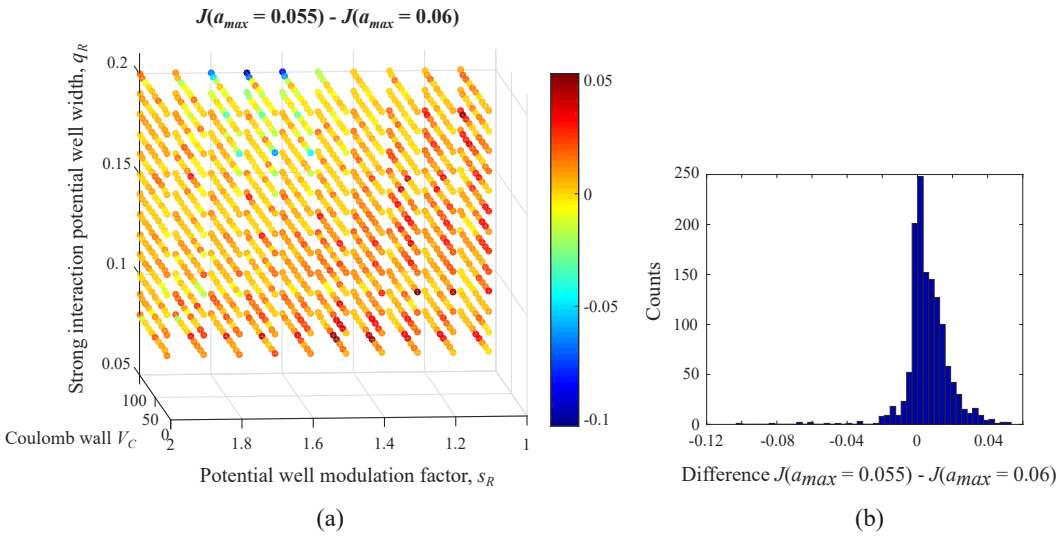


Figure 5.14: Exploring the difference between the cost values calculated with $a_{max} = 0.055$ and $a_{max} = 0.06$ (both cases are displayed in the first and the second panel of Fig. 5.13, respectively). (a) Three-dimensional distribution of the difference $J(a_{max} = 0.055) - J(a_{max} = 0.06)$. (b) Histogram of the calculated difference values.

The optimum chosen in this first step is marked with a red arrow in Fig. 5.13(b).

In addition, the regions in which an increase of the cost is observed are further examined by inspecting the results of such parameter choices on test data (contingency tables and clustering outcome in feature space). Here, the effects underlying the increase of the cost can be divided into two types. On the one hand, undue high values of s_R and q_R, lying in the region marked in light green in Fig. 5.13(b), reduce the ability to detect closely positioned clusters. This effect cannot be compensated by increasing the Coulomb wall and occurs for all tested V_C values. Thus, both clusters in the samples of Type 2 are recognized as one group as shown in Fig. 5.15(d) (the corresponding original data labeled with the ground truth is depicted in Fig. 5.15(c)).

On the other hand, combinations of low s_R and q_R within the region marked in gray in Fig. 5.13(b) lead to the splitting of large cohesive clusters as demonstrated in Fig. 5.15(b) (the corresponding original data labeled with the ground truth is depicted in Fig. 5.15(a)). This effect can be partly compensated by decreasing the height of the Coulomb barrier, however not for all data tested.

5.6.3 Estimation of Optimal a and η

After determining $q_R = 0.15$, $V_C = 50$, $s_R = 1.5$ and $a_{max} = 0.06$ as optimal values for those parameters in the first part of the analysis, the remaining parameters, a and η, are next variated as follows:

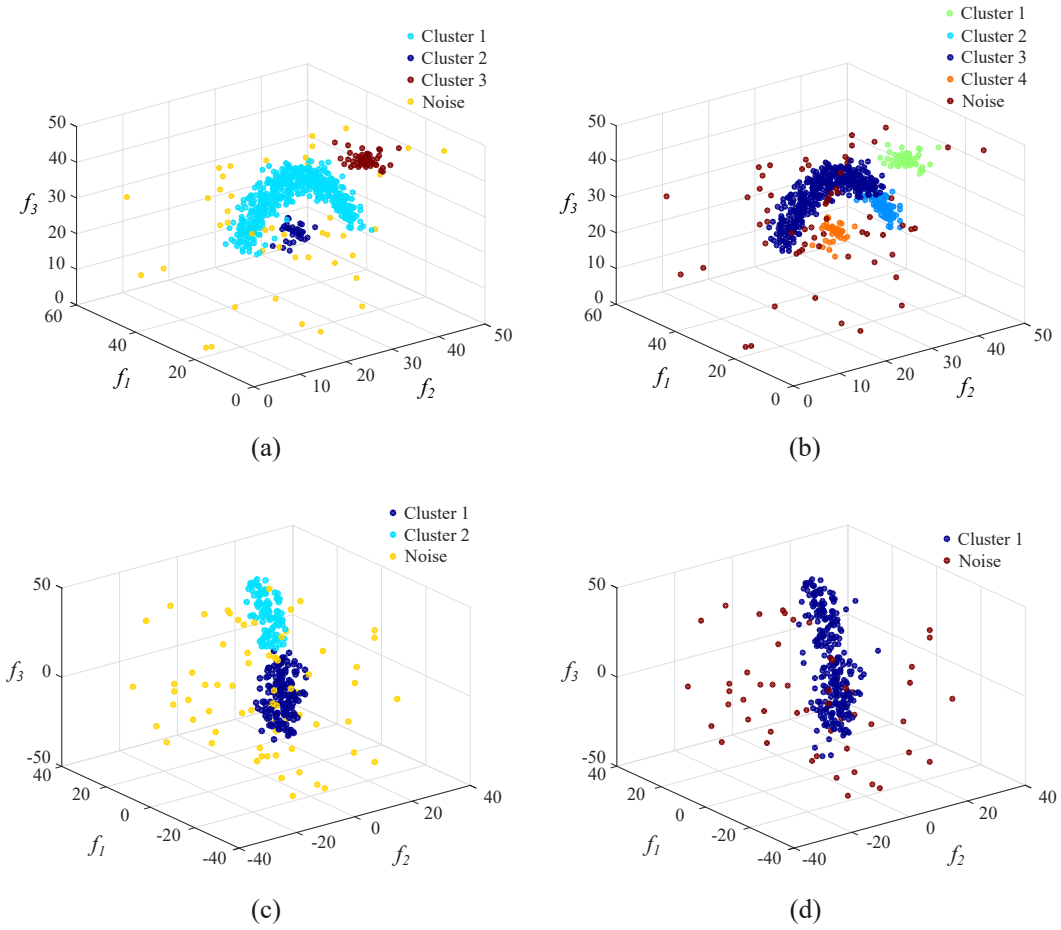


Figure 5.15: Effects underlying the increased cost values in s_R-V_C-q_R space (the remaining parameters are fixed to $V_0 = 3500$, $a = 0.04$, $\eta = 120$). (a) Test set of Type 1 labeled with the ground truth. (b) Results obtained with $a_{max} = 0.06$, $s_R = 1.3$, $V_C = 70$, $q_R = 0.1$, a point located in the region of parameter space marked in gray in Fig. 5.13(b). (d) Results obtained with $a_{max} = 0.06$, $s_R = 1.8$, $V_C = 30$, $q_R = 0.18$, a point located in the region marked in light green in Fig. 5.13(b).

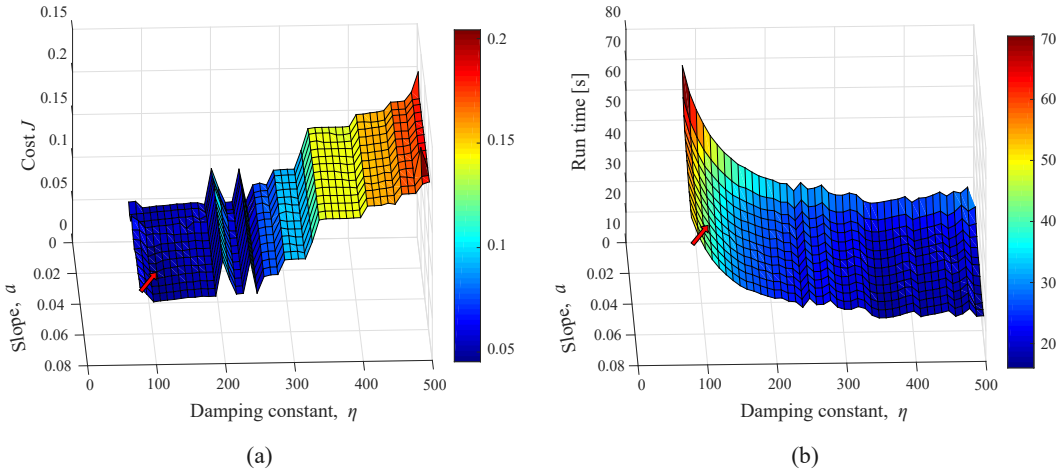


Figure 5.16: Second NPC optimization step concerning the parameters a and η. (a) Cost, J, calculated as a function of a and η. (b) Runtime as a function of a and η. The remaining parameters are fixed to $q_R = 0.15$, $V_C = 50$, $s_R = 1.5$ and $a_{max} = 0.06$. The chosen parameter optimum is marked with a red arrow in both plots.

$$\begin{aligned} a &= 0.01, 0.015 \ldots, 0.06 \\ \eta &= 80, 90, \ldots, 500 \end{aligned} \tag{5.21}$$

Figure 5.16(a) shows the resulting cost, J, as a function of a and η. Here, the average error rate exhibits a minimum at $a = 0.045$ and $\eta = 120$ which are thus chosen to be the default NPC values for these parameters.

In addition to the cost, J, the average runtime as function of a and η is assessed. As previously discussed, values $\eta < 70$ prevent the NPC algorithm from reaching the breaking condition as they do not sufficiently damp the particles and cause the systems to oscillate infinitely. This ultimately affects the NPC performance as can be clearly observed in Fig. 5.16(b): in the case of $\eta < 80$ the runtime increases rapidly, which confirms the initial considerations.

5.6.4 Iteration Number K

All of the parameter selection experiments described so far, are conducted by running the NPC procedure once and assuming the particles start at rest. The default number of NPC iterations is therefore $K = 1$. However, as described in Sec. 4.2.5, the dynamic modeling can be expanded by repeating the whole algorithm K times with different initial velocities. This extension aims to minimize the incidental assignment of noise points to one of the clusters formed. It affects solely nucleons located near the maximum of the potential barrier delimiting each cluster. Figure 5.17 schematically demonstrates the particular effects of the three possible scenarios that can occur. Here, a two-dimensional toy problem is shown in which a noise point (red square) is positioned at the rim of a cluster (green stars). The range of the

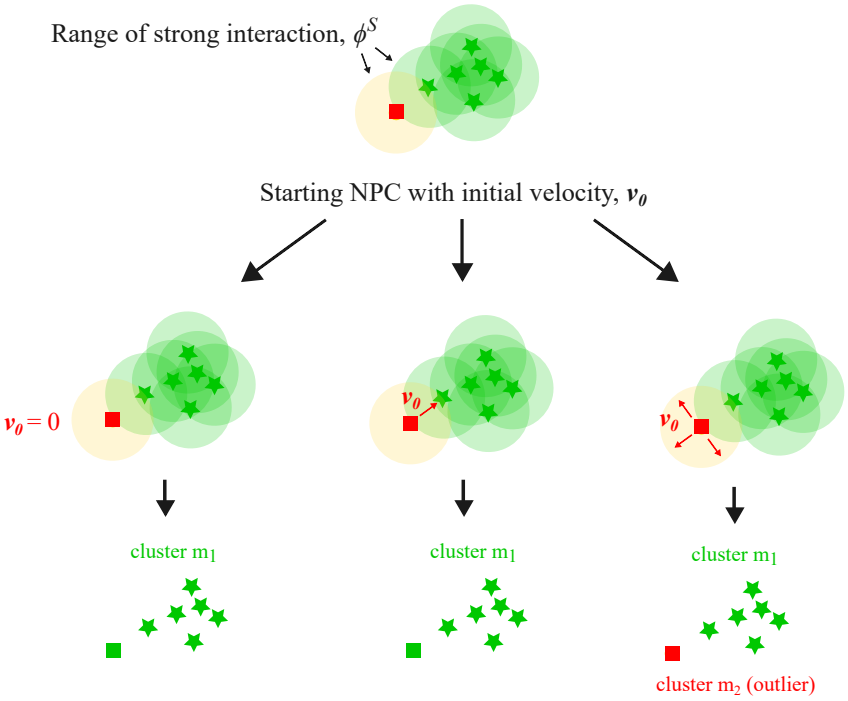


Figure 5.17: Schematic demonstration of the effects resulting from starting NPC with different initial particles' velocity, $\mathbf{v_0}$. The range of attractive strong interaction of each particle is represented by a circle. A noise point (red square) is positioned at the border of a cluster (green stars). Starting NPC with $\mathbf{v_0} = 0$ leads to the detection of a single cluster, including the outlier. The same clustering outcome results in the cases where the initial velocity vector points towards the potential well of the cluster (green area). In all other cases, the red particle will be detected as noise (as long as the associated kinetic energy is higher than the effective strong interaction potential of nearby nucleons).

strong attractive interaction of each nucleon is depicted by a circle (green for the cluster points and orange for the noise point). Due to the moderate overlap of the green and the orange areas, starting NPC with zero initial velocity, $\mathbf{v_0} = 0$, leads to the detection of a single cluster, including the outlier point. The same clustering outcome results in the cases when the noise point is assigned a nonzero initial velocity directed towards the potential well of the cluster (green area in Fig. 5.17). In all other cases, the red particle will be detected separately as noise. The latter effect is enhanced when increasing the data dimensionality d, as each dimension brings along an additional translational degree of freedom. Thus, the chance of a particle to be initially "tapped" towards the cluster decreases and outliers are detected more effectively.

If one is interested in closely exploring nucleons located in marginal regions, running NPC with $K = 10$ is recommended. As stated in Eq. 4.11, the final partitioning of the data is based on the final potential energy of each particle averaged over the K iterations. At equilibrium, P particles that have built a nucleus have final potential energies $U_i \approx PV_0$. Clearly, a nucleon accelerated towards the clus-

ter during the first NPC run has $U_i^1 \approx PV_0$, as well. If in some of the following iterations the same particle has the final potential energy typical for a noise point, $U_i^1 \approx 0$, the value employed for cluster assignment, $U_i = \overline{U_i^{(k)}}$, will be $U_i \neq 0$ and $U_i \neq PV_0$, lying in the interval $PV_0 < U_i < 0$. Thus, it will be assigned a distinct label and appears in the results as a singleton cluster. If the researcher is interested to disclose which is the noise point's neighboring cluster, he/she can compare the outcome with the results based on U_i^1 where the particle is assigned to the cluster nearby. Note that the resulting $U_i = \overline{U_i^{(k)}}$ can theoretically happen to be equal to a final potential energy characteristic for some other cluster. As, however, the peaks built in the final potential energy distribution are very sharp (the particles in a nucleus have U_i values typically equal up to the fifth digit), this practically is not observed in any of the test data.

5.7 Advice for Applying NPC

When analyzing data using NPC, it is recommended to employ the default parameter values as estimated in the current chapter and summarized in Table 5.6.

To sum up, NPC offers the opportunities of

- detecting major clusters and noise (by running it with the default parameter set and $K = 1$);
- closely studying marginal noise points (by increasing the number of iterations, K, as discussed in Sec. 5.6.4);
- exploring sub-groups (as described in Sec. 4.2.6)
  - either by considering noise in every additional clustering step
  - or by reassigning the outliers.

In addition, as the effects of the NPC parameters are well studied, the researcher can vary them to deliberately achieve different effects. For example, higher sensitivity to local density drops can be achieved by

- increasing V_C (e.g. $V_C = 100$);
- decreasing s_R and a_{max} (e.g. by testing $s_R = 1.1$, $a_{max} = 0.055$ or disable potential modulation completely by choosing $s_R = 1$ and $a_{max} = a$);
- increasing the damping constant, η.

Note that increasing η has the additional advantage of speeding up the algorithm (see Fig. 5.16). Thus, to attain a deliberate cluster fragmentation (for example if one wants to explore which points inside of a cohesive cluster are more closely positioned), it is recommended to vary the damping constant first.

On the other hand, one can cause the opposite effect and increase the robustness against density fluctuations (for example if one is explicitly interested to cluster closely positioned groups together) by:

- decreasing V_C (e.g. $V_C = 10$); note that $V_C = 0$ leads to the detection of a single group containing all points as discussed in Sec. 5.5.1 and is thus not suitable for any clustering task;
- increasing s_R and a_{max} (e.g. by testing $s_R = 1.7$, $a_{max} = 0.07$).

Recall, that reducing the value of the damping constant, $\eta < 70$, prevents the NPC from reaching the breaking condition and has thus not the effect of increasing robustness against local density fluctuations.

Table 5.6: Summary of the default NPC parameters' and the general effects of increasing or decreasing their values. Choosing a value greater than the default is marked with an upward pointing arrow, while choosing a value lower than the default is marked with a downward pointing arrow.

	Parameter						
	V_0	V_C	a	a_{max}	q_R	s_R	η
Default value	3500	50	0.04	0.06	0.15	1.5	120
Increase sensitivity to density fluctuations	3500	↑	↓	↓	↓	↓	↑
Increase robustness against density fluctuations	3500	↓	↑	↑	↑	↑	$*$

$*$ decreasing η prevents the NPC from reaching the breaking condition

Chapter 6

NPC Performance Evaluation and Comparison with Standard Clustering Methods

Contents

The current chapter presents a performance study and a comparison of the developed NPC algorithm with four standard clustering methods. First, the application details concerning the considered approaches are briefly discussed in Sec. 6.1. Next, two types of clustering problems that might be potentially problematic for NPC are closely studied in Sec. 6.2 and Sec. 6.3, respectively. Section 6.4 presents a comparison of the clustering algorithms considered. Based on a series of artificially generated samples representing different (nontrivial) clustering problems, the performance is evaluated and compared with regard to accuracy and time efficiency in Sec. 6.4.1. In Sec. 6.4.2 a multivariate benchmark real data set is considered, commonly employed for the evaluation and comparison of machine learning techniques.

6.1 Application Details of the Considered Algorithms

To achieve an informative comparison, NPC is tested along with the most prominent representatives of the clustering algorithm family whose strengths and weaknesses are well known to the research community: K-means as a representative of hard clustering methods; hierarchical clustering with single and complete linkage; and DBSCAN, which is the most prominent density based clustering approach. No fuzzy clustering algorithms are discussed in this context, as the NPC produces only a hard and/or hierarchical partitioning of the input data. Though briefly discussed

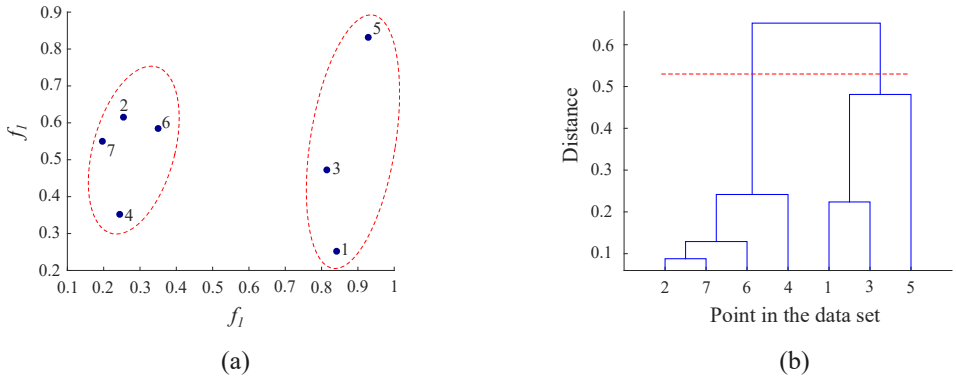


Figure 6.1: Demonstration of agglomerative hierarchical clustering. (a) Test data. (b) A dendrogram of the test data displayed in (a) generated with complete linkage. Cutting the tree at the level marked with a red dashed line in (b) results in a clustering of the sample in two groups as denoted by the red dashed ellipses in (a).

in Chapter 3, the application details concerning the compared algorithms are given in the following paragraphs.

K-means The K-means algorithm needs as input argument the number of expected groups in the studied data, K. It starts by placing K centroids at random locations in feature space. Next, in an iterative manner, the data points are associated with the nearest centroids and based on these associations, the centroids' positions are consecutively corrected. The process is then terminated when the assignments no longer change [Steinhaus 1956, Lloyd 1957, Lloyd 1982].

Agglomerative hierarchical clustering The basic idea underlying agglomerative hierarchical clustering is to consider every observation as a singleton cluster and then successively combine the most closely positioned clusters into larger group(s) until all sub-clusters are combined. The resulting tree-like structure of the data is typically represented by a dendrogram as shown in Fig. 6.1(b). A horizontal line in a dendrogram represents the merging of two groups, while the ordinate value corresponding to a horizontal line indicates the similarity of the merged clusters. To obtain a hard partitioning of the sample, the tree needs to be cut at some level. Different cutoff criteria exist, such as choosing a prespecified level of similarity or cutting at the largest gap between two successive merging levels (more on this topic can be found e.g. in [Jain 1988, Duda 2001]). In the following experiments, the number of clusters in the test data is employed as such criterion.

For the hierarchical clustering procedure, an evaluation of pair-wise distances between the clusters is needed. Different criteria measuring the similarity of sets exist, which are called linkages. The most commonly employed types are single linkage and complete linkage. In the case of single linkage, the distance between sets of observations, e.g. sets $\mathcal{A}$ and $\mathcal{B}$, is estimated as the minimum pairwise

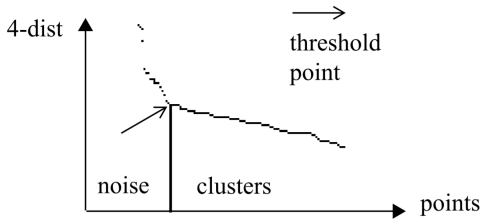


Figure 6.2: Sorted 4-dist graph for an exemplary sample database used to demonstrate the recommendation for choosing the *Eps* parameter for DBSCAN in [Ester 1996].

distance between the points in the different sets,

$$\min\{dist(a,b) : a \in \mathcal{A}, b \in \mathcal{B}\},$$

where *dist* is the metric chosen to compute the distance between points a and b (typically the Euclidean distance). In contrast, in the case of complete linkage the distance between sets of observations is given by

$$\max\{dist(a,b) : a \in \mathcal{A}, b \in \mathcal{B}\}.$$

DBSCAN Originally proposed by [Ester 1996], DBSCAN has two parameters that need to be specified. The first one, *Eps*, defines the radius of neighborhood around each data point in a sample, while the second one, $MinPts$, specifies a minimum number of neighbors within the *Eps* radius. Based on these two arguments, the points in a sample are recognized either as a core point (whose number of neighbors is greater than or equal to $MinPts$), or border point (whose number of neighbors is less than $MinPts$ but it is in the neighborhood of another core point), or noise (neither a core nor a border point). The authors propose fixing the minimum number of neighbors to $MinPts = 4$ and estimating *Eps* interactively. Here, a function, k-dist, is defined which maps each point to the distance from its k-th nearest neighbor. A so called sorted k-dist graph is then constructed where the points are ordered in a descending order with respect to their k-dist values. The authors recommend to choose the *Eps* parameter as the ordinate value corresponding to the "knee" of the curve[1].

In the tests presented in 6.4.2, a modification of DBSCAN avoiding the manual selection of *Eps* as introduced by [Daszykowski 2001] is also considered. In this NP (natural patterns) algorithm, only the $MinPts$ has to be prespecified which makes it advantageous in cases where no distinct "knee" of the sorted k-dist graph can be estimated.

[1] The original recommendation is "The threshold point is the first point in the first "valley" of the sorted k-dist graph [...]. All points with a higher k-dist value (left of the threshold) are considered to be noise, all other points (right of the threshold) are assigned to some cluster.".

NPC NPC does not require *any* a priori knowledge or an interactive decision concerning the analyzed data. Being based on the strong interaction acting between nearby positioned data points in feature space, it is most closely related to the group of density based clustering algorithms. As previously discussed in Sec. 3.2.2, the common criticism concerning this type of algorithms is that they are sensitive to differences in the clusters' densities. Thus, the following performance study starts with an experimental series aiming to explore the feasibility of NPC to handle groups of different densities.

6.2 NPC Applied on Samples Containing Clusters of Different Densities

To explore effects related to density changes of a cluster, an experimental setting previously employed in Sec. 5.5.3 is employed. Here, test samples of Type 2 (cf. Table 5.2) are generated with $d = 3$, $n_{noise} = 50$, $n_{g_1} = 50$, $r_{g_1} = 5$, $l_{g_1} = 5$, $n_{g_2} = 50$ and $l_{g_2} = 5$. The radius of the second cluster is gradually increased as follows:

$$r_{g_2} = 10, 12, \ldots, 30 \tag{6.1}$$

Keeping the number of points $n_{g_2} = 50$ constant leads to a drop of the cluster density, ρ_{g_2}, when increasing r_{g_2}, as ρ_{g_2} is inversely proportional to $r_{g_2}^2$ (cf. Eq. 5.14).

Up to $r_{g_2} = 20$ (which corresponds to a difference in the clusters' densities by the factor $\frac{\rho_{g_2}}{\rho_{g_2}} = 15$) both clusters and the noise points are correctly detected by the NPC method. Yet, increasing r_{g_2} leads to a continuous increase of the run time. In the cases of $r_{g_2} > 20$, the systems cool down very slowly, so that the NPC breaking condition cannot be reached in a reasonable computational time. However, when analyzing clusters of very different densities, the temperature evolution curve exhibits a characteristic form (different from the T-curves typical for clusters of similar densities, such as the one demonstrated e.g. in Fig. 4.6(b) or Fig. 6.6(d)). This phenomenon can be employed to define an additional breaking condition and cure the emerging problem as described in the following.

As shown in Fig. 6.3, the $T(t)$-curve typically starts with a rapid increase of the temperature. Observing the particles' motion in feature space and in U-E_{kin} space as demonstrated in Fig. 5.1 and Fig. 5.2, respectively, reveals that this rapid temperature increase corresponds to the fusion process of the dense cluster. This region is then followed by a plateau. Its form is typically not flat, but rather "cluttered" with minor maxima and minima, whereby the maxima reach increasingly high values at later time points and the minima do not reach low temperature values (Fig. 6.3). This temperature evolution results from the fact that within the more sparsely populated cluster, the particles' mutual attraction at the beginning of the dynamic modeling is weaker than the interaction within the dense cluster. Thus, the points in the sparse cluster are fist slowly accelerated towards local potential minima in feature space building small groups, which then fuse to build larger and larger groups. When these small groups are accelerated towards each

other, the system temperature rises and this corresponds to the minor maxima in the $T(t)$-curve. At the same time, the minima separating them do not reach low temperature values, indicating that the fusion of the subgroups is ongoing and the system has not yet reached equilibrium. However, it is also typical for such samples that the final cool-down phase is strongly prolonged. It is characterized by an offset between the temperature "tail" at later time points and the zero line (red line in Fig. 6.3). The temperature values in this region decrease very slowly thus preventing the systems from reaching the temperature-based breaking condition. Merely increasing the damping constant η cannot cure this problem, as this would stick the particles of a sparse cluster into local minima and prevent them from fusing to a nucleus at all. Thus, another additional stopping condition is introduced. Here, the first derivative of the temperature as a function of time, $\frac{d}{dt}T(t)$ (green dashed line in Fig. 6.3), is employed and the process is terminated, when its value pertains negative $\frac{d}{dt}T(t) < 0$ for more than 1500 consecutive time steps. This large number of time steps is deliberately chosen, so as to assure that all points in local potential minima have enough time to fuse and build nuclei. If the NPC analysis is terminated because this breaking condition is met, the researcher/user is notified that the found groups are presumably of different densities.

Introducing this additional stopping condition allows to examine sparse clusters generated with values $r_{g_2} > 20$. The results show that up to $r_{g_2} \approx 22$, which corresponds to a difference in the clusters' densities approximately by the factor $\frac{\rho_{g2}}{\rho_{g_2}} = 17$, both groups are correctly identified (Fig. 6.4(a), (c) and (e)). Due to the rapidly decreasing density, larger r_{g_2} values increasingly lead to the recognition of some fractions of the sparse cluster as separate groups and marginal cluster points as noise (as in the sample demonstrated in Fig. 6.4(b), (d) and (f)). This, however, is not necessarily to be considered as a drawback, as these points typically have distances towards the rest of the group lying in the range of the distances between the noise points.

Furthermore, the introduced additional breaking condition is generally met in the cases of sparse samples where the particles' motion is largely or primarily dominated by the repulsive Coulomb force. This occurs for example, in the cases of unstructured data or high dimensional data including many non-discriminative features. One such sample comprised of 20 points randomly positioned in a 15-dimensional space is demonstrated in Fig. 6.5. In such cases, no clusters or only small clusters comprised of a few nucleons are built. Note, however, that the temperature here starts with a rapid increase and the typical plateau at the beginning of the curve as described previously in this section is not observed (see Fig. 6.5(a)). A possible (though uncommon) constellation here may also be that even the constitutes of the already fused small clusters repel each other to such extend that all of them reach a positive final potential energy. In the current implementation of NPC, the user is noted that this effect is observed and has the possibility to decide interactively how to treat these points. Since the cluster-building particles here fuse initially and thus exhibit at early time points the typical negative potential energy of approximately $N_i V_0$ (N_i number of members in the i-th cluster), the standard NPC routine detects

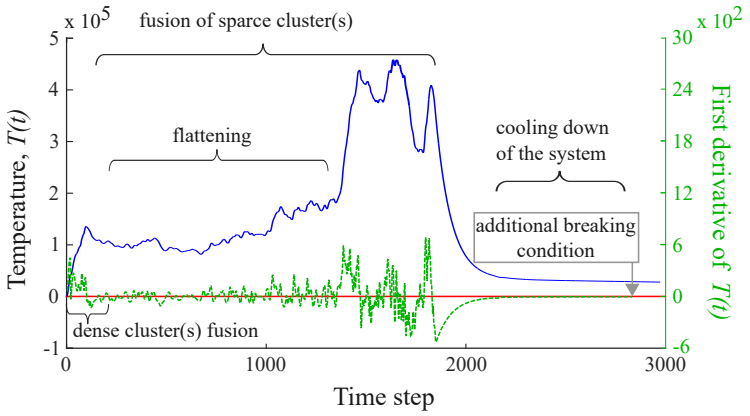


Figure 6.3: A temperature curve (blue) typical for samples containing cluster of different densities. It starts with a rapid increase corresponding to the fusion of the dense cluster(s), followed by a flat region characterized by multiple minor minima and maxima. Typically, the maxima values increase at later time points due to the fusion of sparce clusters and the minima do not reach low temperature values. In the final phase, the system cools down very slowly, so that an additional breaking condition based on the first derivative of the temperature (green curve) is introduced.

the cluster(s) found based on the minimal potential energy reached during dynamic simulation. The other possibility is to treat these points as one unordered group of noise.

6.3 NPC Applied on Prolonged Clusters

In the case of NPC, the particles interaction is governed not only by the short-ranged strong attractive force but also by the long range repulsive Coulomb force. Thus, a series of tests is conducted that aims to explore whether the cumulative long range repulsion leads to an undue split of prolonged clusters. It is analogous to the experiments described earlier in Sec. 5.5.3 aiming to study the effects related to changes in the clusters' form for the purposes of the NPC parameter estimation.

Here, three-dimensional ($d = 3$) test samples of Type 2 are considered with one of the clusters chosen to be a compact spheroid generated with $n_{g_1} = 50$, $r_{g_1} = 5$ and $l_{g_1} = 5$. The second cluster has $n_{g_2} = 50$ points, width specified by $r_{g_2} = 5$ and its length is gradually increased as follows:

$$l_{g_2} = 10, 20, \ldots, 110 \tag{6.2}$$

As in Sec. 5.5.3, here again solely the effects of varying the cluster form are studied by keeping the average intra-cluster particles density approximately constant through increasing the number of points, n_{g_2} proportional to the length, l_{g_2} (cf. Eq. 5.13 and Eq. 5.14). Furthermore, both cohesive groups are overlaid with $n_{noise} = 50$ random noise points.

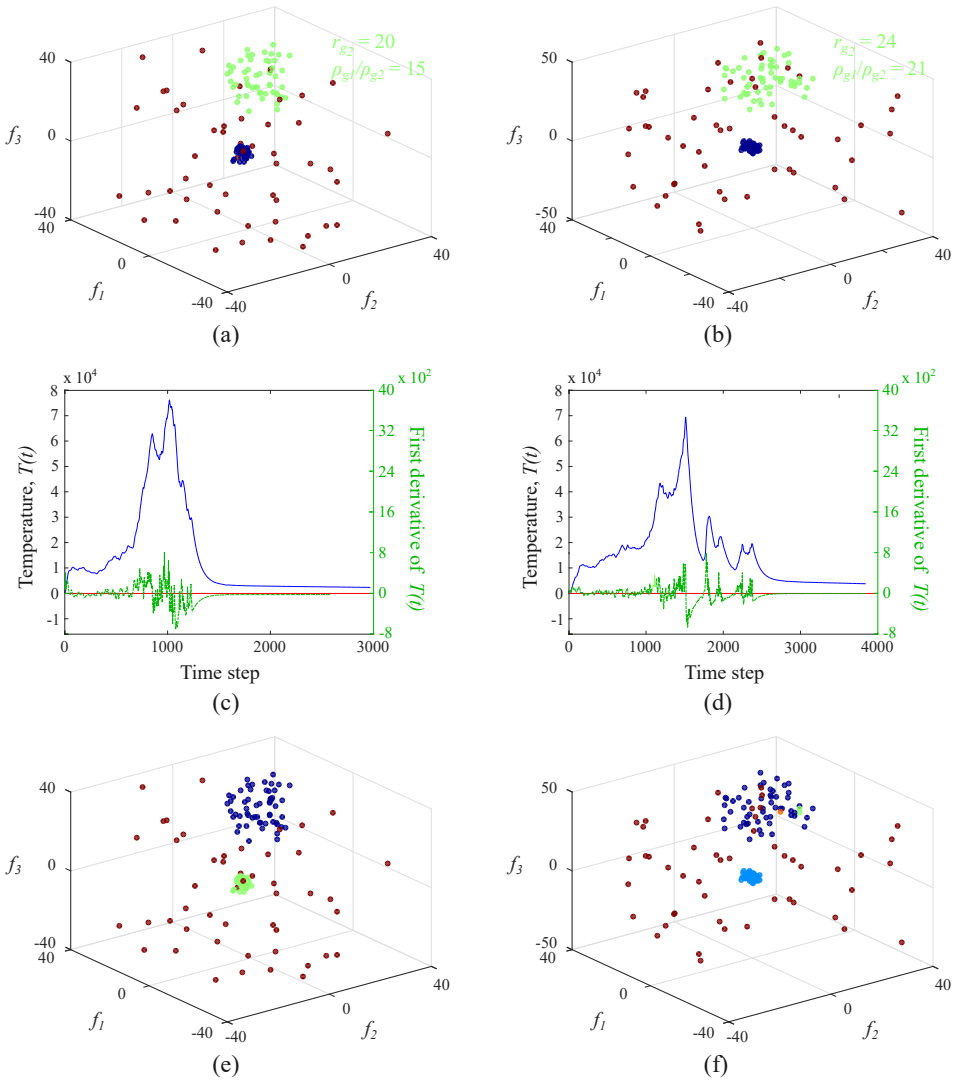


Figure 6.4: NPC performance in the case of samples containing groups of different densities. (a) and (b) show the test data labeled with the ground truth. (c) and (d) demonstrate the temporal temperature evolutions and the first derivatives, $\frac{d}{dt}T(t)$, for the samples depicted in (a) and (b), respectively. In the last row, (e) and (f) show the corresponding NPC results.

For all test samples studied, NPC correctly detects the two groups regardless of the grade of prolongation. This is due to the potential modulation introduced in Sec. 4.2.3 and further discussed in Sec. 5.5.2 and Sec. 5.5.3, that prevents the fragmentation of the prolonged formations. This is amplified in the following three examples shown in Fig. 6.6. The first one, Fig. 6.6(a), contains two compact clusters generated with the above listed parameters and $l_{g_2} = 10$. As it can be seen in Fig. 6.6(d), the temperature curve in such cases, typically exhibits one main maximum after which the system reaches equilibrium with no apparent oscillations.

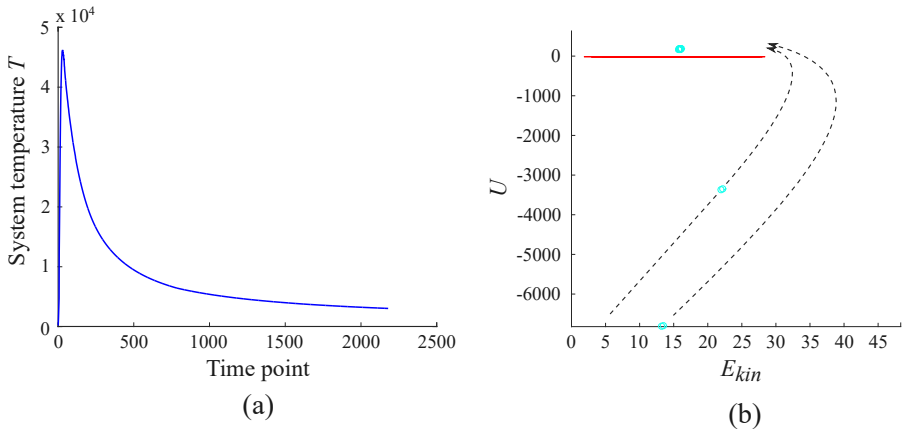


Figure 6.5: Energy evolution of a sample containing unstructured data where only two groups, each comprised of two nucleons, are found. (a) Temperature evolution curve. (b) The points fusing to a cluster initially reach a negative potential energy of $2V_0$. Due to the dominant Coulomb force in this case, they are eventually accelerated to such extend that they reach a positive final potential energy. The observed clusters' motion trajectory in $U - E_kin$ space is indicated by the dashed arrows.

Furthermore, a test sample of Type 1 is displayed in addition to the samples of Type 2 studied so far. Here, the clusters also have similar densities (generated with $r_h = 5$, $n_h = 550$, $r_{si} = 3$, $n_{si} = 50$, $r_{so} = 4$, $n_{so} = 70$, $n_{noise} = 35$) and the length of the curvilinear formation is approximately $l_h = 60$. Though the bent form of the elongated structure has apparently no influence on the ability of NPC to detect the group, the increased length leads to an increased oscillation of the system (indicated by the additional multiple minor minima and maxima) and a shift of the simulation termination to later time points (Fig. 6.6(e)).

Fig. 6.6(f) displays the T-curve of a sample of Type 2 generated with the above listed parameters and $l_{g_2} = 110$ (the data labeled with the ground truth is displayed in Fig. 6.6(c)). It shows that the effect of increased systems oscillations enhances when increasing the cluster's length. Despite the resulting increased computational time, the observed pattern of the temperature evolution curve may be utilized as indicator that the sample studied with NPC may contain (strongly) prolonged cluster(s). Thus, it is advantageous to not only behold the NPC clustering outcome, i.e. the points' labels, but also the temperature evolution of the studied sample.

6.4 Comparison with Standard Clustering Algorithms

6.4.1 Artificial Data

The following performance evaluation and comparison of NPC with other clustering algorithms is based on five test samples with different specifications. The details concerning their generation are summarized in Table 6.1.

The clustering results for all considered algorithms are summarized in Fig. 6.7.

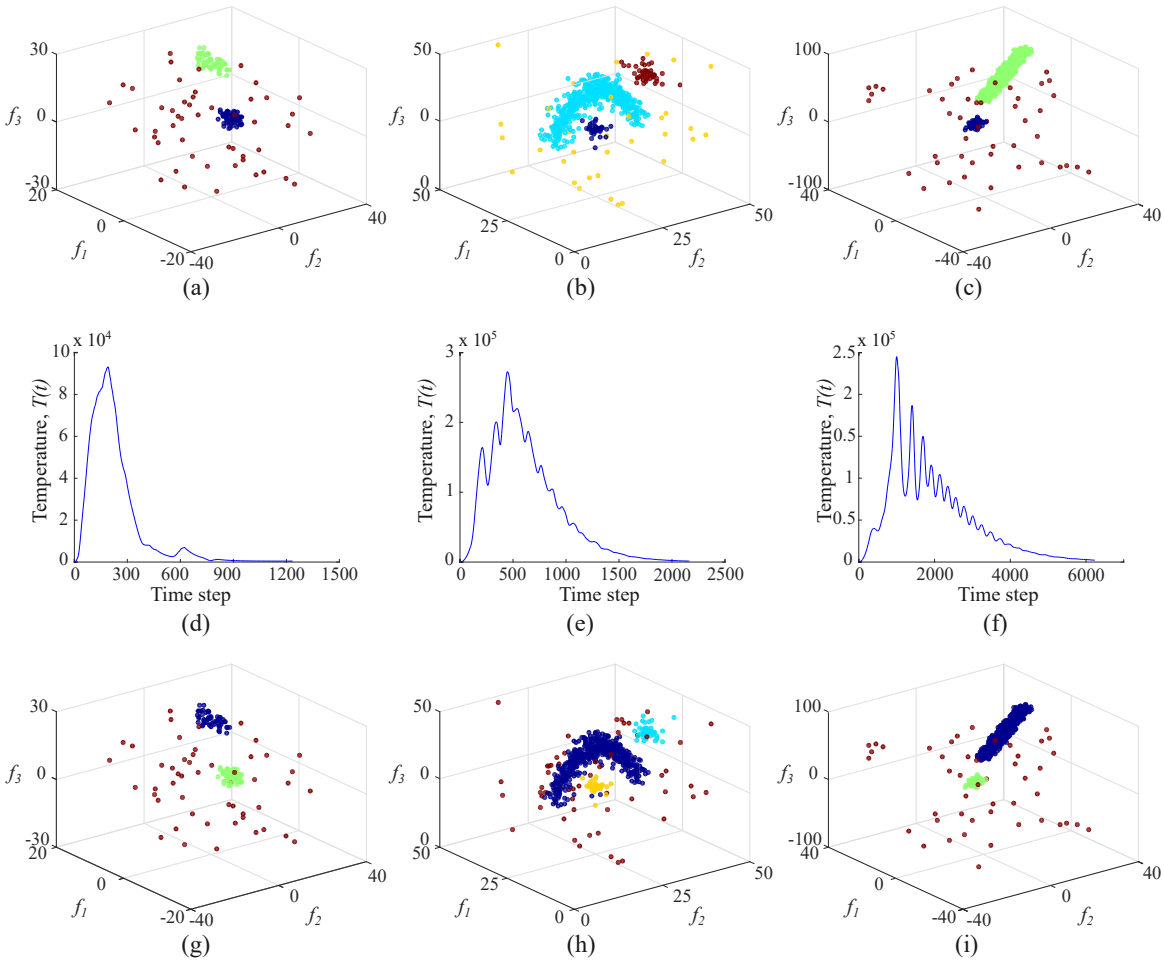


Figure 6.6: Exploring the efficiency of NPC to detect clusters of prolonged form. Three test samples are shown which are generated as described in the text. (a) Three-dimensional test data of Type 2 containing two compact spheroids overlaid with noise. (b) Test sample of Type 1, where one of the clusters is elongated ($l_h = 60$) and of irregular form. (c) Three-dimensional test data of Type 2 containing a compact spheroid, a prolonged cluster of length $l_h = 120$, both overlaid with noise. (d), (e) and (f) display the temperature evolution curves corresponding to the samples (a), (b) and (c), respectively. The last panels, (g), (h) and (i) show the results obtained with NPC.

Here, each column corresponds to a test sample. In the first row, the data sets labeled with the ground truth are presented, followed by the NPC results in the second one and the corresponding systems' temperature evolution curves in the third one. The fourth row displays the DBSCAN 4-dist curves employed to manually select a threshold (marked with a red square) as explained in Sec. 6.1. The corresponding DCSCAN results are displayed in the following fifth row. Finally, the results obtained with single linkage and complete linkage hierarchical clustering are shown in the sixth and the seventh row, respectively, followed by the K-means partitionings shown in the last, eight row.

The corresponding Rand index values for all experiments are summarized in Ta-

Table 6.1: Comparison of NPC with four standard clustering methods for a series of five different clustering problems.

Sample	Data specifications
1.	Type 2 test data with $d = 3$, $r_{g_1} = 5$, $l_{g_1} = 5$, $n_{g_1} = 50$, $r_{g_2} = 20$, $l_{g_2} = 5$, $n_{g_2} = 50$, $n_{noise} = 50$ (total number of data points $N = 150$).
2.	Type 1 test data with $r_h = 5$, $n_h = 550$, $r_{si} = 3$, $n_{si} = 50$, $r_{so} = 4$, $n_{so} = 70$, $n_{noise} = 35$ ($N = 805$).
3.	Three spheroids generated analogously to the clusters constructed in test data of Type 2 overlaid by $n_{noise} = 50$ noise points. All of them are comprised of 50 points and have radii $r_s = 5$. Two of the spheroids are closely positioned, while the third one is well separated ($N = 200$).
4.	$N = 500$ points randomly positioned in feature space, no cluster structure emulated.
5.	A spheroid containing $n_s = 30$ points, with radius $r_s = 1$ surrounded by a torus specified by $d_{torus} = 2$ and $n_{torus} = 400$ as described in the text. Both clusters are overlaid by 20 randomly positioned noise points ($N = 450$).

ble 6.2. Furthermore, the assessed run times in all considered cases are displayed in Table 6.3. For a more informative comparison, the measured time values are displayed along with the general time complexity for each of the considered algorithms as a function of the input parameters and/or test sample size and dimensionality.

1. Sample The first sample consists of two spheroids of different densities (Type 2 test data with $d = 3$, $r_{g_1} = 5$, $l_{g_1} = 5$, $n_{g_1} = 50$, $r_{g_2} = 20$, $l_{g_2} = 5$, $n_{g_2} = 50$) overlaid by random noise ($n_{noise} = 50$). The clustering results show that in this case only NPC and DBSCAN are capable of detecting correctly both groups and separating them from the noise. Based on the temperature evolution of the sample displayed in the third panel of the first column of Fig. 6.7, NPC provides the additional information that the groups found are presumably of largely different densities.

In the case of DBSCAN, an interactive manual selection is needed, whereby in the concrete case a threshold at 4-dist=7.09 is selected (see third panel of the first column in Fig. 6.7). Note that in cases of clusters with different densities overlaid by noise such as in the current test data set, it is not trivial to select "the first point in the first "valley" of the sorted k-dist graph" as recommended originally by [Ester 1996]. Selecting for example the first "knee" value of the curve marked in green in the 4-dist plot, leads to the detection of merely 5 noise points while the rest is clustered as one group.

Considering the remaining three algorithms tested, single linkage hierarchical clustering fails to detect both groups and separate them from the noise, while com-

plete linkage hierarchical clustering and K-means split the data set into two groups containing the clusters. However, the surrounding noise is assigned to the nearby positioned group, thereby failing to identify the group of noise points.

2. Sample The second sample tested consists of three clusters overlaid by random noise: two compact spheroids, one of which is partly enclosed by a curvilinear cluster (test data of Type 1 with $r_h = 5$, $n_h = 550$, $r_{si} = 3$, $n_{si} = 50$, $r_{so} = 4$, $n_{so} = 70$, $n_{noise} = 35$). As in the previous case, here again only NPC and DBSCAN correctly detect the cohesive groups and separate them from the noise points. Even though the number of the searched groups is provided to the rest of the clustering methods, they all fail to correctly cluster them. Concretely, single linkage hierarchical clustering detects two singleton clusters and groups the rest of the sample together. In the case of complete linkage hierarchical clustering, the group positioned outside of the bent group is recognized but the elongated group is split into two fragments, one of which is clustered together with the enclosed spheroid. K-means fragments the "horseshoe" into three parts and clusters two of them together with one of the spheroids. Furthermore, both hierarchical clustering methods and K-means fail to detect the noise points and assign them to one of the found groups.

3. Sample The third sample is used to examine the ability to distinguish touching clusters in the presence of noise. Here, three spheroids are generated in an analogous way as the clusters in the samples of Type 2 (cf. Sec. 5.1) and overlaid by $n_{noise} = 50$. They are all comprised of 50 points and have radii $r_s = 5$. The centers of two of the spheroids are chosen to have distance $2r_s$, while the third one is well separated. In this case, best results are achieved by NPC which can successfully detect all clusters and separate them from the noise. DBSCAN cannot distinguish the two closely positioned groups and clusters them together. A threshold value of 4-dist=2.92 is chosen here, which is rather the second "knee" of the curve. However it is visually the more prominent one and leads indeed to better results as choosing for example the value marked in green in the corresponding sorted 4-dist graph (an assertion which is generally only possible in the presence of ground truth information). The three remaining methods examined cannot properly detect the groups. Single linkage hierarchical clustering detects a singleton cluster, one group consisting of points labeled as noise in the ground truth and one group combining the rest (the three clusters and the remaining noise points). K-means detects correctly the well separated group, but combines the two closely positioned spheroids and the surrounding noise into a single cluster. Furthermore, as in the case of single linkage, it clusters the same group of noise points together.

Finally, complete linkage hierarchical clustering detects one of the closely positioned groups as a separate cluster but groups the other one together with the well separated spheroid. In addition, it fails to detect the noise points as such.

4. Sample The fourth sample consists of 500 unordered points and aims to examine the performance of the different algorithms in the case where no cluster structure is present. Apart from one singleton cluster, NPC detects here all points as one group. Considering DBSCAN, the question of selecting a threshold in this case becomes increasingly problematic. The value selected here is 4-dist=7.35 which most likely corresponds to the first "knee" of the 4-dist curve. It leads to the detection of 15 points as noise and clustering the remaining data together.

Single linkage hierarchical clustering leads to similar results as NPC and detects a singleton cluster and a cluster containing the rest of the points. Complete linkage hierarchical clustering and K-means lead to a splitting of the sample into K groups (in this case $K = 2$) whereby K-means results in different partitionings of the sample when run multiple times (due to the random initialization of the cluster centroids, see Sec. 6.1).

5. Sample The last artificial test set examined contains a compact group surrounded by a completely enclosed cluster, both overlaid with random noise. The spheroid is constructed in an analogous way as the ones included in test data of Type 1 (cf. Sec. 5.1). The radius and the number of points within this group are set to $r_s = 2$ and $n_s = 50$, respectively. The torus is constructed by placing $n_{torus} = 400$ points at random positions on a circle of radius $r_s + d_{torus}$, where $d_{torus} = 5$ is the distance between the circle and the center of the spheroid. The points' radial distances towards the circle are sampled from a normal distribution and scaled by radius $r_{thorus} = 1$.

In this case, NPC detects both groups as a single cluster. The reason for this is that the particles constituting the enclosed outer structure start to shrink due to the mutual attractive strong intra-cluster forces. Thus, the distance between the outer and the inner cluster is continuously reduced. Though the Coulomb force provides for a repelling between distinct cohesive groups, in such cases it cannot prevail over the strong interaction, and both "nuclei" fuse to one structure, resulting in the distinction only between particles belonging to a cluster and noise points.

Considering the rest of the tested algorithms, only DBSCAN allows for the successful detection of both groups and the outliers. The rest of the methods do not lead to a correct detection of the groups and the noise points. Furthermore, as in the case of the previous test sample, K-means results in different partitionings when run multiple times on the same sample.

Table 6.3 shows the observed run times along with the general time complexity for each of the considered algorithms as a function of the input parameters and/or test sample size and dimensionality. As can be noted, NPC has the highest run times and a time complexity of the order $\mathcal{O}(N^2 d)$. However, a more thorough evaluation reveals that the computational time needed is mostly influenced by the spatial distribution of the data in feature space and only to a smaller extend by the sample size. Systems containing elongated, irregularly formed and/or nested clusters as well as unordered data tend to need more time to reach equilibrium. In the case

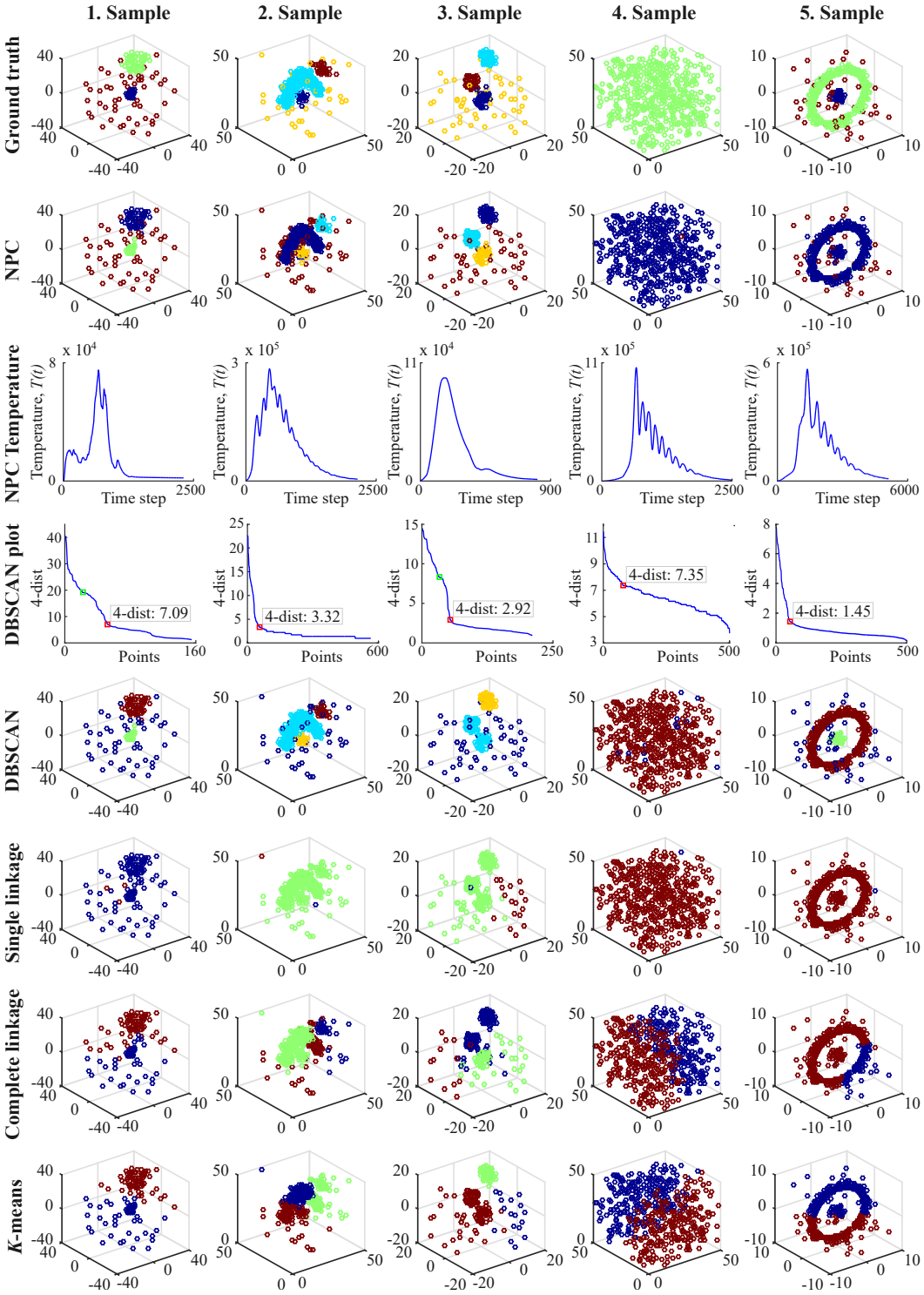


Figure 6.7: Comparison of NPC with four standard clustering methods based on a series of five test samples, representing different clustering problems. Details are given in Sec. 6.4.1.

Table 6.2: Rand index for NPC and four standard clustering methods calculated for a series of five different clustering problems.

Clustering method	Sample				
	1.	2.	3.	4.	5.
NPC	0.975	0.915	0.998	0.996	0.798
DBSCAN	0.975	0.973	0.991	0.946	0.989
HC single linkage	0.344	0.666	0.352	0.996	0.663
HC complete linkage	0.731	0.603	0.691	0.507	0.537
K-means	0.736	0.513	0.711	0.499	0.503

of DBSCAN, the threshold values for all samples are separately estimated and the time required for this evaluation is not included in the measurements displayed in Table 6.3. Single linkage and complete linkage hierarchical clustering, as well as K-means have the best time performance. However, they all fail to detect the studied clusters correctly in the presence of noise.

Table 6.3: Comparison of the time complexity for NPC and four standard clustering methods. Columns two to six display the measured times for the five considered test samples in seconds. The last column displays the time complexity for each of the considered algorithms as a function of the input parameters and/or test sample size and dimensionality. Here, N is the number of points in a sample, d the sample dimensionality (number of features) and K the number of groups to be found (needed to be prespecified only in the case of K-means).

Clustering algorithm	Sample					Time complexity
	1.	2.	3.	4.	5.	
NPC	33.615	79.032	13.828	79.453	156.474	$\mathcal{O}(N^2 d)$
DBSCAN	12.575	9.306	6.771	7.882	8.849	$\mathcal{O}(N \log(N))$ †
Single linkage	0.087	0.010	0.003	0.006	0.008	$\mathcal{O}(N^2)$ *
Complete linkage	0.006	0.007	0.003	0.006	0.006	$\mathcal{O}(N^2)$ *
K-means	0.006	0.010	0.006	0.014	0.015	$\mathcal{O}(NKd)$ * $\mathcal{O}(N^{dK+1})$ ‡

† [Ester 1996] * [Xu 2005] ‡ [Inaba 1994]

6.4.2 Fisher Iris Data

In addition to the series of different artificial clustering problems, the performance of the developed NPC algorithm is further tested on the Fisher Iris data: the benchmark dataset most commonly used to examine and compare novel clustering methods. Three classes of iris flowers are contained in the sample, Setosa, Versicolor and

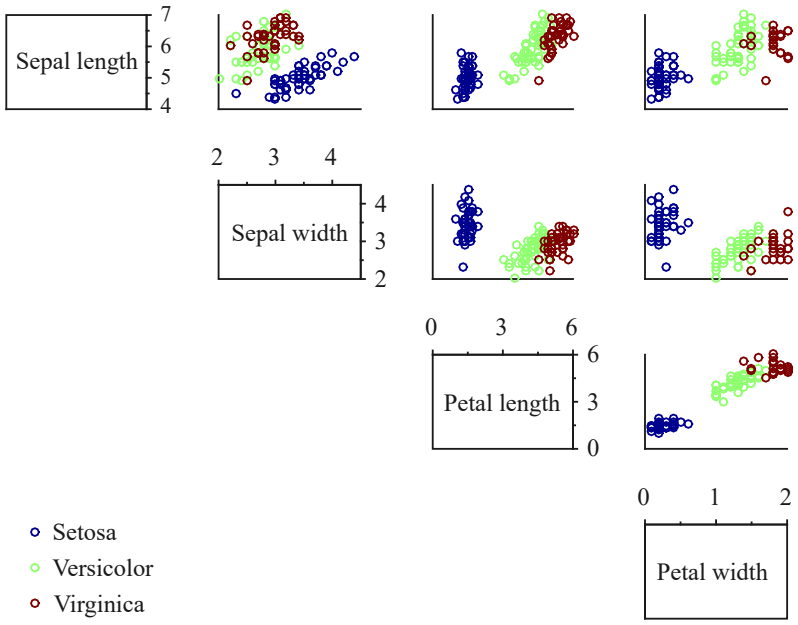


Figure 6.8: Fisher Iris data labeled in color with the ground truth.

Virginica, each having 50 observations and four features measured in centimeters: sepal length, sepal width, petal length and petal width. Importantly, the Setosa class is linearly separable from the rest, while the Virginica and Versicolor groups cannot be linearly separated from each other. The complete data set as published originally by [Fisher 1936] and employed in the current work is given in Appendix D, Table D.1[2]. Fig. 6.8 depicts the data labeled with the ground truth.

Figure 6.9 displays the results for the Fisher Iris sample obtained when applying NPC (denoted in the following as one-step NPC). Furthermore, a two-step NPC analysis is performed. Here, a search for subgroups is conducted without additional noise detection within the nested clusters (cf. Sec. 4.2.6) and the results are displayed in Fig. 6.10. The contingency tables corresponding to both cases are given in Table 6.4.

Table 6.4: NPC Analysis of Fisher's Iris Data. C_i denotes the i-th found cluster.

	One-step NPC			Two-step-NPC			
	C_1	C_2	Noise	C_1	C_2	C_3	Noise
Setosa	47	0	3	47	0	0	3
Versicolor	0	49	1	0	50	0	0
Virginica	0	41	9	0	16	25	9

[2]Note that there are different databases referring to as Fisher Iris Data, where some of the measurements are not consistent [Bezdek 1999]. In the current thesis, the sample is employed in its original version as published by [Fisher 1936].

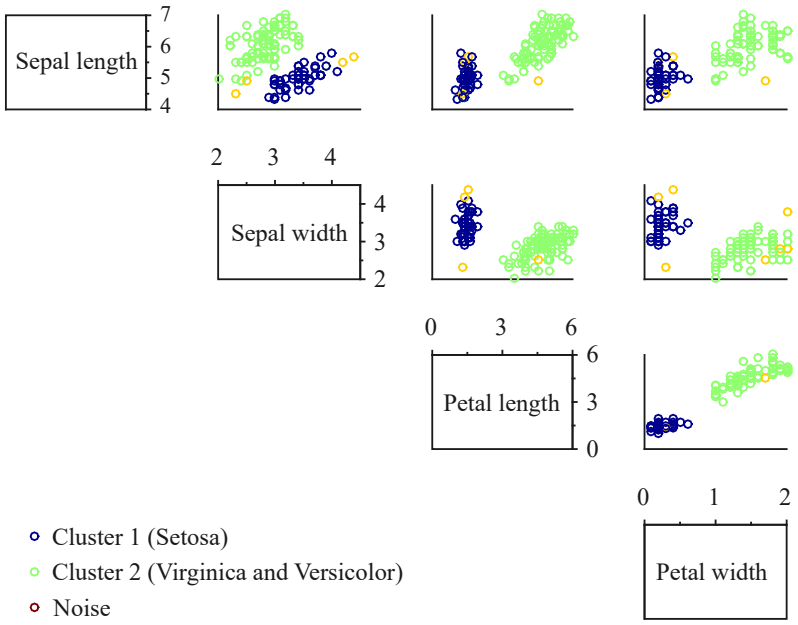


Figure 6.9: Fisher Iris data processed with one-step NPC.

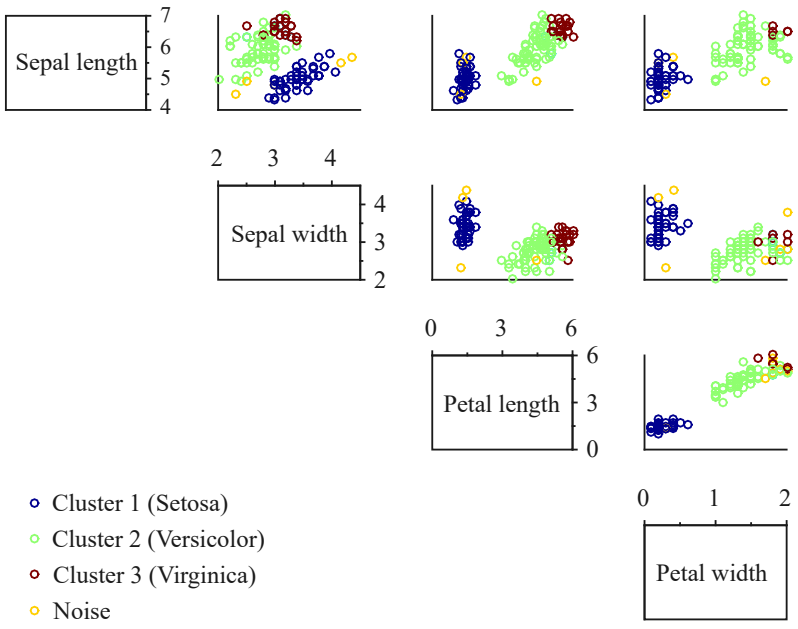


Figure 6.10: Fisher Iris data processed with NPC, where an additional search for subgroups is conducted (two-step NPC). Here, outliers are detected only during the first clustering step, while noise detected when searching for subgroups is reassigned as described in Sec. 4.2.6.

Processing the data with one-step NPC leads to a Rand accuracy of 0.77, as the adjoining Virginica and Versicolor clusters are pooled together (cf. Table 6.4, Table 6.5 and Fig. 6.9). Applying a two-step NPC, however, detects two further subgroups, resulting in a Rand accuracy of 0.85 (cf. Table 6.4, Table 6.5 and Fig. 6.10).

Next, the NPC results obtained for the Fisher Iris data are compared with DBSCAN, hierarchical clustering with single and complete linkage, as well as K-means.

In the case of DBSCAN as proposed by [Ester 1996], Fig. 6.12 shows clearly that the sorted 4-dist curve does not exhibit a distinct "knee" that can be employed to manually select the Eps threshold required for the further cluster analysis. In order to estimate whether there is any threshold leading to the recognition of the three classes, all 4-dist values in the range [0.1 1.1] are tested. Indeed, there is a threshold value that leads to the distinction of the three clusters as displayed in Fig. 6.11 at $Eps = 0.42$. However, as seen in Fig. 6.12 the corresponding region of the sorted 4-dist curve cannot be associated with any kind of a "knee" type curve. Moreover, selecting a slightly higher threshold, e.g. $Eps = 0.44$ (marked in green in Fig. 6.12), leads to the detection of Virginica and Versicolor as one cluster surrounded by 22 noise points, while selecting a slightly lower threshold, e.g. $Eps = 0.41$ (marked in yellow in Fig. 6.12), leads to the fragmentation of the points within Virginica and Versicolor into three clusters surrounded by 29 outliers (i.e. one third of the points in both groups are clustered as outliers). As there is no distinct visual criterion suggesting the best threshold value, this method cannot be employed for the unsupervised analysis of the Fisher Iris data. For this reason, a modification on DBSCAN without manual interaction is employed for the further analysis: the NP DBSCAN algorithm proposed by [Daszykowski 2001].

As previously stated in 6.1, the parameter $NumPts$ has to be prespecified for this method. It refers to the minimum number of points that is to be considered as a cluster. Since a low $NumPts$ value allows for the detection of smaller, possibly closely positioned groups, as it is the case of the classes Virginica and Versicolor, the following analysis is held by employing $NumPts = 1$. It leads to a Rand accuracy of 0.78. The well separated Setosa class is correctly recognized but Virginica and Versicolor are clustered together (apart from one detected noise point and a cluster comprised of four Versicolor points). To assure an objective comparison, the algorithm is then run for each of the groups found in analogous manner as the two-step NPC. The aim here is to ascertain whether NP DBSCAN can detect Virginica and Versicolor as separate clusters. The results from this two-step process are displayed in Fig. 6.14. It can be noted that, apart from the detection of additional noise points, Virginica and Versicolor are still not separated. The Rand index in this case is 0.74.

For the rest of the considered methods, hierarchical clustering with single linkage, hierarchical clustering with complete linkage and K-means, the best possible results are taken by specifying the number of clusters to $M = 3$.

As seen in Fig. 6.15(a), single linkage hierarchical clustering fails to disclose properly the structure of the data, detecting the well separated Setosa class correctly,

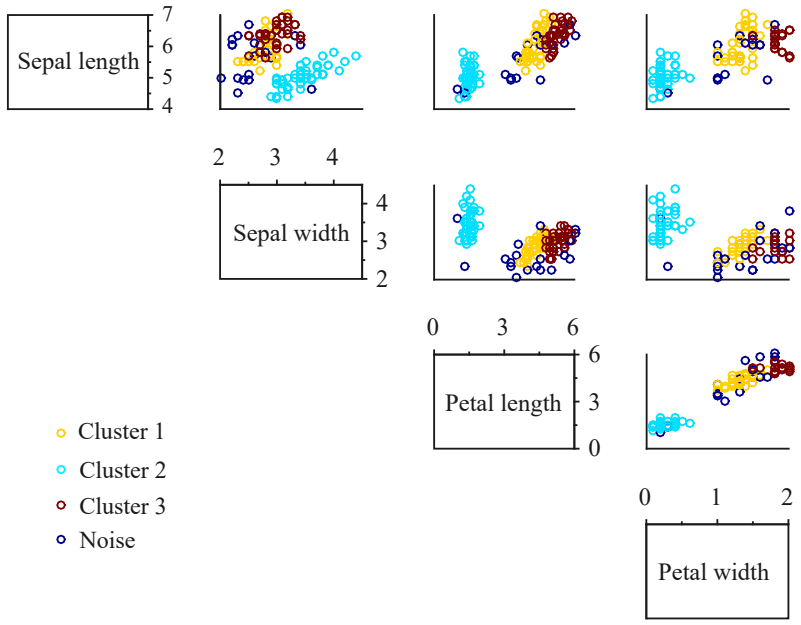


Figure 6.11: Best possible partitioning of the Fisher Iris data that can be achieved with DBSCAN as published by [Ester 1996] and setting the threshold for partitioning the data to 4-dist=0.42. It is achieved by gradually varying the 4-dist threshold between 0.1 and 1.1 and choosing the results corresponding to the maximal RI value.

but splitting the other two groups into a singleton cluster (one Virginica species) and a group containing the rest of the Virginica points and the Versicolor group. In this case, a Rand index of 0.78 is attained. On the other hand, complete linkage hierarchical clustering exhibits a Rand index of 0.84 and is capable of detecting that there are three clusters in the sample analyzed (Fig. 6.15(b)).

In the case of K-means, the analysis of the Fisher Iris data results in two different partitionings of the sample as shown in Fig. 6.16(a) and Fig. 6.16(b). This is an effect of the random initialization of the clusters' centroids (cf. Sec. 6.1). For the current algorithm performance comparison, the results displayed in Fig. 6.16(b), corresponding to the higher Rand index of $RI = 0.84$, are further considered.

The obtained Rand indices for all clustering techniques tested on the Fisher Iris data are summarized and displayed in Table 6.5.

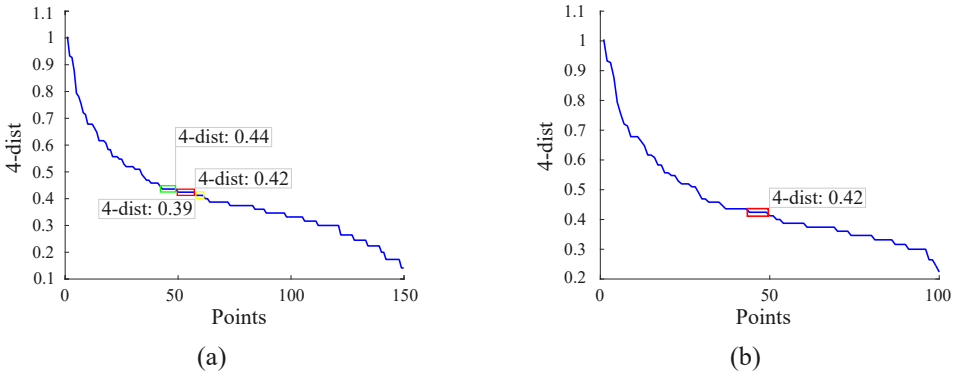


Figure 6.12: Sorted 4-dist graphs for (a) the complete Fisher Iris data set and (b) for the overlapping Setosa and Versicolor classes. In both cases, no distinct elbow of the function is observed that can be employed for choosing the *Eps* parameter for DBSCAN.

Table 6.5: Fisher Iris Rand indices for NPC and four commonly employed clustering methods.

Algorithm	Rand index	Number of clusters correct
NPC, one-step	0.7685	no
NPC, two-step	0.8446	yes
DBSCAN (NP)	0.7773	no
DBSCAN (NP), 2steps	0.7399	no
Hierarchical clustering, single linkage	0.7766	no
Hierarchical clustering, complete linkage	0.8368	yes
K-means ($K = 3$)	0.8368	no/yes

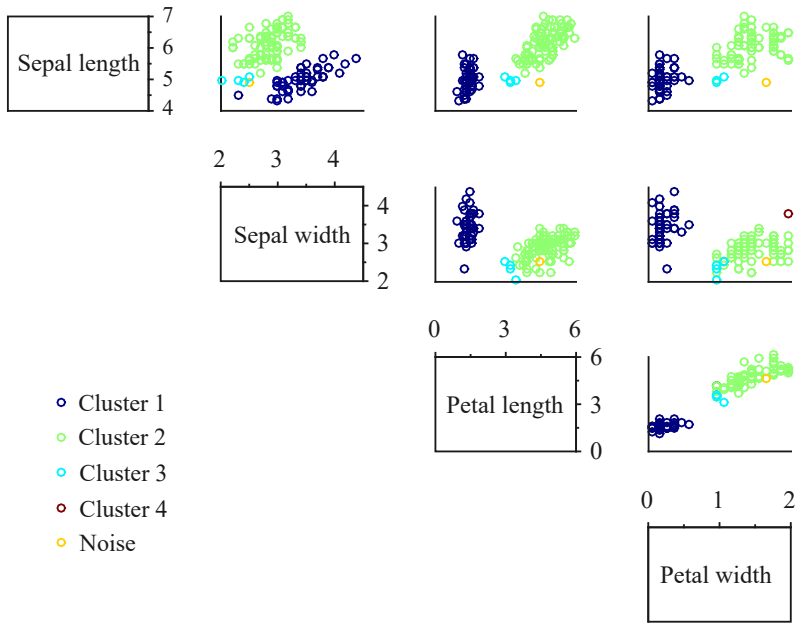


Figure 6.13: Fisher Iris data processed with NP DBSCAN.

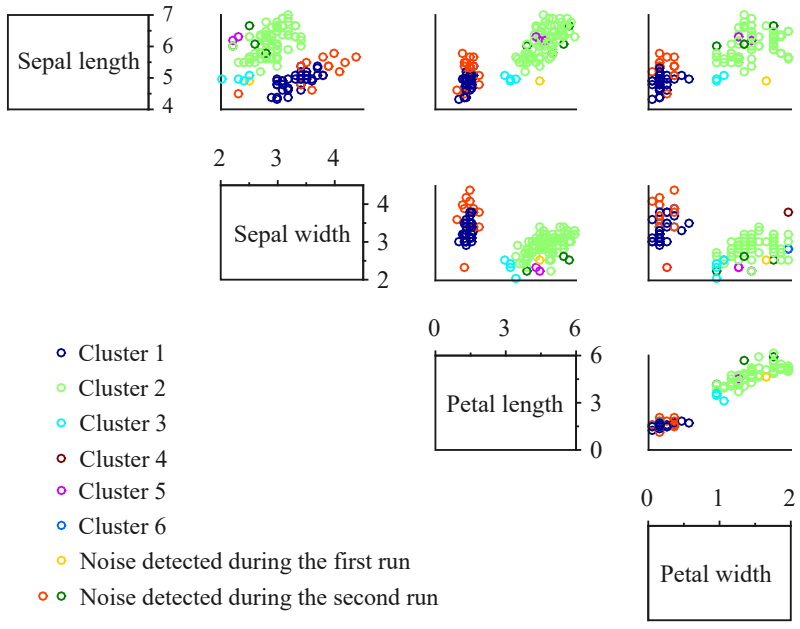
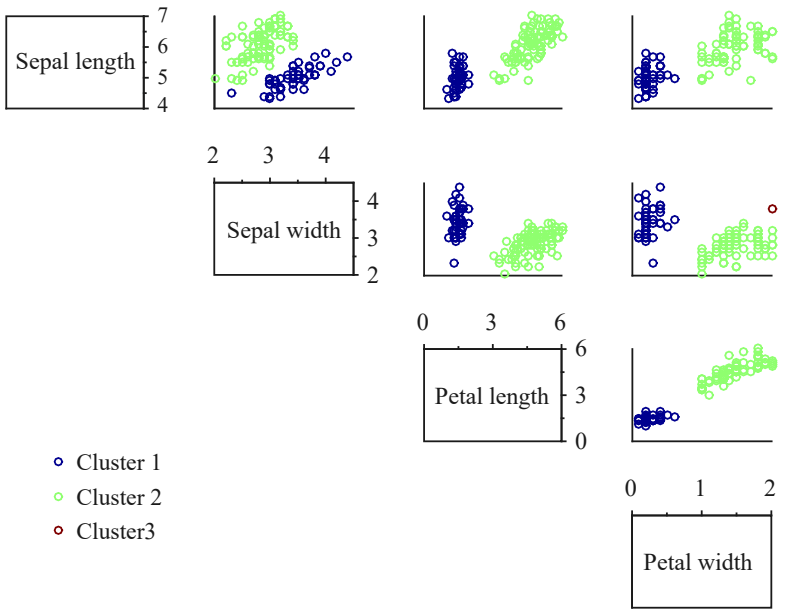
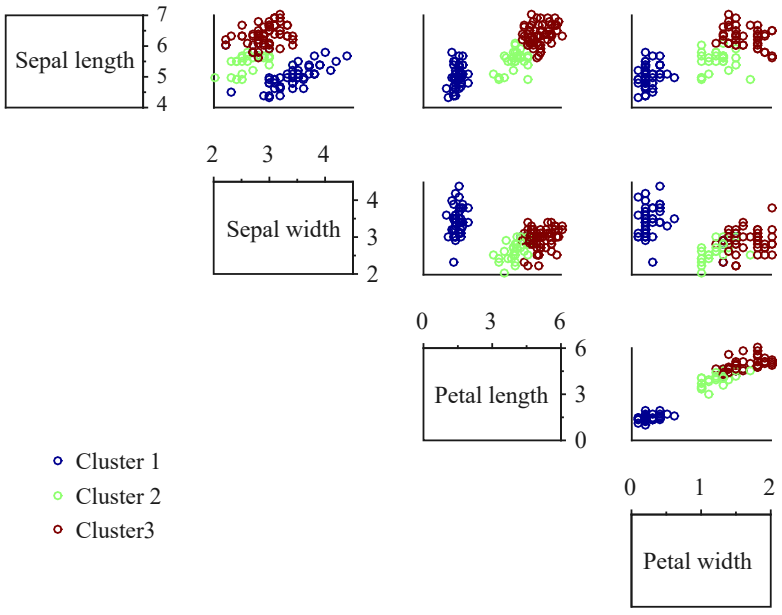


Figure 6.14: Fisher Iris data processed with NP DBSCAN, where an additional search for subgroups is conducted in a similar way as in the case of two-step NPC. Here, the classes Virginica and Versicolor are clustered together even when running the algorithm repeatedly after excluding the class Setosa.

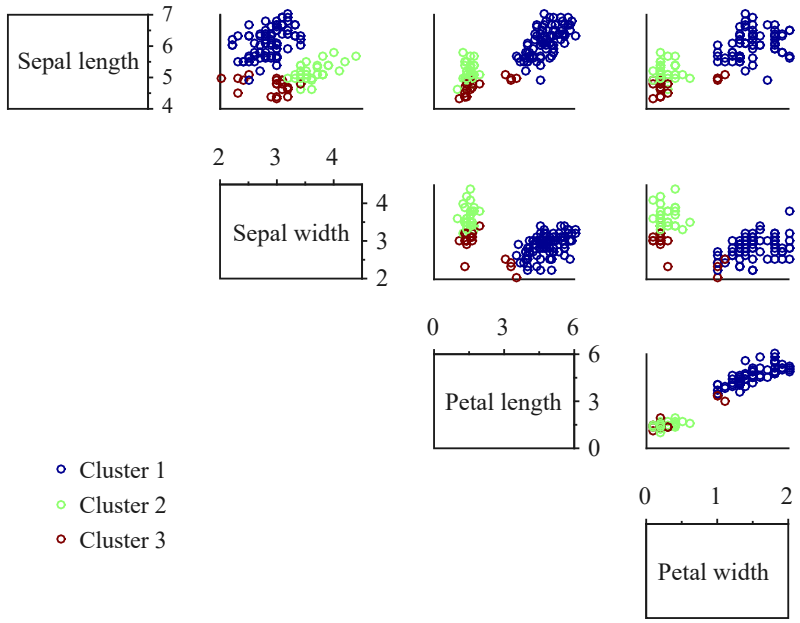


(a)

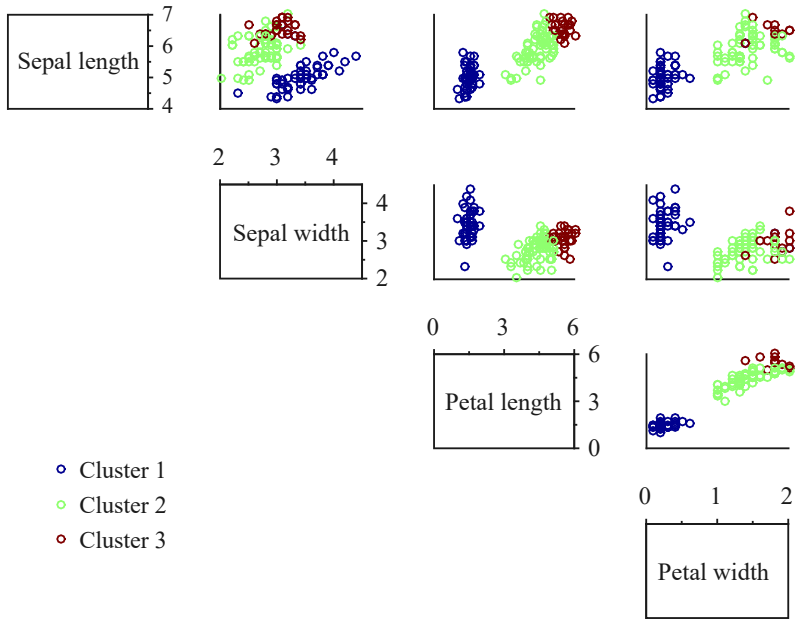


(b)

Figure 6.15: Fisher Iris data processed with hierarchical clustering, where the hierarchy is cut so as to produce three groups either with (a) single linkage or (b) complete linkage.



(a)



(b)

Figure 6.16: Fisher Iris data processed with K-means, where the number of groups is prespecified to $K = 3$. Two different results, displayed in (a) and (b), are observed due to the random initialization of the clusters' centroids.

Chapter 7

Multiple Sclerosis Image Data Analysis

Contents

The analysis presented in the current chapter rehandles the problem that has initially inspired the development of NPC: the clinico-radiological paradox in multiple sclerosis and the attempt to address it by means of multimodal quantitative MRI. The acquisition protocol employed has been optimized for the rapid high-resolution mapping of the parameters T_1, T_2^* and absolute water content, H_2O^{total} [Neeb 2008]. As previously described in Chapter 2, in a first step of the current translational research work, it is successfully extended by the full-brain measurement of myelin water content, H_2O^{myelin}. With this expansion, a single individual's scan allows for the reconstruction of four different quantitative maps depicting the spacial distribution of parameters sensitive to inflammatory processes as well as tissue myelination. Thus, in the context of a chronic inflammatory demyelinating disease such as multiple sclerosis, the quantitative nature of this imaging data set is expected to be particularly advantageous. As the individual's lesion load observed in standard MRI images does not necessarily correspond to the actual clinical findings, quantitative information characterizing the *normal appearing brain tissue* can be analyzed to additionally characterize the disease state of an individual subject. Generally, the ultimate goals in this context are to improve the personalized diagnosis and therapy, thus delaying disease progression. Advances considering the following two aspects are of particular importance here:

- expedite and facilitate an early diagnosis since MS tends to lead to more damage in the first year than in later stages of the disease [Kuhlmann 2002];

- attain a medical stratification and enable detection of patients who are at risk to develop a MS relapse and/or steady disability progression.

The first aspect has been successfully addressed by [Neeb 2014] who investigate whether quantitative MRI of normal appearing brain matter can be employed to automatically differentiate between MS patients and healthy subjects. They present a straightforward and robust dimensionality-reduction scheme in combination with supervised learning models. The results here clearly demonstrate that the normal appearing neural tissue *is* experiencing MS-induced changes and that the employed MRI technique is an examination that allows for a quantitative and objective assessment of such subtle pathological abnormalities.

The avenue pursued in the current chapter addresses the second aspect stated above. It aims to study whether there is some natural tendency of patient grouping that corresponds to the individual disease progression (i.e. impairment as measured by the clinical EDSS score) when solely quantitative MRI of normal appearing brain tissue is employed. Such disease stratification could enable the indication of a potential exacerbation in cases where no new lesions or no MS relapse is yet observed. Advances in this area could be especially advantageous for the disease monitoring since the genesis of brain pathology typically predates the onset of clinical symptoms so that earlier and putatively more effective treatments can be offered to patients.

Whether quantitative MRI data could be employed to attain such a stratification or not is the last question posed within the current translational research work.

7.1 Data Acquisition and Preprocessing

For the current analysis, 18 male and 24 female MS patients aged between 18 and 77 years (mean age 40.52 ± 11.34 years) are scanned on a standard clinical 3T MRI scanner (TRIO, Siemens AG, Erlangen) at the Radiological Institute Hohenzollernstrasse Koblenz. Informed written consent is obtained from each subject prior to examination. The individuals' expanded disability status scores at the time of MRI scan range between EDSS = 0 and EDSS = 6.5. Figure 7.3(a) shows the frequency of the different EDSS values within the studied cohort. For each subject, full brain quantitative T_1, T_2^*, H_2O^{total} and H_2O^{myelin} maps are acquired as described in detail in Chapter 2.

All constructed parameter maps are segmented using the following thresholds: voxels having $T_1 \in [500, 900]$ ms and $H_2O^{total} \in [60, 85]$ % are assigned to the white matter (WM) segment and voxels having $T_1 \in [901, 1300]$ ms and $H_2O^{total} \in [70, 93]$ % to the gray matter (GM) segment. The rest is considered as non-brain tissue and is not relevant for the further analysis.

7.2 Feature Extraction

To extract potentially discriminative features, the following procedure is followed:

- For each of the four quantitative MR parameters, H_2O^{total}, H_2O^{myelin}, T_1 and T_2^*, the mean and median values as well as the standard deviations in gray and white matter are determined. These characteristic values are calculated for each transversal slice as well as for the whole three-dimensional WM and GM segments.

- To assess brain atrophy, which is a clinically relevant component of MS progression, the number of voxels within the white and gray matter segments is determined. Again, this is calculated for each transversal slice as well as for the whole three-dimensional WM and GM segments.

- As previously mentioned, the attributes derived based on the employed quantitative MR protocol are biophysically correlated. The next set of features derived are based on the study of [Hilkert 2010] who has demonstrated that the correlation angles between the quantitative parameters assessed are indicative of MS pathologies. Thus, for the following analysis the average correlation angles between T_1 and T_2^* are calculated along with the standard deviations as described in [Hilkert 2010] for each transversal slice as well as for the whole three-dimensional WM and GM segments.

- Information concerning the distributions of T_1 and T_2^* is included in the form of histogram counts. For each scanned slice, a histogram of the longitudinal relaxation times within the interval $T_1 \in [500, 1300]$ ms is generated and the observed frequencies are employed as features. The same procedure is conducted for the transverse relaxation times within the range $T_2^* \in [35, 130]$ ms.

- Finally, the average values and the standard deviations of the fast and the slow transverse relaxing time, $T_2^{*,f}$ and $T_2^{*,s}$, respectively, are considered. These four values are calculated for the three-dimensional WM and GM segments.

This procedure results in a total number of 5287 features. A full list of the attributes derived can be found in Appendix E (VariableNames_minContrast_0.txt). As the number of individuals examined is significantly lower ($N = 42$) than the number of features, analyzing the whole sample is not expected to lead to meaningful results, an effect commonly known as the curse of dimensionality. To select a subset of parameters that are more distinguishing with respect to the individual MS-induced disability, the contrast between the group of patients G_1 having an $EDSS > 3$ and the group of patients G_2 having an $EDSS \leq 3$ is calculates for each attribute as follows:

$$K_G = \frac{|\mu(G_1) - \mu(G_2)|}{\sqrt{\sigma(G_1) - \sigma(G_2)}} \tag{7.1}$$

Selecting only features which exhibit $K_G \geq 0.4$ results in a subset containing 583 attributes. A list of the features included here can be found in Appendix E (file VariableNames_minContrast_0_4.txt). In comparison, choosing $K_G \geq 0.2$ results in a

Table 7.1: Summary of the MS data sets analyzed and their specifications.

Data set	Number features	Contrast threshold	Variance explained by the first 15 PCs	Variance explained by the first three PCs
1	583	$K_G \geq 0.4$	99.75%	89.88%
2	2259	$K_G \geq 0.2$	99.97%	99.74%
3	5287	$K_G = 0$	99.97%	99.77%

selection of 2259 features (Appendix E, file VariableNames_minContrast_0_2.txt). These two data sets along with the full sample including all 5287 attributes are the basis for the following analysis. Table 7.1 summarizes the specifications of the three samples.

7.3 Data Analysis

For visualization purposes as well as additional explorative investigations, a principal component analysis of each explored data set is conducted. The space spanned by the fist three principal components (PC) is employed to graphically display the clustering results obtained as well as the phenotypic data which enables a visual comparison of both.

The data is clustered with the NPC default parameter set (cf. Table 5.6) and an additional investigation of the groups found by employing a two-step NPC without noise reassignment as described in Sec. 4.2.6.

For a comparison, the same procedure is repeated for all three data sets generated (c.f. Table 7.1) as well as for the first 15 principal components of each sample (details concerning the total variance explained in each case are given in Table 7.1, as well). Furthermore, the 583-dimensional subset is clustered with DBSCAN by employing different values for *Eps* as discussed in Sec. 6.1.

7.4 Results

7.4.1 Data extracted with $K_G \geq 0.4$

The results obtained by clustering the 583-dimensional data with NPC are summarized in Table 7.2 and graphically displayed in Fig. 7.2.

Two cohesive clusters are identified during the first analysis step: the first one is comprised of two patients with no neurological deficits and mean age of 36.00 ± 8.49 years, the second one - of six patients with mean EDSS of 0.60 ± 0.89 and mean age of 33.00 ± 9.45 years. The rest of the patients builds the group of outliers with a mean EDSS of 1.94 ± 2.14 and a mean age of 42.12 ± 11.46 years. The second NPC step detects no subgroups within the found cohesive groups (denoted as Cluster 1 and Cluster 2 in Table 7.2 and Fig. 7.2). However, exploring additionally the noise group detects one additional cohesive group consisting of two individuals with no

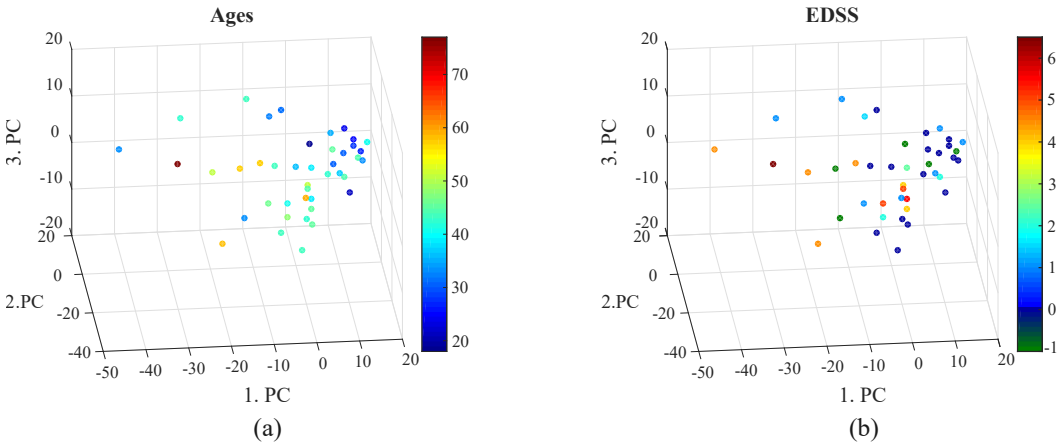


Figure 7.1: Patients phenotypes depicted in a space spanned by the first three principal components of the data set generated with a minimum contrast threshold of $K_G \geq 0.4$: (a) Ages and (b) EDSS assessed at the time of examination. EDSS $= -1$ indicates an unknown value.

Table 7.2: Characteristic values for the patient (sub)groups found with two-step NPC applied on a dataset including 585 features.

NPC step	Cluster	Number elements	μ_{EDSS}	σ_{EDSS}	μ_{age}	σ_{age}
1	1	2	0.00	0.00	36.00	8.49
	2	6	0.60	0.89	33.00	9.45
	3 (noise)	34	1.94	2.14	42.12	11.46
2	3.1	2	0.00	0.00	39.50	4.95
	3.2 (noise)	32	2.07	2.15	42.28	11.77

neurological deficits and mean age of 39.50 ± 4.95 years (denoted as Cluster 3.1 in Table 7.2 and as Cluster 3.1 (step 2) in Fig. 7.2). The rest of the patients is recognized again as outliers.

Concerning the temperature curves in both NPC analysis steps, it is noticeable that they are very similar. Both exhibit a rapid increase of $T(t)$ at early time points and a long cooling-down phase. Thus, the algorithm terminates in both cases based on the additional breaking condition discussed in Sec. 6.2, and provides an indication stating that the groups found are presumably of (vastly) different densities. Observing the temporal energy evolution of the points (as discussed in Sec. 5.2) reveals that the systems' kinetic energy evolution is primarily determined by the mutual repulsion of the noise points which indeed constitute the majority of observations in both cases.

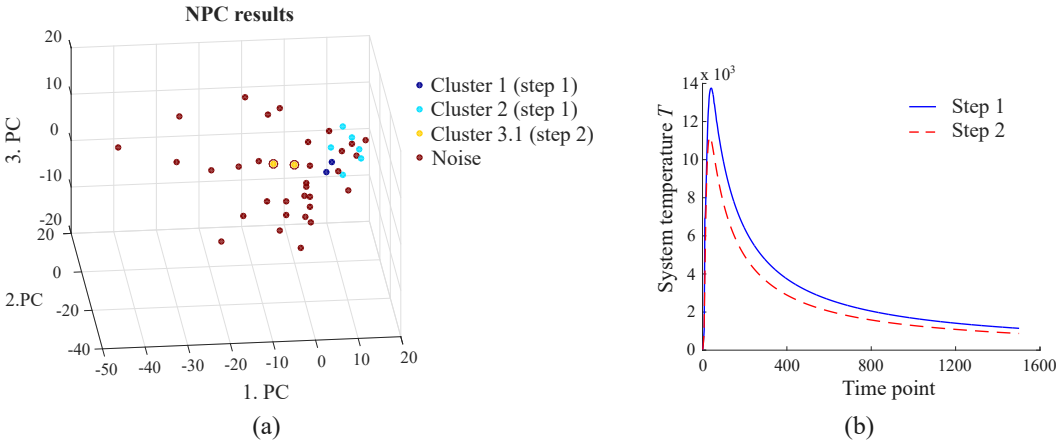


Figure 7.2: NPC results for the dataset generated with $K_G \geq 0.4$ (585 features) as described in Sec. 7.2. (a) Scatter plot of the results depicted in a space spanned by the first three principal components of the data subset. During the fist NPC step the points belonging to Cluster 3.1 are recognized as noise (labeled in red) and during the second NPC step - as members of a cohesive cluster (additionally labeled in yellow). (b) System temperature evolution during the first (blue) and the second (red) NPC step.

Taking a closer look at the results displayed in Fig. 7.2(a) and comparing them to the phenotypic data (Fig. 7.1) reveals that the region in feature space in which the cohesive groups detected are located correspond to a region of younger age and lower EDSS. To explore this observation more closely, the distances w.r.t. the centroid of the group with EDSS $= 0.00 \pm 0.00$ and age of 36.00 ± 8.49 years (turquoise points in Fig. 7.2(a)) are studied. Figure 7.3(b) displays a box plot of these distances for three groups of patients: individuals with EDSS $= 0$, with $1 \leq$ EDSS ≤ 3 and with EDSS > 3 (note that unknown EDSS scores denoted with EDSS $= -1$ are not included in this consideration). It reveals the tendency that highly impaired cases are on average located further away in feature space from the low-EDSS region. It can be also seen that the median values of the individual distances tend to increase with higher EDSS but the variability within the groups increases as well.

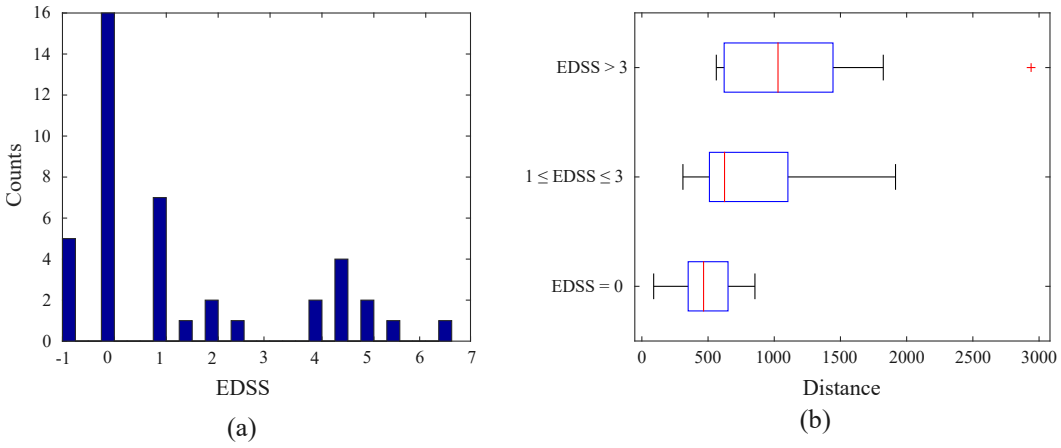


Figure 7.3: (a) Frequency distribution of the EDSS values within the studied cohort. (b) Box plot of the distance w.r.t. the centroid of the group with EDSS = 0 (depicted in turquoise in Fig. 7.2(a)) detected by clustering data generated with $K_G \geq 0.4$ (585 features). Here, three groups of patients are considered: patients with EDSS = 0, with $1 \leq \text{EDSS} \leq 3$ and with EDSS > 3.

Conducting the same analysis with the first 15 principal components of the dataset generated with $K_G \geq 0.4$ (explaining 99.75% of the total variance) leads to very similar results. As seen in Table 7.3 and Fig. 7.4 in this case one additional cohesive cluster comprised of only two points is detected during the first NPC step (mean EDSS $= 2.00 \pm 0.00$, mean age 37.00 ± 9.90 years). The remaining points are again grouped in two low-EDSS clusters and noise. During the second NPC step, only one subgroup is detected containing the same two patients already recognized as a cohesive group when using the full $K_G \geq 0.4$ dataset instead of its 15 first principal components (see Cluster 3.1 in Table 7.2 and Cluster 4.1 in Table 7.3). The rest is again recognized as noise.

Table 7.3: Characteristic values for the patients' (sub)groups found with NPC applied on a dataset including the first 15 principal components of the data selected with $K_G \geq 0.4$.

NPC step	Cluster	Number elements	μ_{EDSS}	σ_{EDSS}	μ_{age}	σ_{age}
1	1	2	2.00	0.00	37.00	9.90
	2	2	0.00	0.00	34.50	10.61
	3	6	0.17	0.41	32.83	9.99
	4 (noise)	32	2.07	2.15	42.56	11.39
2	4.1	2	0.00	0.00	39.50	4.95
	4.2 (noise)	30	2.22	2.15	42.77	11.71

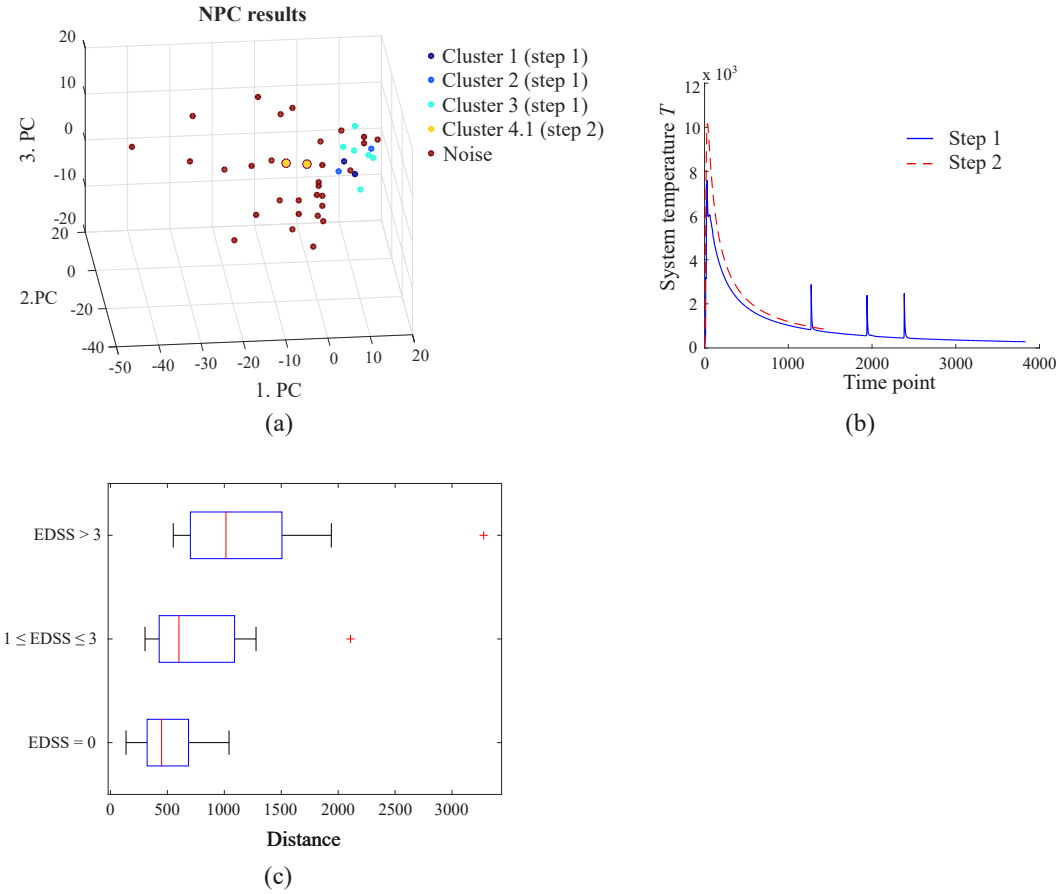


Figure 7.4: NPC results for a dataset generated with $K_G \geq 0.4$. Here the first 15 principal components are employed as features. (a) Scatter plot of the results depicted in a space spanned by the first three principal components. During the fist NPC step the points belonging to Cluster 4.1 are recognized as noise (labeled in red) and during the second NPC step - as members of a cohesive cluster (additionally labeled in yellow). (b) System temperature evolution during the first (blue) and the second (red) NPC step. (c) Box plot of the distance w.r.t. the centroid of the group with $\mu_{EDSS} = 0$ (light blue in panel (a)). Here, three groups of patients are considered: patients with EDSS $= 0$, with $1 \leq$ EDSS ≤ 3 and with EDSS > 3.

Table 7.4: Characteristic values for the (sub)groups found with NPC based on a sample containing 2259 features exhibiting a minimal contrast of 0.2 between the patients' groups with $EDSS > 3$ and $EDSS \leq 3$.

NPC step	Cluster	Number elements	μ_{EDSS}	σ_{EDSS}	μ_{age}	σ_{age}
1	1	2	2.00	0.00	37.00	9.90
	2	40	1.69	2.08	40.70	11.54
2	2.1	2	0.50	0.71	24.50	0.71
	2.2 (noise)	38	1.77	2.12	41.55	11.20

7.4.2 Data extracted with $K_G \geq 0.2$

In exploring the MS data, an additional analysis is conducted with the sample generated by employing a contrast threshold of $K \geq 0.2$, resulting in 2259 features. Here again, the complete data set is clustered, as well as its first 15 principal components, explaining 99.97% of the total variance. As stated in Sec. 7.3, the results and the phenotypic data are displayed in the space spanned by the first three principal components (Fig. 7.5).

The results from clustering the complete data set are listed in Table 7.4 and graphically displayed in Fig. 7.5(c). In this case, only two cohesive groups each comprised of two points are detected: one during the first NPC step with mean EDSS $= 2.00 \pm 0.00$ and mean age 37.00 ± 9.90 years, as well as a second one during the second NPC step with lower mean EDSS $= 0.50 \pm 0.71$ and of younger age of 24.00 ± 0.71 years.

A box plot of the distances toward the centroid of the low-EDSS group (green in Fig. 7.5(c)) for the same three groups of patients considered previously in Sec. 7.4.1 is displayed in Fig. 7.5(e). Though the median values still reveal a tendency to shift toward higher distanced with increasing EDSS, including features less sensitive to the impairment also leads to higher variability within the groups when compared with the previous results (see Sec. 7.4.1).

Concerning the temporal temperature evolution in both NPC analysis steps, it is noticeable that they are almost identical, exhibiting a rapid increase of $T(t)$ at early time points and a long cooling-down phase. This is due to the fact that in both cases only two nucleons are exploring mutual attraction and the systems' kinetic energy evolution is primarily determined by the mutual repulsion of the noise points which constitute the majority of observations. Here again, the algorithm terminates in both NPC steps based on the additional breaking condition discussed in Sec. 6.2, and provides an indication stating that the groups found are presumably of (vastly) different densities.

For comparison, the first 15 principal components of the data set are also clustered. As shown in Table 7.5 and graphically displayed in Fig. 7.6, in this case five clusters are detected during the first NPC step, all but one with $\mu_{EDSS} < 1$. In the

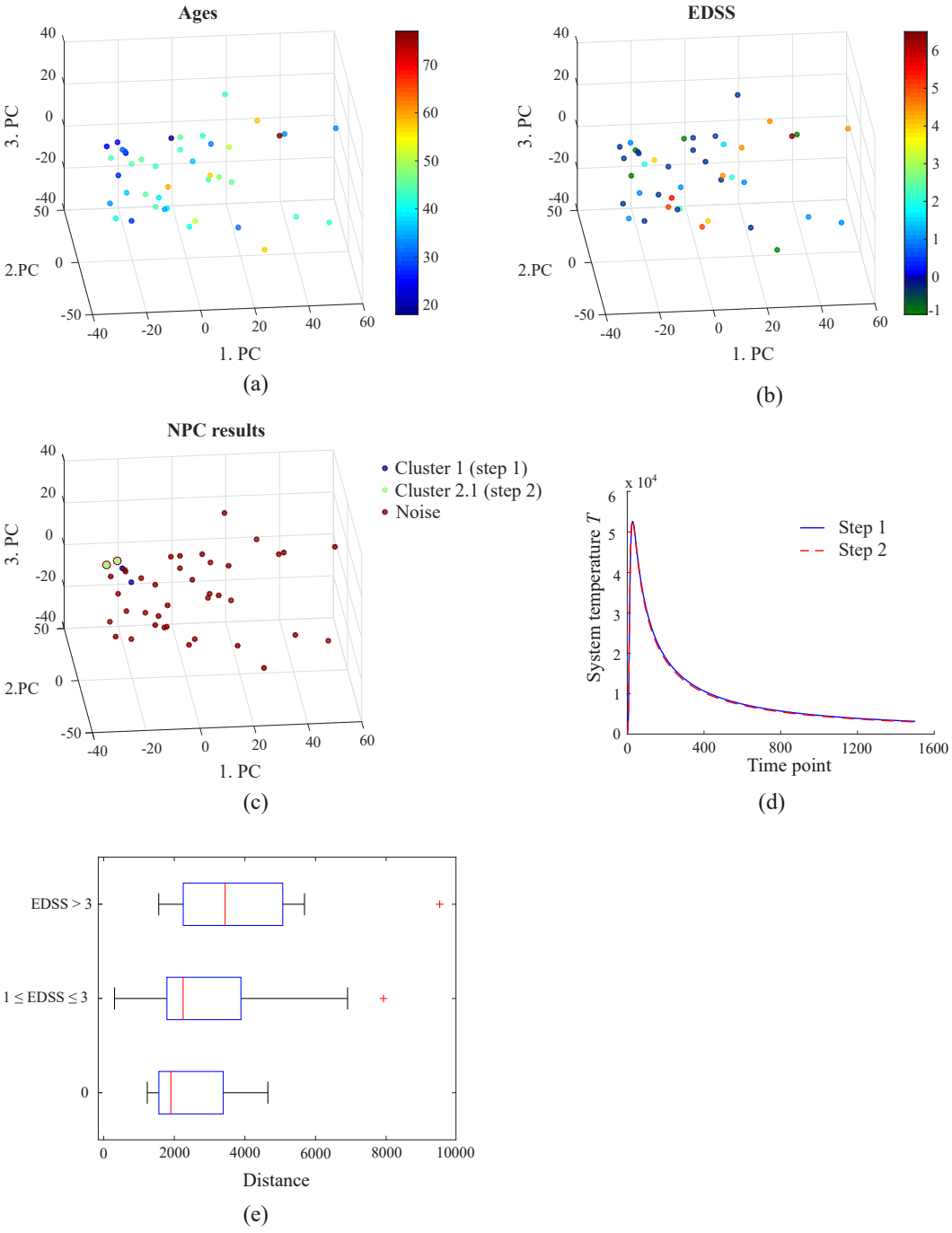


Figure 7.5: NPC results for the dataset generated with $K_G \geq 0.2$ (2259 features) as described in Sec. 7.2. The obtained clusters as well as the patients' phenotypes are depicted in a space spanned by the first three principal components of the data subset: (a) Ages and (b) EDSS assessed at the time of examination. (c) NPC results. During the fist NPC step the points belonging to Cluster 2.1 are recognized as noise (labeled in red) and during the second NPC step - as members of a cohesive cluster (additionally labeled in green).
(d) System temperature evolution during the first (blue) and the second (red) NPC steps.
(e) Box plot of the observation distances toward the centroid of Cluster 2.1 (green points in panel (c), EDSS $= 0.50 \pm 0.71$, age of 24.00 ± 0.71 years) for three groups: patients with EDSS $= 0$, with $1 \leq$ EDSS ≤ 3 and with EDSS > 3.

second NPC step all of the points have a positive final potential energy leading to the detection of no additional subgroups. The temperature evolutions are similar to the previous experiments and the NPC provides a notification that the clusters found in the first NPC step might be of different densities. Also, the distances towards the centroid of the $EDSS = 0$ points detected (singleton clusters 2 and 3 as listed in Table 7.5) are calculated and a box plot of their distribution for the previously considered patients groups is presented in Fig. 7.6(c). Here, the same tendency is observed as in the case of clustering the whole dataset including all 2259 features.

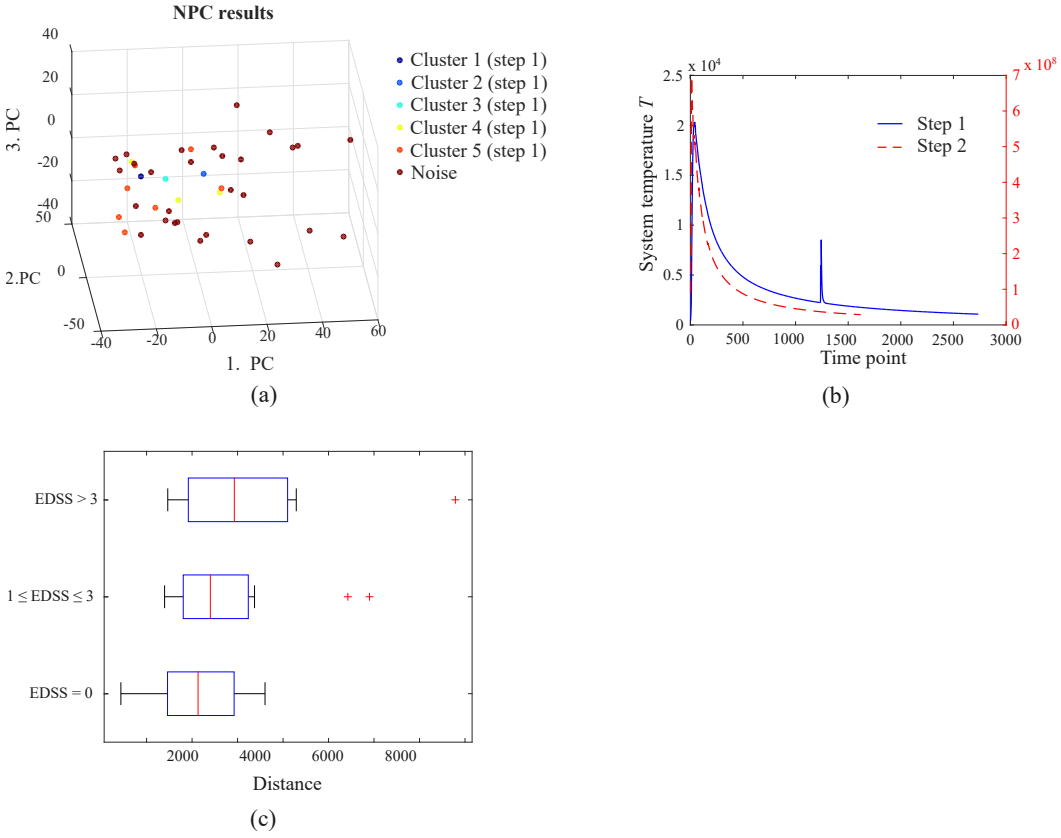


Figure 7.6: NPC results for a dataset generated with $K_G \geq 0.2$. Here the first 15 principal components are employed as features. (a) Scatter plot of the results depicted in a space spanned by the first three principal components. Cohesive groups are detected only during the first NPC step, no subgroups are detected during the second NPC step. (b) System temperature evolution during the first (blue) and the second (red) NPC steps. (c) Box plot of the observation distances toward the centroid of the points with $\mu_{EDSS} = 0$ (turquoise and light blue points in panel (a)) for three groups: patients with EDSS = 0, with $1 \leq$ EDSS ≤ 3 and with EDSS > 3.

Table 7.5: Characteristic values for the patient groups found with NPC. Here, the first 15 principal components of a sample generated with $K_G \geq 0.2$ as described in Sec. 7.2 are employed as features. Cohesive groups are detected only during the first NPC step, no subgroups are detected during the second NPC step.

Cluster	Number elements	μ_{EDSS}	σ_{EDSS}	μ_{age}	σ_{age}
1	1	2.00	0.00	44.00	0.00
2	1	0.00	0.00	36.00	0.00
3	1	0.00	0.00	42.00	0.00
4	3	0.50	0.71	44.00	14.53
5	7	0.92	1.80	38.43	12.26
6 (noise)	29	2.10	2.18	40.66	11.79

7.4.3 Full sample

Last but not least, clustering the full sample including all 5287 features does not lead to any grouping. Repeating the analysis with the first 15 principal components of this data results in the grouping presented in Table 7.6. Interestingly, here five clusters are detected that correspond to groups characterized by different age and degree of disability:

- a singleton cluster with zero EDSS and age of 36 years,
- a second cluster comprised of four patients with low EDSS of 0.67 ± 0.58 and higher age of 43.75 ± 11.98 years,
- a third group including nine patients with intermediate EDSS of 1.31 ± 1.95 and higher age of 40.44 ± 9.54 years,
- a fourth cluster including three patients with high EDSS of 2.67 ± 1.75 years and younger age of 33.00 ± 10.39 years,
- a fifth group including six patients with a low to intermediate EDSS of 1.67 ± 1.75 and of younger age
- and finally a sixth (noise) group including patients of higher disability, $EDSS = 2.03 \pm 2.34$, and higher age of 42.32 ± 12.97 years.

No additional subgroups are detected when employing a two-step NPC approach. The obtained results are graphically displayed in Fig. 7.7(c).

As in the previous cases, the distances towards the centroid of the cluster with the lowest average age and $EDSS$ (cluster 2 in Table 7.6 displayed in light blue in Fig. 7.6(c)) are calculated and a box plot of their distribution for the different EDSS groups is presented in Fig. 7.7(e). Here, a similar tendency is observed as in the case of clustering the dataset including 2259 features.

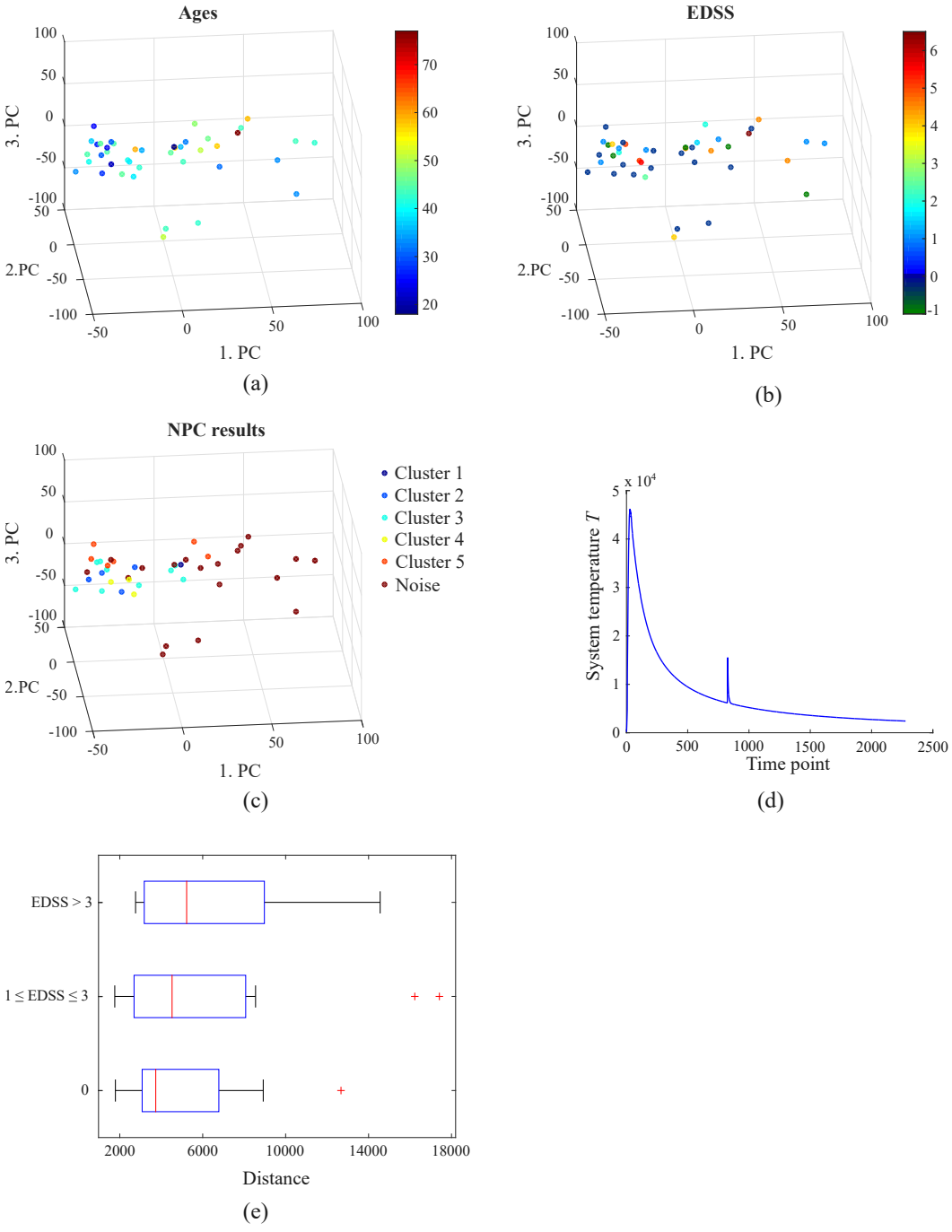


Figure 7.7: NPC results for the full sample including all 5287 features. Here the first 15 principal components are employed as features. The obtained clusters as well as the patients' phenotypes are depicted in a space spanned by the first three principal components of the data subset: (a) Ages, (b) EDSS assessed at the time of examination and (c) NPC results. (d) System temperature evolution. (e) Box plot of the observation distances toward the centroid of Cluster 2 (light blue points in panel (c)) for three groups: patients with EDSS $= 0$, with $1 \leq$ EDSS ≤ 3 and with EDSS > 3.

Table 7.6: Characteristic values for the patient groups found with NPC. Here, the first 15 principal components of the complete 5287-dimensional data are employed as features (explaining 99.97% of the total variance).

Cluster	Number elements	μ_{EDSS}	σ_{EDSS}	μ_{age}	σ_{age}
1	1	0.00	0.00	36.00	0.00
2	4	0.67	0.58	43.75	11.98
3	9	1.31	1.95	40.44	9.54
4	3	2.67	2.75	33.00	10.39
5	6	1.67	1.75	37.33	10.62
6 (noise)	19	2.03	2.34	42.32	12.97

7.4.4 Comparison with DBSCAN

Finally, the NPC results obtained for the 585-dimensional data are compared with DBSCAN as proposed by [Ester 1996]. Two parameters need to be specified here: $NumPts$ and Eps (cf. Sec. 6.1). The first is set to $NumPts = 4$ as recommended in [Ester 1996]. Considering the second parameter, Fig. 7.8(a) shows that the sorted 4-dist curve does not exhibit a single distinct but rather multiple "knees" that can be employed to manually select the Eps threshold. The different values considered here are marked in color in Fig. 7.8(a) and the corresponding results obtained with DBSCAN are displayed in the panels Fig. 7.8(b)-(f). Clustering the data with $Eps = 5180$ leads to no detection of any group. In all other cases ($Eps = 4250, 3309, 2907, 2575$) the sample is split into one cluster and noise. Reducing Eps results in increasing the number of noise points detected. Comparing the results with the patients phenotypes displayed in Fig. 7.1 reveals a similarity to the NPC results, as the noise points detected correspond to patients of higher age and EDSS. The characteristic phenotypic values for the groups found with the different Eps thresholds are summarized in Table 7.7.

Table 7.7: Characteristic values for the patients groups found when clustering 585-dimensional data with DBSCAN and employing different Eps values.

	Cluster			Noise		
Eps	Number elements	$EDSS$	Age (years)	Number elements	$EDSS$	$Age(years)$
2575	25	1.82 ± 1.99	39.80 ± 9.23	17	1.53 ± 2.20	41.59 ± 14.23
2907	27	1.67 ± 1.97	39.00 ± 9.33	15	1.77 ± 2.28	43.27 ± 14.34
3309	32	1.59 ± 1.98	38.69 ± 10.34	10	2.06 ± 2.36	46.40 ± 13.13
4250	39	1.41 ± 1.85	38.87 ± 9.55	3	5.00 ± 1.32	62.00 ± 13.45

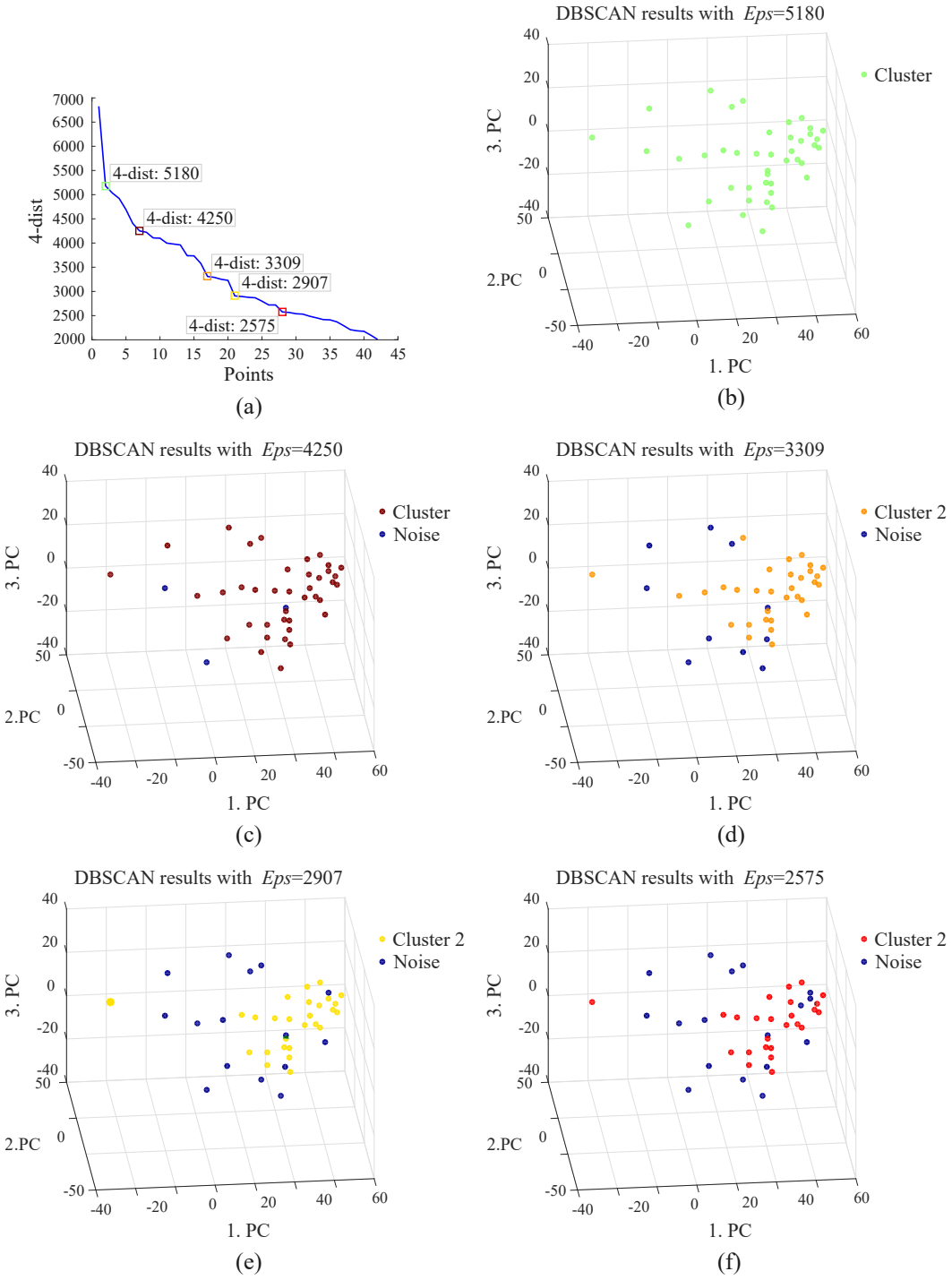


Figure 7.8: Clustering results obtained with DBSCAN applied on a dataset including 585 features (cf. Table 5.6). (a) Sorted 4-dist curve, where the different "knee"-values employed are marked by colored squares. The value marked in green, dark red, orange, yellow and light red result in the clusters presented in panel (b), (c), (d), (e) and (f), respectively.

7.5 Discussion

The experiments conducted reveal that there is a consistent general grouping tendency of the data points within the MS sample: NPC detects as a cohesive cluster a fraction representing the low disease grade patients of younger age, while the patients of higher age and EDSS are recognized as outliers. As expected, this tendency becomes increasingly prominent when employing features more sensitive to the individual EDSS: In the case of the data selected with $K_G \geq 0.4$ all cluster constitute only of patients with $EDSS < 1$ and very similar results are obtained when processing the first 15 principal components of this data due to the high percentage of explained variance in this case. In comparison, employing $K_G \geq 0.2$ results in approx. a four-fold increase in the number of features and leads to the detection of only one group with $EDSS < 1$ when clustering the complete data (or the detection of multiple smaller groups when considering the first 15 principal components of this sample). Yet, the general tendency of detecting low-EDSS patients as a cluster constituent is still present as four of the five groups found during the first NPC step in this case have $EDSS < 1$.

Concerning the data set including all 5287 derived attributes where no EDSS-based feature selection takes place prior to processing, no meaningful splitting of points is obtained when analyzing the complete dataset. The reason for this observation is that choosing the threshold $K = 0$ includes all attributes, regardless of their pathological relevance. Including a large number of features, many of which possibly insensitive to the patient state, leads to an undue disproportionality between the sample dimensionality and number of observations, $d \gg N$. In such a constellation, commonly referred to as "the curse of dimensionality", the mutual distance between observations representing their similarity increases unduly so that the data becomes sparse and no patterns can be detected. Indeed, we have $d \gg N$ in the samples created with $K \neq 0$, as well. However, selecting attributes sensitive (to a different extent) to the patients' EDSS allows NPC to detect similarities with respect to the problem of interest.

Interestingly, in contrast to the preselected datasets, clustering the first 15 principal components of the sample including all 5287 attributes results in the recognition of groups characterized by different age and degree of disability. This suggests that features included in the full sample, may be distinctive to disease characteristics other than the overall physical impairment as assessed by the EDSS, which forms the bases for the samples obtained with $K > 0$.

Taking a closer look at the results obtained for the 585-dimensional data (Fig. 7.2) and comparing them to the phenotypic information (Fig. 7.1), reveals that the similarity within the group of less affected patients (represented by the distance measurement in feature space) can be employed to identify a region in feature space characteristic for patients with normal neurological findings. This idea is also behind the examination of the individuals' distances towards the centroid of the points corresponding to this region in feature space. Indeed, even though not every possible EDSS value is included in the patient sample studied here (see Fig. 7.3(a)), consid-

ering the three groups of patients with EDSS = 0, $1 \leq$ EDSS ≤ 3 and EDSS > 3 confirms that highly impaired cases tend to be located further away in feature space from the low-EDSS area. Furthermore, it can also be seen that along with the increase of the median distance with higher EDSS, the variability within the groups increases as well. This tendencies can be observed in all samples studied but are particularly pronounced in the $K_G \geq 0.4$ dataset. Both effects are not unexpected and demonstrate that patients experiencing no disease-induced pathologies are more similar with respect to the employed quantitative features on one side, whereas patients of higher age, experiencing a longer disease course, different therapies and possibly additional age-related comorbidities are not. Based on this observation, it can be presumed that studying a larger cohort including observations of all EDSS (and especially a greater number of higher EDSS values) could lead to a more specific disease stratification. Note also that here solely absolute distances are considered and not the particular direction/region in which a point is located with respect to the low-EDSS region.

However, as suggested by the results obtained with the third data set ($K = 0$), restricting the attributes to not only those sensitive to EDSS leads to the definition of more specific regions in feature space. It is of great interest, e.g. to study if a patient position in this feature space shifts e.g during an exacerbation from the region of younger individuals with no neurological findings (Cluster 1 in Table 7.6) towards a group of higher impairment (Cluster 5 in Table 7.6). Conversely, a displacement in the opposite direction could be employed to monitor whether a patient responds to a therapy or not. Such possibilities would be very advantageous for patients with a relapsing MS course. But they would be also extremely beneficial for the monitoring of progressive MS (either primary or secondary) as these disease courses are characterized by damage to the nerve cells which are sometimes "silent" i.e. occurring without any noticeable symptoms. In both cases, the results obtained suggest new opportunities for the strategic use of image data characterizing quantitatively the *normal appearing* white matter for optimizing personalized treatment and ultimately improving patient outcomes.

Last but not least, comparing the clustering results with additional phenotypic data may enable the association with different disease manifestations and also allow to examine for correlations with more specific mental and sensory functional system scores. Tracking a patient's position in a feature space associated with such information could on the one hand be employed for supporting the initial estimation of the MS subtype. On the other hand, an individual's motion therein could disclose a tendency to convert e.g. from a relapsing remitting MS to a secondary progressive MS. According to the National Multiple Sclerosis Society, 50% of the patients experience this conversion within 10 years after disease onset and 90% within 25 years. Unfortunately, it is not known if and when this transition takes place. Thus, in addition to the positive psychological impact, the objective estimation of such a disease conversion would be very beneficial in supporting treatment decisions as different therapies are recommended for the different MS types. Up to date, there is no method capable of automatically suggesting the MS disease type, expected

disease course or time of transition from relapsing to progressive manifestation.

Finally, clustering the 585-dimensional data with DBSCAN reveals that, similarly to the NPC results, the noise points detected correspond to patients of higher age and EDSS for all of the employed *Eps* thresholds, suggesting again that patients with no or mild neurological findings are more similar with respect to the image-derived features than the rest. However, as seen in Table 7.7, none of the obtained clustering results represent a grouping that could be potentially employed for MS patient stratification. This is in contrast to the NPC results obtained which allow the definition of region that could be employed as a reference for disease evaluation.

Chapter 8

Discussion and Future Prospects

The idea of exploring data by merely imitating a natural process has led to the development of a simple but quite powerful tool able of simultaneously meeting most of the common demands on a clustering routine: it does not require any *a priori* knowledge, it is able to determine the number of clusters in the sample, it can handle high dimensional correlated data and offers reliable noise detection.

The NPC algorithm is designed in a way that allows not only the detection of disjoined groups, but also to obtain hierarchical clusters by running it successively for every cluster found. In this way one can explore the sub-structure of a sample in a manner similar to the traditional hierarchical clustering, however having the advantage of being capable to handle correlated attributes and sub-formations of irregular shape. As discussed in Sec. 4.2.6, taking advantage of this additional option allows the user to choose between a search for subgroups with or without a "zoomed-in" detection of noise in the sub-clusters. Whether to employ this two-step procedure or not is a question of setting up the clustering task which, as discussed in Chapter 3, is a complex problem itself. Generally, if one is interested in exploring the hierarchical structure of a sample by employing the traditional methods, a decision must be made which linkage to use. This choice will influence the final outcome and the researcher has to trade-off between the strengths and the limitations of the different possibilities (e.g single linkage can handle non-elliptical shapes but is sensitive to noise and outliers, while complete link is less susceptible to noise and outliers but tends to break large clusters; group average and Ward's method are less susceptible to noise and outliers, but are biased towards globular clusters). In the absence of any *a priori* information about the data to be clustered, making such a decision is a nontrivial task. In the case of NPC, no such guess has to be made. The only detail that must be specified by the beholder is weather to detect noise in every sub-step or not. In any case, the outcome can be beneficial for the researcher, as this decision is taken deliberately. One has however to bear in mind that employing NPC for hierarchical clustering can significantly increase the computational cost, in a way that strongly depends on the data investigated, i.e. the number of sample points as well as the number and shape of the clusters (cf. Sec. 6.3). In this context, it is also important to state that the time complexity of NPC is $\mathcal{O}(N^2 d)$ and it is thus not recommendable for clustering of very large datasets where methods specially designed for such tasks are preferable (see e.g. [Zhang 1996a]).

Still, if one is rather interested in analyzing a problem associated with small to moderately large data sets, NPC is a very good choice. Here, beside the ability to handle well-separated groups of comparable densities (either in a hard or in

a hierarchical clustering manner), NPC manages to properly recognize clusters of largely different densities while still preserving its capability of properly detecting noise (cf. Sec. 6.2). This was not necessarily an expected feature of the method, as NPC is most closely related to the density-based clustering algorithms which are very advantageous when handling clusters of irregular shape without *a priori* information, but have limitations concerning high-dimensional data and data containing clusters of widely different densities (cf. Sec. 3.2.2).

Furthermore, NPC can generally properly handle clusters of irregular shape. The only exception here are samples containing a cluster completely enclosed by another (e.g spherical or thorus-formed) one. In such peculiar cases, the outer and the inner group fuse to one nucleon and cannot be differentiated due to the mutual attractive strong forces acting between cluster constituents. Here, however, still a very good distinction between points belonging to a cluster and noise is observed (cf. Sec. 6.4.1).

The study of the NPC performance shows that its main strength is in the detection of cohesive clusters regardless of their dimensionality and shape (with the only exception mentioned in the previous paragraph). As it is seen in the case of the Iris data, this can be disadvantageous when a strong overlap of groups is present, since NPC recognizes them as one group of irregular form. However even in this tricky clustering problem, the two-step NPC outperforms the commonly applied clustering algorithms tested. The experiments conducted with artificial data containing touching clusters intentionally overlaid with noise (3. Sample listed in Table 6.1) also confirm the NPC ability to handle such problems. Of all algorithms compared and presented in Sec. 6.4, solely NPC led to the proper recognition of the clusters within the third test sample (touching clusters intentionally overlaid with noise, cf. Fig. 6.7). The results show that the proposed method is able to reveal the natural shape of clusters, even if they have interwined or complicated multidimensional topologies, which is exactly what NPC aims at. Due to the data-driven local scale R, sensitive to the collective behavior of the system, one has on the one hand the advantage of being able to detect cohesive groups. On the other side, the two-step NPC is still able to handle moderate overlaps, since R is adapted to the intra-cluster distances of the subgroups.

In contrast to all of the tested methods, NPC does not only provide the standard clustering outcome (i.e. labels). It also offers a series of auxiliary notification and possibilities to decide how to proceed depending on the clustering task at hand (such as the above mentioned option whether to detect noise within sub-clusters ot not). Valuable insights into the data structure can be obtained not only by the clustering results themselves, but also the system's temperature evolution during dynamic simulation. Based on this information, two cases are automatically detected and reported accordingly to the user: (1) the presence of clusters of different densities and (2) data whose motion in feature space is dominated by the repulsive Coulomb force, meaning that it contains a majority of unstructured points and only few cluster constituents. This additional information can be particularly useful for the interpretation of the results (e.g. lower density within a group indicates a lower

average similarity of its members when compared to a "populous" one) and/or when deciding how to interpret the data (e.g. whether to treat small sub-cluster(s) surrounded by a majority of unstructured points as a cohesive group and noise or not). Furthermore, as the influence of the NPC parameters on its performance is well studied, the researcher is free to vary them to deliberately achieve different effects and explore the system's response. Generally, it is recommended to start with the default parameter values given in Table 5.6. Depending on whether it is desired to increase or decrease the method's sensitivity towards local density fluctuations, one can then continue experimenting by following the instructions listed in Sec. 5.7.

The easily conceivable nature of the NPC procedure as well as its straightforward way of getting a thorough insight even in complex high dimensional data, make the method intuitive and are beneficial for its use of a wide range of scientists from different fields. These properties allow also for the implementation of improvements and extensions which could extend the range of NPC potential applications. For example, one further development of the NPC approach could be realized based on the strategy of running the basic NPC step multiple times with different initial particles' velocities aiming in the current implementation solely at improving noise detection. This gives the opportunity of generating an additional clustering outcome: a fuzzy partitioning of the input data. A conceivable approach to realize this is to determine how often point i is been assigned to the j-th of M clusters, n_j. Eventually, n_j can be employed to calculate the relative frequency of this event and determine the degree of membership of pattern i to cluster j:

$$u_j(\mathbf{x}^{(i)}) = \frac{n_j}{K}, j = 1, ..., M. \tag{8.1}$$

where K is the number if NPC iterations.

Another possibility that can be easily implemented and potentially widen the range of applications of NPC is the adoption of different proximity measures. In any case, however, the effects introduced by such extensions have to be thoroughly studied.

Concerning the computational time, one potential improvement of the designed algorithm could be attained through the introduction of a data specific damping, for example by making it dependent on the sample size. In this way, e.g. samples containing small cohesive groups of similar sizes might be more efficiently processed with respect to the run time. In such cases, the cumulative strong interaction force that is primarily responsible for the rapid increase of a system's temperature in the beginning of NPC simulation, is lower than in cases of samples containing large cohesive groups. Thus, the maximal temperature reached and the terminating temperature are lower as well, resulting in stopping NPC at a later time point (this effect is most prominent for small clusters). However, this applies only if all groups within a sample are small (e. g. in the range of 10 or 15 points). In contrast, the points in data containing cohesive group(s) of large sizes are strongly accelerated (especially at the beginning of the simulation) due to the cumulative strong attractive force and the NPC breaking condition as described in Sec. 4.2

leads to very good results. Note, however, that potential improvements based on such observations should be carefully studied so as to affect solely the NPC run time and not its accuracy.

The range of possibilities to analyze and behold data by means of Nuclear Potential Clustering as well as its very good performance in all tests conducted makes it a valuable tool. It is one of the few novel physics-inspired synergy methods that could stimulate other researchers from virtually all disciplines to search for further intedisciplinary ideas. The current work shows that even though cluster analysis and nuclear physics seem to have nothing in common, such an avenue of combining knowledge from both fields can lead to almost unexpectedly efficient results. I truly believe that the future of science lies in the "melting" of knowledge and deriving new insights into a certain problem by viewing it from different perspectives. Indeed, the history of natural sciences has its roots in philosophy and has started by observing the world as a whole unity. Only later the differentiation of the various subfields takes place leading to the segregation of knowledge and instruments (in the broadest sense). This, of course, is an indispensable part of research, as the only way of investigating a certain issue is to dissect it and explore every observed detail and sub-detail. And that is why this process still goes on. But during the last century, a simultaneous process increasingly takes place in which the different disciplines start to meet again in an extremely fruitful manner. The few physics-inspired clustering methods mentioned in Chapter 3, as well as this work itself are proofs of how well this can function. And though this is not a huge number of works, I believe that there is a high potential of this new type of synergy methods.

Appendix A

List of Notations

a	Thickness of the transition zone between strong interaction and Coulomb interaction (NPC parameter)
$\mathbf{a}$	Acceleration
a_f	Constraint controlling the desired maximum amplitudes for the fast relaxing pool in the context of myelin water content mapping
a_s	Constraint controlling the desired maximum amplitudes for the slow relaxing pool in the context of myelin water content mapping
A	Mass number, indicates the number of nucleons in a nucleus
b	Data-specific local distance scale factor in feature space (NPC)
$\mathbf{B_0}$	Main magnetic field (MRI physics)
$\mathbf{B_1}$	Transmitting magnetic field (MRI physics)
C	Total number of class types, i.e labels
C_i	i-th detected cluster in a crisp clustering procedure
$\mathcal{C}$	Set of C ground truth labels, $\mathcal{C} = \{c_1, c_2, \cdots, c_C\}$, (gold standard partitioning used for evaluation of NPC when processing synthetically generated test data)
$\mathcal{C}'$	Set of M labels resulting from NPC, $\mathcal{C}' = \{c'_1, c'_2, \cdots, c'_M\}$
d	Number of assessed attributes/features per sample, dimensionality of feature space
d_{min}	Minimum mutual particles' distance in a sample
e	Elementary charge
Eps	Input parameter of the clustering method DBSCAN; defines the radius of neighborhood around each data point in a sample
E_{kin}	Kinetic energy
f_i	i-th feature
f_{ij}	Matrix element
$\mathbf{F_i}$	Force acting on the i-th nucleon
$\hbar$	Reduced Planck constant, Dirac constant
H_i	i-th detected hierarchical cluster
H_2O^{myelin}	Myelin bound water fraction

H_2O^{total}	Total water content
I^f	Constraint intervals for the allowed amplitudes the fast relaxing pool in the context of myelin water content mapping
I^s	Constraint intervals for the allowed amplitudes the slow relaxing pool in the context of myelin water content mapping
J	Cost function
K	Number of NPC iterations
K_G	Contrast between two groups of patients, G_1 and G_2, for a given feature
l_{g_1}, l_{g_2}	Parameters used to generate test data of Type 2: lengths of the two generated clusters
l_m	Label of the m-th found cluster
$\mathcal{L}$	Set of discrete labels, e.g. $\mathcal{L} = \{l_1, l_2, ..., l_M\}$ contains M discrete labels
M	Number of groups in a clustering outcome
$MinPts$	Input parameter of the clustering method DBSCAN; specifies a minimum number of neighbors within a predefined radius Eps
n_{noise}	Parameters used to generate test data of Type 1 or Type 2: number random noise points
n_{g_1}, n_{g_2}	Parameters used to generate test data of Type 2: number of points in the two generated clusters
n_{is}, n_{os}	Parameters used to generate test data of Type 1: number of points within the two spheroid-formed clusters
N	In the context of data analysis: sample size, number of observations; in the context of nuclear physics: number of neutrons in a nucleus
$\mathcal{N}(\mu, \sigma)$	Normal distribution with mean μ and standard deviation σ
P	Number of independent samples employed in the NPC optimization procedure when estimating the cost J
p	Order of an ODE solver
q	ODE order
q_R	q_R-quantile of the distance distribution within a sample employed for definition of the range of strong interaction R (NPC)
r	Distance
$\mathbf{r_i}$	Position of the i-th data point (also nucleon in the context of NPC) in feature space

r_h	Parameter used to generate test data of Type 1: average radial distance of the points within the horseshoe-formed cluster towards the central parable
r_{g_1}, r_{g_2}	Parameters used to generate test data of Type 2: radii of the clusters generated along lines g_1 and g_2, respectively
r_{is}, r_{os}	Parameters used to generate test data of Type 1: average radial distance of the points within the two spheroid-formed clusters towards their centers
R	Range of strong interaction (NPC parameter)
s_j	j-th signal compartment (Myelin water content mapping)
s_R	Scaling factor modulating the range of strong interaction R (NPC parameter)
$\mathbf{s}^*$	Signal amplitude vector (Myelin water content mapping)
$S_{0,SE}$	Extrapolated initial signal intensity based on a fit to a single exponential function (used in the context of quantitative MRI)
S_{My}	Myelin amplitude in the context of myelin water content mapping
S_{Tot}	Total signal amplitude in the context of myelin water content mapping
t	Time
t_i	i-th time point (in the context of MRI i-th time point of a measurement i.e echo time)
T	Temperature
T_1	Longitudinal relaxation time
T_2	Transverse relaxation time constant characterizing the decay in spin-echo MRI measurements/sequences
T_2^*	Transverse relaxation time constant characterizing the decay in gradient-echo MRI measurements/sequences
$T_2^{*,f}$	Transverse relaxation time constant characterizing a fast relaxing component
$T_2^{*,s}$	Transverse relaxation time constant characterizing a slow relaxing component
$\widehat{T_2^{*,f}}$	Parameter constraining the amplitude of a fast relaxing pool (quadratic programming in the context of myelin water content mapping)
$\widehat{T_2^{*,s}}$	Parameter constraining the amplitude of a slow relaxing pool (quadratic programming in the context of myelin water content mapping)

$T^*_{2,SE}$	Transverse relaxation time constant estimated based on a fit to a single exponential function (used in the context of myelin water content mapping)
$T_2^{*,3rd-pool}$	Transverse relaxation time constant characterizing a third relaxation pool in the context of myelin water content mapping
u_i	Membership function in the context of fuzzy clustering
U	Potential energy
$U_i^{(k)}$	Final potential energy value for the i-th particle and the k-th NPC iteration
$\mathbf{v}$	Velocity
$\mathbf{v_0}$	Initial velocity
V_0	Maximal depth of the strong interaction potential (NPC)
V_C	Normalized electromagnetic coupling constant (NPC)
$\widehat{w_f}$	Constraint for the total width of the interval where a nonzero amplitude is allowed for the fast relaxing pool in the context of myelin water content mapping
$\widehat{w_s}$	Constraint for the total width of the interval where a nonzero amplitude is allowed for the slow relaxing pool in the context of myelin water content mapping
W_{My}^{rel}	Relative myelin water content
$\mathbf{x}^{(i)}$	Feature vector for the i-th observation
$\mathbf{X}$	Input data of a machine learning algorithm containing all observations, design matrix
$y^{(i)}$	Class type/label of the i-th observation in supervised learning
y_i^{meas}	i-th measurement of an artificial decay signal (Myelin water content mapping)
y_i^{theo}	i-th measurement of a theoretical decay signal (Myelin water content mapping)
Z	Atomic number, indicates the number of protons in a nucleus
α	Flip angle (MRI sequence parameter)
$\mathbf{\Delta B}(x, y, z)$	Magnetic field offset maps in the context of myelin water content mapping
$\mathbf{\Delta G}(x, y, z)$	Gradient map along the slice-select direction in the context of myelin water content mapping
Δ_t	Time interval at which numerical approximation of an integrand is calculated
Δz	Voxel size in z-direction (MRI)

ε_{tol}	Tolerance level of truncation error for an adaptive step ODE solver
ζ	Data-specific local scale in feature space
η	Damping coefficient
λ	Regularization parameter
μ	Mean value
ρ	Density
σ	Standard deviation
ϕ	Potential

Appendix B

List of Abbreviations

ACS	Ant Colony Systems
BIC	Bayesian information criterion for mixture models
BIRCH	A hierarchical clustering algorithm
BW	Band width (MRI sequence parameter)
CLIQUE	A grid-based clustering method
CNS	Cental nervous system
CSF	Cerebrospinal fluid
CURE	A hierarchical clustering algorithm
DBSCAN	A density based clustering algorithm
DENCLUE	A density based clustering algorithm
DQC	Dynamic Quantum Clustering
EA	Evolutionary algorithms
EDSS	Expanded disability status scale
EPI	Echo planar imaging (MRI sequence)
FCM	Fuzzy C-means, a fuzzy clustering algorithm
FCQS	A fuzzy clustering algorithm
FCR	A fuzzy clustering algorithm
FCRS	A fuzzy clustering algorithm
FCSS	A fuzzy clustering algorithm
FID	Free induction decay (MRI)
FLAIR	Fluid attenuated inversion recovery (MRI sequence)
FN	False negatives
FP	False positives
FS	Functional system in the context of MS as defined by [Kurtzke 1983]
FSC	A fuzzy clustering algorithm
ISODATA	A hard clustering algorithm
GM	Gray matter

GRIDCLUST	A density based clustering algorithm
K-means	A hard clustering algorithm
K-medians	A hard clustering algorithm
MAFIA	A density based clustering algorithm
mcDESPOT	Multicomponent driven equilibrium single pulse observation of $T1/T2$ MRI sequence)
MEGE	Multi-echo gradient echo (MRI sequence)
MRI	Magnetic resonance imaging
MS	Multiple Sclerosis
MWF	Myelin bound water fraction
NPC	Nuclear Potential Clustering
ODE	Ordinary differential equation
PAM	A hard clustering algorithm
PPMS	Primary progressive multiple sclerosis
PRMS	Progressive relapsing multiple sclerosis
PSO	Particle Swarm Optimization
QP	Quadratic programming
$QUTE - ST_2^*$	Quantitative T_2^* Image MRI sequence
RI	Rand Index
ROCK	A hierarchical clustering algorithm
RRMS	Relapsing-remitting multiple sclerosis
SI	Swarm Intelligence
SNR	Signal-to-noise-ratio
SPMS	Secondary-progressive multiple sclerosis
STING	A density based clustering algorithm
TE	Echo time (MRI sequence parameter)
TN	True negatives
TP	True positives
TR	Repetition time (MRI sequence parameter)
WM	White matter
X-means	A hard clustering algorithm

Appendix C

Numerical methods for ordinary differential equations

Four methods for approximation to the solutions of ordinary differential equations are discussed in the current work. In this Appendix their computational schemes are given.

To study a system numerically, it is convenient to split the second order differential equation in two coupled first-order equations of the form

$$\frac{dy}{dt} = f(t, y) \tag{C.1}$$

The integration process starts with an initial value $y_0 = y(t_0)$ and an approximation of the integrand is calculated at equally spaced time intervals Δ_t. In the specific case of NPC, all particles have mass normalized to unity, so that for each nucleon the velocity and acceleration are given by:

$$\begin{aligned} \frac{d}{dt}\mathbf{r}(t) &= \mathbf{v}(t) \\ \frac{d}{dt}\mathbf{v}(t) &= \mathbf{F}\big[\mathbf{r}(t)\big] \end{aligned} \tag{C.2}$$

For reasons of clarity and comprehensibility, in the next sections first the general scheme of a presented solver is given as applied for an ODE of the form given by Eq. C.1. This part is then followed by the specific realization of this numerical integration method for the classical equation of motion Eq. 4.5 as rewritten in the form of Eq. C.2.

C.1 Euler Method

The general scheme of the solver is given by:

$$y_{k+1} = y_k + \Delta_t f(t_k, y_k) \tag{C.3}$$

For the Newtonian equation of motion Eq. 4.5:

$$\begin{aligned} \mathbf{r}_{k+1} &= \mathbf{r}_k + \Delta_t \mathbf{v}_k \\ \mathbf{v}_{k+1} &= \mathbf{v}_k + \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k) \end{aligned} \tag{C.4}$$

The discretization error in this case amounts to $\mathcal{O}({\Delta_t}^2)$.

C.2 Runge-Kutta of Second Order

General scheme:

$$\begin{aligned} k_1 &= \Delta_t f(t_k, y_k) \\ k_2 &= \Delta_t f(t_{k+\frac{1}{2}}, y_k + \frac{k_1}{2}) \\ y_{k+1} &= y_k + k_2 \end{aligned} \tag{C.5}$$

For the Newtonian equation of motion Eq. 4.5:

$$\begin{aligned} k_1 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k) \\ l_1 &= \Delta_t \mathbf{v}_k \\ k_2 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{l_1}{2}, \mathbf{v}_k + \frac{k_1}{2}) \\ l_2 &= \Delta_t (\mathbf{v}_k + \frac{k_1}{2}) \end{aligned} \tag{C.6}$$

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{v}_k + k_2 \\ \mathbf{r}_{k+1} &= \mathbf{r}_k + l_2 \end{aligned}$$

The discretization error amounts to $\mathcal{O}({\Delta_t}^3)$.

C.3 Runge-Kutta of Fourth Order

General scheme:

$$\begin{aligned} k_1 &= \Delta_t f(t_k, y_k) \\ k_2 &= \Delta_t f(t_{k+\frac{1}{2}}, y_k + \frac{k_1}{2}) \\ k_3 &= \Delta_t f(t_{k+\frac{1}{2}}, y_k + \frac{k_2}{2}) \\ k_4 &= \Delta_t f(t_{k+1}, y_k + k_3) \\ y_{k+1} &= y_k + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{aligned} \tag{C.7}$$

For the Newtonian equation of motion:

$$
\begin{aligned}
k_1 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k) \\
l_1 &= \Delta_t \mathbf{v}_k \\
k_2 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{l_1}{2}, \mathbf{v}_k + \frac{k_1}{2}) \\
l_2 &= \Delta_t (\mathbf{v}_k + \frac{k_1}{2}) \\
k_3 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{l_2}{2}, \mathbf{v}_k + \frac{k_2}{2}) \\
l_3 &= \Delta_t (\mathbf{v}_k + \frac{k_2}{2}) \\
k_4 &= \Delta_t \mathbf{F}(\mathbf{r}_k + l_3, \mathbf{v}_k + k_3) \\
l_4 &= \Delta_t (\mathbf{v}_k + k_3)
\end{aligned}
\tag{C.8}
$$

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \mathbf{v}_k + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
\mathbf{r}_{k+1} &= \mathbf{r}_k + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)
\end{aligned}
$$

Discretization error amounts to $\mathcal{O}({\Delta_t}^5)$.

Furthermore, for the purposes of NPC the a modification of the solver is considered in which the force, $\mathbf{F}$, acting on a nucleon within a single time step, Δ_t is considered to be constant. This results in the following implementation scheme:

$$
\begin{aligned}
k_1 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k) \\
l_1 &= \Delta_t \mathbf{v}_k \\
k_2 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k + \frac{k_1}{2}) \\
l_2 &= \Delta_t (\mathbf{v}_k + \frac{k_1}{2}) \\
k_3 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k + \frac{k_2}{2}) \\
l_3 &= \Delta_t (\mathbf{v}_k + \frac{k_2}{2}) \\
k_4 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k + k_3) \\
l_4 &= \Delta_t (\mathbf{v}_k + k_3)
\end{aligned}
\tag{C.9}
$$

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \mathbf{v}_k + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
\mathbf{r}_{k+1} &= \mathbf{r}_k + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)
\end{aligned}
$$

In this version of the solver (referred to in the remaining text as modified Runge-

Kutta of fourth order) the force update takes place once per time step, thus resulting in reduced computational time. The reasons motivating and justifying this modification of the solver for the purposes of NPC are discussed in Sec. 5.4.2.

C.4 Runge-Kutta-Fehlberg Method

Runge-Kutta-Fehlberg algorithm which uses a combination of fourth and fifth order Runge-Kutta methods, thus typically abbreviated to RKF45. This method is based on the evaluating of function f in such a way that the function values can be used for both the fourth order and the fifth order method, avoiding thereby additional computations. The Runge-Kutta-Fehlberg method requires at each step the computations of the following six values [Butcher 2008]:

$$\begin{aligned}
k_1 &= \Delta_t f(t_k, y_k) \\
k_2 &= \Delta_t f(t_k + \frac{1}{4}\Delta_t, y_k + \frac{1}{4}k_1) \\
k_3 &= \Delta_t f(t_k + \frac{3}{8}\Delta_t, y_k + \frac{3}{32}k_1 + \frac{9}{32}k_2) \\
k_4 &= \Delta_t f(t_k + \frac{12}{13}\Delta_t, y_k + \frac{1932}{2197}k_1 + \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3) \\
k_5 &= \Delta_t f(t_k + \Delta_t, y_k + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4) \\
k_6 &= \Delta_t f(t_k + 12\Delta_t, y_k - \frac{8}{27}k_1 - 2k_2 + \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5)
\end{aligned} \tag{C.10}$$

In the next step an approximation to the solution is made using a Runge-Kutta method of order four:

$$y_{k+1}^{RK4} = y_k + \frac{2}{25}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4101}k_4 - \frac{1}{5}k_5 \tag{C.11}$$

Next, a more exact value for the solution is obtained using a Runge-Kutta step of order five:

$$y_{k+1}^{RK5} = y_k + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6 \tag{C.12}$$

Given a desire accuracy ε_{tol}, the optimal step size, sh, can be determined by multiplying the current step size Δ_t by the scalar s which is calculated as follows:

$$s = \left(\frac{\varepsilon_{tol}\Delta_t}{2|y_{k+1}^{RK5} - y_{k+1}^{RK4}|} \right)^{\frac{1}{4}} \tag{C.13}$$

Applying this method for the coupled equations Eq. C.2 yields the following procedure:

$$
\begin{aligned}
k_1 &= \Delta_t \mathbf{F}(\mathbf{r}_k, \mathbf{v}_k) \\
l_1 &= \Delta_t \mathbf{v}_k \\
k_2 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{l_1}{4}, \mathbf{v}_k + \frac{k_1}{4}) \\
l_2 &= \Delta_t (\mathbf{v}_k + \frac{k_1}{4}) \\
k_3 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{3}{32} l_1 + \frac{9}{32} l_2, \mathbf{v}_k + \frac{3}{32} k_1 + \frac{9}{32} k_2) \\
l_3 &= \Delta_t (\mathbf{v}_k + \frac{3}{32} k_1 + \frac{9}{32} k_2) \\
k_4 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{1932}{2197} l_1 + \frac{7200}{2197} l_2 + \frac{7296}{2197} l_3, \mathbf{v}_k + \frac{1932}{2197} k_1 + \frac{7200}{2197} k_2 + \frac{7296}{2197} k_3) \\
l_4 &= \Delta_t (\mathbf{v}_k + \frac{1932}{2197} k_1 + \frac{7200}{2197} k_2 + \frac{7296}{2197} k_3) \\
k_5 &= \Delta_t \mathbf{F}(\mathbf{r}_k + \frac{439}{216} l_1 - 8 l_2 + \frac{3680}{513} l_3 - \frac{845}{4104} l_4, \mathbf{v}_k + \frac{439}{216} k_1 - 8 k_2 + \frac{3680}{513} k_3 - \frac{845}{4104} k_4) \\
l_5 &= \Delta_t (\mathbf{v}_k + \frac{439}{216} k_1 - 8 k_2 + \frac{3680}{513} k_3 - \frac{845}{4104} k_4) \\
k_2 &= \Delta_t \mathbf{F}(\mathbf{r}_k - \frac{8}{27} l_1 - 2 l_2 + \frac{3544}{2565} l_3 + \frac{1859}{4104} l_4 - \frac{11}{40} l_5, \mathbf{v}_k - \frac{8}{27} k_1 - 2 k_2 + \frac{3544}{2565} k_3 + \frac{1859}{4104} k_4 - \frac{11}{40} k_5) \\
l_2 &= \Delta_t (\mathbf{v}_k - \frac{8}{27} k_1 - 2 k_2 + \frac{3544}{2565} k_3 + \frac{1859}{4104} k_4 - \frac{11}{40} k_5)
\end{aligned}
\tag{C.14}
$$

Obtain fourth-order Runge-Kutta solution:

$$
\begin{aligned}
\mathbf{v}_{k+1}^{RK4} &= \mathbf{v}_k + \frac{2}{25} k_1 + \frac{1408}{2565} k_3 + \frac{2197}{4101} k_4 - \frac{1}{5} k_5 \\
\mathbf{r}_{k+1}^{RK4} &= \mathbf{r}_k + \frac{2}{25} l_1 + \frac{1408}{2565} l_3 + \frac{2197}{4101} l_4 - \frac{1}{5} l_5
\end{aligned}
\tag{C.15}
$$

Obtain fifth-order Runge-Kutta solution:

$$
\begin{aligned}
\mathbf{v}_{k+1}^{RK5} &= \mathbf{v}_k + \frac{16}{135} k_1 + \frac{6656}{12825} k_3 + \frac{28561}{56430} k_4 - \frac{9}{50} k_5 + \frac{2}{55} k_6 \\
\mathbf{r}_{k+1}^{RK5} &= \mathbf{r}_k + \frac{16}{135} l_1 + \frac{6656}{12825} l_3 + \frac{28561}{56430} l_4 - \frac{9}{50} l_5 + \frac{2}{55} l_6
\end{aligned}
\tag{C.16}
$$

Define time step correction factor:

$$
s = \left(\frac{\varepsilon_{tol} \Delta_t}{2 |(\mathbf{v}_{k+1}^{RK5} || \mathbf{r}_{k+1}^{RK5}) - (\mathbf{v}_{k+1}^{RK4} || \mathbf{r}_{k+1}^{RK4})|} \right)^{\frac{1}{4}} \tag{C.17}
$$

Here, a 5th-order-accurate solution is computed of $\mathcal{O}({\Delta_t}^6)$, while controlling the error the less accurate 4-th-order solution related to it.

Appendix D

Fischer Iris Data

Table D.1: Original Fisher's Iris Data.

Setosa				Versicolor				Virginica			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Appendix E

Digital Data

Bibliography

[Šášik 2001] R. Šášik, T. Hwa, N. Iranfar and W. F. Loomis. *Percolation clustering: a novel approach to the clustering of gene expression patterns in Dictyostelium development.* In Biocomputing 2001, pages 335–347, 2001. (Cited on page 38.)

[Achlioptas 2005] Dimitris Achlioptas and Frank McSherry. *On Spectral Learning of Mixtures of Distributions.* In Learning Theory, 18th Annual Conference on Learning Theory, COLT Proceedings, pages 458–469, 2005. (Cited on page 34.)

[Aderberg 1998] M.R. Aderberg. Cluster analysis for applications. Academic Press, Inc., 1998. (Cited on pages 27 and 30.)

[Aggarwal 2013] Charu C. Aggarwal and Chandan K. Reddy. Data clustering: Algorithms and applications. Chapman & Hall/CRC, 1st édition, 2013. (Cited on page 35.)

[Agrawal 1998] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. *Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications.* SIGMOD Record, vol. 27, no. 2, pages 94–105, June 1998. (Cited on page 35.)

[Aïmeur 2007] Esma Aïmeur, Gilles Brassard and Sébastien Gambs. *Quantum Clustering Algorithms.* In Proceedings of the 24th International Conference on Machine Learning, ICML '07, pages 1–8, New York, NY, USA, 2007. ACM. (Cited on page 38.)

[Alves 2006] V.S. Alves, R.J.G.B. Campello and E.R. Hruschka. *Towards a Fast Evolutionary Algorithm for Clustering.* In IEEE Congress on Evolutionary Computation (CEC), 2006, pages 1776–1783, 2006. (Cited on page 36.)

[Alves 2007] V.S. Alves, R.J.G.B. Campello and E.R. Hruschka. *A Fuzzy Variant of an Evolutionary Algorithm for Clustering.* In IEEE International Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007., pages 1–6, 2007. (Cited on page 36.)

[Arbelaitz 2013] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez and Iñigo Perona. *An Extensive Comparative Study of Cluster Validity Indices.* Pattern Recognition, vol. 46, no. 1, pages 243–256, 2013. (Cited on page 59.)

[Bakshi 2008] Rohit Bakshi, Alan J. Thompson, Maria A. Rocca, Daniel Pelletier, Vincent Dousset, Frederik Barkhof, Matilde Inglese, Charles R.G. Guttmann, Mark A. Horsfield and Massimo Filippi. *MRI in Multiple Sclerosis: Current Status and Future Prospects.* The Lancet Neurology, vol. 7, no. 7, pages 615–625, 2008. (Cited on pages 9 and 12.)

[Ball 1967] Geoffrey H. Ball and David J. Hall. *A Clustering Technique for Summarizing Multivariate Data.* Behavioral Science, vol. 12, no. 2, pages 153–155, 1967. (Cited on page 33.)

[Bandyopadhyay 2001] Sanghamitra Bandyopadhyay and Ujjwal Maulik. *Nonparametric genetic clustering: comparison of validity indices.* IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 31, no. 1, pages 120–125, 2001. (Cited on page 36.)

[Bandyopadhyay 2002a] Sanghamitra Bandyopadhyay and Ujjwal Maulik. *An Evolutionary Technique Based on K-means Algorithm for Optimal Clustering in R^N.* Information Sciences - Informatics and Computer Science, Intelligent Systems, Applications: An International Journal, vol. 146, no. 1-4, pages 221–237, 2002. (Cited on page 36.)

[Bandyopadhyay 2002b] Sanghamitra Bandyopadhyay and Ujjwal Maulik. *Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification.* Pattern Recognition, vol. 35, no. 6, pages 1197–1208, 2002. (Cited on page 36.)

[Baraldi 1999] Andrea Baraldi and Palma Blonda. *A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. Part I and II.* IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 6, pages 778–801, 1999. (Cited on page 32.)

[Barkhof 1997] Frederik Barkhof, Massimo Filippi, David H. Miller, Philip Scheltens, Adriana Campi, Chris H. Polman, Giancarlo Comi, Herman J. Ader, Nick Losseff and Jacob Valk. *Comparison of MRI Criteria at First Presentation to Predict Conversion to Clinically Definite Multiple Sclerosis.* Brain, vol. 120, no. 11, pages 2059–2069, 1997. (Cited on page 9.)

[Ben-Hur 2002] Asa Ben-Hur, Andre Elisseeff and Isabelle Guyon. *A Stability Based Method for Discovering Structure in Clustered Data.* Pacific Symposium on Biocomputing, pages 6–17, 2002. (Cited on page 35.)

[Bender 2010] B. Bender and U. Klose. *The in vivo influence of white matter fiber orientation towards B_0 on T_2^* in the human brain.* NMR in Biomedicine, vol. 23, no. 9, pages 1071–1076, 2010. (Cited on page 11.)

[Beni 1989] G. Beni and J. Wang. *Swarm Intelligence in Cellular Robotic Systems.* In Proceedings of NATO Advanced Workshop on Robots and Biological Systems, pages 26–30, 1989. (Cited on page 37.)

[Bezdek 1981] J.C. Bezdek. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, Norwell, MA, USA, 1981. (Cited on page 34.)

[Bezdek 1992] J.C. Bezdek and R.J. Hathaway. *Numerical Convergence and Interpretation of the Fuzzy C-shells Clustering Algorithm.* IEEE Transactions on Neural Networks, vol. 3, no. 5, pages 787–793, 1992. (Cited on page 34.)

[Bezdek 1994a] J.C. Bezdek, S. Boggavarapu, L.O. Hall and A. Bensaid. *Genetic Algorithm Guided Clustering.* In Proceedings of the First IEEE Conference on Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., volume 1, pages 34–39, 1994. (Cited on page 36.)

[Bezdek 1994b] J.C. Bezdek and R.J. Hathaway. *Optimization of Fuzzy Clustering Criteria Using Genetic Algorithms.* In Proceedings of the First IEEE Conference on Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., volume 2, pages 589–594, 1994. (Cited on page 36.)

[Bezdek 1999] J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva and N.R. Pal. *Will the Real Iris Data Please Stand Up?* Transactions on Fuzzy Systems, vol. 7, no. 3, pages 368–369, 1999. (Cited on page 105.)

[Blatt 1996] Marcelo Blatt, Shai Wiseman and Eytan Domany. *Superparamagnetic Clustering of Data.* Physical Review Letters, vol. 76, pages 3251–3254, 1996. (Cited on page 38.)

[Bradley 1996] P.S. Bradley, O.L. Mangasarian and W.N. Street. *Clustering via Concave Minimization.* In Proceedings of the 9th International Conference on Neural Information Proceeding Sstems, NIPS'96, pages 368–374, Cambridge, MA, USA, 1996. MIT Press. (Cited on pages 34 and 35.)

[Bradley 1998] Paul S. Bradley and Usama M. Fayyad. *Refining Initial Points for K-Means Clustering.* In Proceedings of the Fifteenth International Conference on Machine Learning, pages 91–99, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. (Cited on page 32.)

[Brucker 1978] P. Brucker. *On the Complexity of Clustering Problems.* In Rudolf Henn, Bernhard Korte and Werner Oettli, editeurs, Optimization and Operations Research, volume 157 of *Lecture Notes in Economics and Mathematical Systems*, pages 45–54. Springer Berlin Heidelberg, 1978. (Cited on page 36.)

[Butcher 2008] John C. Butcher. Numerical methods for ordinary differential equations. John Wiley & Sons, 2008. (Cited on page 146.)

[Byrd 1999] Richard H. Byrd, Mary E. Hribar and Jorge Nocedal. *An Interior Point Algorithm for Large-scale Nonlinear Programming.* SIAM Journal on Optimization, vol. 9, no. 4, pages 877–900, 1999. (Cited on page 81.)

[Byrd 2000] Richard H. Byrd, Jean Charles Gilbert and Jorge Nocedal. *A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming.* Mathematical Programming, vol. 89, no. 1, pages 149–185, 2000. (Cited on page 81.)

[Campello 2009] Ricardo J.G.B. Campello, Eduardo R. Hruschka and Vinícius S. Alves. *On the Efficiency of Evolutionary Fuzzy Clustering*. Journal of Heuristics, vol. 15, no. 1, pages 43–75, 2009. (Cited on page 36.)

[Carlisle 2001] A. Carlisle and G. Dozier. *An Off-The-Shelf PSO*. In PSO Workshop, Indianapolis, IN, 2001. (Cited on page 37.)

[Casillas 2003] Arantza Casillas, M.T. González De Lena and R. Martí nez. *Document Clustering Into an Unknown Number of Clusters Using a Genetic Algorithm*. In Text, speech and dialogue, pages 43–49. Springer, 2003. (Cited on page 36.)

[Cherkassky 1998] V. Cherkassky and F. Müller. Learning from data: Concepts, theory and methods. John Wiley & Sons, 1998. (Cited on page 28.)

[Chiş 2008] Monica Chiş. *Hierarchical Clustering Using Evolutionary Algorithms*. Mathematical Methods for Knowledge Discovery and Data Mining, pages 146–156, 2008. (Cited on page 36.)

[Cole 1998] Rowena M. Cole. *Clustering with Genetic Algorithms*, 1998. (Cited on page 36.)

[Compston 2002] Alastair Compston and Alasdair Coles. *Multiple Sclerosis*. The Lancet, vol. 359, no. 9313, pages 1221–1231, 2002. (Cited on page 7.)

[Cowgill 1999] M.C. Cowgill, R.J. Harvey and L.T. Watson. *A genetic algorithm approach to cluster analysis*. Computers & Mathematics with Applications, vol. 37, no. 7, pages 99–108, 1999. (Cited on page 36.)

[Cui 2005] Xiaohui Cui and Thomas E. Potok. *Document Clustering Analysis Based on Hybrid PSO+ K-means Algorithm*. Journal of Computer Sciences (special issue), vol. 27, page 33, 2005. (Cited on page 37.)

[Das 2006] Swagatam Das, Ajith Abraham and Amit Konar. *Spatial Information Based Image Segmentation Using a Modified Particle Swarm Optimization Algorithm*. In Sixth International Conference on Intelligent Systems Design and Applications, 2006. ISDA'06., volume 2, pages 438–444. IEEE, 2006. (Cited on page 37.)

[Dasgupta 2000] Sanjoy Dasgupta. *Experiments with Random Projection*. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI '00, pages 143–151, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. (Cited on page 34.)

[Daszykowski 2001] M. Daszykowski, B. Walczak and D.L. Massart. *Looking for Natural Patterns in Data: Part 1. Density-based Approach*. Chemometrics and Intelligent Laboratory Systems, vol. 56, no. 2, pages 83–92, 2001. (Cited on pages 93 and 107.)

[Davé 1997] Rajesh N. Davé and Raghuram Krishnapuram. *Robust clustering methods: a unified view.* IEEE Transactions on Fuzzy Systems, vol. 5, no. 2, pages 270–293, 1997. (Cited on page 34.)

[De Jong 2006] Kenneth A. De Jong. Evolutionary computation: A unified approach. MIT press, 2006. (Cited on page 36.)

[Denk 2011] Christian Denk, Enedino Hernandez Torres, Alex MacKay and Alexander Rauscher. *The Influence of White Matter Fibre Orientation on MR Signal Phase and Decay.* NMR in Biomedicine, vol. 24, no. 3, pages 246–252, 2011. (Cited on pages 22 and 23.)

[Deoni 2008] Sean C.L. Deoni, Brian K. Rutt, Tarunya Arun, Carlo Pierpaoli and Derek K. Jones. *Gleaning Multicomponent T_1 and T_2 Information from Steady-state Imaging Data.* Magnetic Resonance in Medicine, vol. 60, no. 6, pages 1372–1387, 2008. (Cited on pages 11, 15 and 17.)

[Dolnicar 2010] Sara Dolnicar and Friedrich Leisch. *Evaluation of structure and reproducibility of cluster solutions using the bootstrap.* Marketing Letters, vol. 21, no. 1, pages 83–101, 2010. (Cited on page 31.)

[Dorigo 1996] Marco Dorigo, Vittorio Maniezzo and Alberto Colorni. *The Ant System: Optimization by a Colony of Cooperating Agents.* IEEE Transactions on Systems, Man, and Cybernetics - Part B, vol. 26, no. 1, pages 29–41, 1996. (Cited on page 37.)

[Dorigo 1997] Marco Dorigo and Luca Maria Gambardella. *Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem.* IEEE Transactions on Evolutionary Computation, 1997. (Cited on page 37.)

[Du 2007] Yiping P. Du, Renxin Chu, Dosik Hwang, Mark S. Brown, Bette K. Kleinschmidt-DeMasters, Debra Singel and Jack H. Simon. *Fast Multislice Mapping of the Myelin Water Fraction Using Multicompartment Analysis of T_2^* Decay at 3T: A Preliminary Postmortem Study.* Magnetic Resonance in Medicine, vol. 58, no. 5, pages 865–870, 2007. (Cited on pages 10, 11, 15, 20 and 23.)

[Duda 2001] R.O. Duda, P.E. Hart and D.G. Stork. Pattern classification, 2nd edition. John Wiley & Sons, 2001. (Cited on pages 28, 35 and 92.)

[Dunn 1973] J.C. Dunn. *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-separated Clusters.* Journal of Cybernetics, vol. 3, no. 3, pages 32–57, 1973. (Cited on page 34.)

[Eberhart 2000] R.C. Eberhart and Y. Shi. *Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization.* In Proceedings of the 2000 Congress on Evolutionary Computation, 2000., volume 1, pages 84–88, 2000. (Cited on page 37.)

[Egan 1998] M.A. Egan, M. Krishnamoorthy and K. Rajan. *Comparative study of a genetic fuzzy c-means algorithm and a validity guided fuzzy c-means algorithm for locating clusters in noisy data.* In The 1998 IEEE International Conference on Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., pages 440–445, 1998. (Cited on page 36.)

[Ester 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei 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, pages 226–231. AAAI Press, 1996. (Cited on pages 93, 100, 104, 107, 108 and 126.)

[Estivill-Castro 1997] Vladimir Estivill-Castro and Alan T. Murray. *Spatial Clustering for Data Mining with Genetic Algorithms.* In Proceedings of the International ICSC Symposium on Engineering of Intelligent Systems, EIS-98, pages 317–323. Wiley, 1997. (Cited on page 36.)

[Estivill-Castro 2000] Vladimir Estivill-Castro and Jianhua Yang. *Fast and Robust General Purpose Clustering Algorithms.* In Riichiro Mizoguchi and John Slaney, editeurs, PRICAI 2000 Topics in Artificial Intelligence, volume 1886 of *Lecture Notes in Computer Science*, pages 208–218. Springer Berlin Heidelberg, 2000. (Cited on page 33.)

[Everitt 2011] Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl. Cluster analysis. John Wiley & Sons, 2011. (Cited on pages 33 and 35.)

[Falkenauer 1998] Emanuel Falkenauer. Genetic algorithms and grouping problems. John Wiley & Sons, Inc., New York, NY, USA, 1998. (Cited on page 36.)

[Fasulo 1999] Daniel Fasulo. *An Analysis of Recent Work on Clustering Algorithms.* Department of Computer Science & Engineering, University of Washington, 1999. (Cited on page 32.)

[Fernández-Seara 2000] M.A. Fernández-Seara and F.W. Wehrli. *Postprocessing Technique to Correct for Background Gradients in Image-based R*(2) Measurements.* Magnetic Resonance in Medicine, vol. 44, no. 3, pages 358–366, 2000. (Cited on page 17.)

[Fisher 1936] R. A. Fisher. *The Use of Multiple Measurements in Taxonomic Problems.* Annals of Eugenics, vol. 7, no. 2, pages 179–188, 1936. (Cited on page 105.)

[Forgy 1965] E. W. Forgy. *Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications.* Biometrics, vol. 21, pages 768–769, 1965. (Cited on pages 32 and 34.)

[Fränti 1997] Pasi Fränti, Juha Kivijärvi, Timo Kaukoranta and Olli Nevalainen. *Genetic algorithms for large-scale clustering problems.* The Computer Journal, vol. 40, no. 9, pages 547–554, 1997. (Cited on page 36.)

[Geva 1999] A.B. Geva. *Hierarchical unsupervised fuzzy clustering.* IEEE Transactions on Fuzzy Systems, vol. 7, no. 6, pages 723–733, 1999. (Cited on page 34.)

[Guha 1998] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim. *CURE: An Efficient Clustering Algorithm for Large Databases.* In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98, pages 73–84, New York, NY, USA, 1998. ACM. (Cited on page 33.)

[Guha 1999] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim. *ROCK: A Robust Clustering Algorithm for Categorical Attributes.* In Proceedings of the 15th International Conference on Data Engineering, ICDE '99, pages 512–521, Washington, DC, USA, 1999. IEEE Computer Society. (Cited on page 33.)

[Haber 1985] A. Haber and N.G. LaRocca. *Minimal Record of Disability for Multiple Sclerosis.* New York: National Multiple Sclerosis Society, 1985. (Cited on page 7.)

[Hall 1994] L.O. Hall, J.C. Bezdek, S. Boggavarpu and A. Bensaid. *Genetic Fuzzy Clustering.* In Proceedings of the First International Joint Conference of the North American Fuzzy Information Processing Society Biannual Conference. The Industrial Fuzzy Control and Intelligent Systems Conference, and the NASA Joint Technology Workshop on Neural Networks and Fuzzy Logic (NAFIPS/IFIS/NASA'95), pages 411–415, 1994. (Cited on page 36.)

[Hall 1995] L.O. Hall and B. Ozyurt. *Scaling Genetically Guided Fuzzy Clustering.* In Third International Symposium on Uncertainty Modeling and Analysis, 1995, and Annual Conference of the North American Fuzzy Information Processing Society. Proceedings of ISUMA - NAFIPS '95., pages 328–332, 1995. (Cited on page 36.)

[Hall 1999] L.O. Hall, I.B. Ozyurt and J.C. Bezdek. *Clustering With a Genetically Optimized Approach.* IEEE Transactions on Evolutionary Computation, vol. 3, no. 2, pages 103–112, 1999. (Cited on page 36.)

[Halliday 2014] David Halliday, Robert Resnick and Jearl Walker. Fundamentals of physics, 10th edition. Wiley, 2014. (Cited on page 45.)

[Handl 2002] Julia Handl and Bernd Meyer. *Improved Ant-based Clustering and Sorting in a Document Retrieval Interface.* In Parallel Problem Solving from Nature-PPSN VII, pages 913–923. Springer, 2002. (Cited on page 37.)

[Handl 2003] J. Handl, J. Knowles and M. Dorigo. *On the Performance of Ant-based Clustering.* In Ajith Abraham, Mario Köppen and Katrin Franke, editeurs,

Design and Application of Hybrid Intelligent Systems, pages 204–213. IOS Press, Amsterdam, The Netherlands, 2003. (Cited on page 37.)

[Handl 2007] J. Handl and J. Knowles. *An Evolutionary Approach to Multiobjective Clustering.* IEEE Transactions on Evolutionary Computation, vol. 11, no. 1, pages 56–76, 2007. (Cited on page 36.)

[Hassanzadeh 2010] H.R. Hassanzadeh and M. Rouhani. *A Multi-objective Gravitational Search Algorithm.* In Second International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), 2010, pages 7–12, 2010. (Cited on page 37.)

[Hathaway 2000] R.J. Hathaway, J.C. Bezdek and Yingkang Hu. *Generalized Fuzzy C-means Clustering Strategies Using Lp Norm Distances.* IEEE Transactions on Fuzzy Systems, vol. 8, no. 5, pages 576–582, 2000. (Cited on page 34.)

[He 2009] Xiang He and Dmitriy A. Yablonskiy. *Biophysical Mechanisms of Phase Contrast in Gradient Echo MRI.* Proceedings of the National Academy of Sciences USA, vol. 106, no. 32, pages 13558–13563, 2009. (Cited on page 22.)

[Hilkert 2010] Thomas Hilkert. *Quantitative MR-Korrelationsbildgebung bei Multipler Sklerose*, 2010. (Cited on page 115.)

[Hochbaum 1985] Dorit S. Hochbaum and David B. Shmoys. *A Best Possible Heuristic for the k-Center Problem.* Mathematics of Operational Research, vol. 10, no. 2, pages 180–184, 1985. (Cited on page 34.)

[Hoeppner 1997] F. Hoeppner. *Fuzzy shell clustering algorithms in image processing: fuzzy C-rectangular and 2-rectangular shells.* IEEE Transactions on Fuzzy Systems, vol. 5, no. 4, pages 599–613, 1997. (Cited on page 34.)

[Höppner 1999] F. Höppner, F. Klawonn, R. Kruse and T. Runkler. Fuzzy cluster analysis: Methods for classification, data analysis, and image recognition. John Wiley, New York, 1999. (Cited on page 34.)

[Horn 2001a] David Horn and Assaf Gottlieb. *Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics.* Phys. Rev. Lett., vol. 88, page 018702, Dec 2001. (Cited on page 38.)

[Horn 2001b] David Horn and Assaf Gottlieb. *The Method of Quantum Clustering.* In NIPS, pages 769–776. MIT Press, 2001. (Cited on page 38.)

[Hruschka 2003] Eduardo R. Hruschka and Nelson F. f. Ebecken. *A Genetic Algorithm for Cluster Analysis.* Intelligent Data Analysis, vol. 7, no. 1, pages 15–25, 2003. (Cited on page 36.)

[Hruschka 2004a] Eduardo R. Hruschka, Ricardo J.G.B. Campello and Leandro N. de Castro. *Improving the Efficiency of a Clustering Genetic Algorithm.*

In Advances in Artificial Intelligence–IBERAMIA 2004, pages 861–870. Springer, 2004. (Cited on page 36.)

[Hruschka 2004b] Eduardo R. Hruschka, Leandro N. de Castro and Ricardo J.G.B. Campello. *Evolutionary algorithms for clustering gene-expression data.* Fourth IEEE International Conference on Data Mining (ICDM'04), pages 403–406, 2004. (Cited on page 36.)

[Hruschka 2004c] E.R. Hruschka, R.J.G.B. Campello and L.N. de Castro. *Evolutionary search for optimal fuzzy c-means clustering.* In IEEE International Conference on Fuzzy Systems, 2004. Proceedings., volume 2, pages 685–690, 2004. (Cited on page 36.)

[Hruschka 2006] Eduardo R. Hruschka, Ricardo J.G.B. Campello and Leandro N. De Castro. *Evolving clusters in gene-expression data.* Information Sciences - Informatics and Computer Science, Intelligent Systems, Applications: An International Journal, vol. 176, no. 13, pages 1898–1927, 2006. (Cited on page 36.)

[Hruschka 2009] Eduardo Raul Hruschka, Ricardo J.G.B. Campello, Alex A. Freitas and André C. Ponce Leon F. De Carvalho. *A Survey of Evolutionary Algorithms for Clustering.* Transactios on Systems, Man and Cybernetics Part C, vol. 39, no. 2, pages 133–155, 2009. (Cited on page 36.)

[Hung 2001] Ming-Chuan Hung and Don-Lin Yang. *An Efficient Fuzzy C-Means Clustering Algorithm.* In Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001, pages 225–232, 2001. (Cited on page 34.)

[Hwang 2010] Dosik Hwang, Dong-Hyun Kim and Yiping P. Du. *In Vivo Multi-slice Mapping of Myelin Water Content Using T_2^* Decay.* NeuroImage, vol. 52, no. 1, pages 198–204, 2010. (Cited on pages 10, 11, 15, 20 and 23.)

[Hwang 2011] Dosik Hwang, Hyunjin Chung, Yoonho Nam, Yiping P. Du and Ung Jang. *Robust Mapping of the Myelin Water Fraction in the Presence of Noise: Synergic Combination of Anisotropic Diffusion Filter and Spatially Regularized Nonnegative Least Squares Algorithm.* Journal of Magnetic Resonance Imaging, vol. 34, no. 1, pages 189–195, 2011. (Cited on page 10.)

[Inaba 1994] Mary Inaba, Naoki Katoh and Hiroshi Imai. *Applications of Weighted Voronoi Diagrams and Randomization to Variance-based K-clustering: (Extended Abstract).* In Proceedings of the Tenth Annual Symposium on Computational Geometry, SCG '94, pages 332–339, New York, NY, USA, 1994. ACM. (Cited on page 104.)

[Jain 1988] Anil K. Jain and Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. (Cited on pages 32 and 92.)

[Jain 1999] A.K. Jain, M.N. Murty and P.J. Flynn. *Data Clustering: A Review.* ACM Comput. Surv., vol. 31, no. 3, pages 264–323, 1999. (Cited on page 32.)

[Jain 2000] Anil K. Jain, Robert P. W. Duin and Jianchang Mao. *Statistical Pattern Recognition: A Review.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pages 4–37, 2000. (Cited on page 32.)

[Jain 2010] Anil K. Jain. *Data Clustering: 50 Years Beyond K-means.* Pattern Recognition Letters, vol. 31, no. 8, pages 651–666, 2010. (Cited on pages 28, 30, 31, 32 and 38.)

[Jiang 2007] M. Jiang, Y.P. Luo and S.Y. Yang. *Stochastic Convergence Analysis and Parameter Selection of the Standard Particle Swarm Optimization Algorithm.* Information Processing Letters, vol. 102, no. 1, pages 8–16, 2007. (Cited on page 37.)

[Kalantari 2011] Saeed Kalantari, Cornelia Laule, Thorarin A Bjarnason, Irene M Vavasour and Alex L MacKay. *Insight into in vivo magnetization exchange in human white matter regions.* Magnetic resonance in medicine, vol. 66, no. 4, pages 1142–1151, 2011. (Cited on page 23.)

[Kanade 2003] Parag M. Kanade and Lawrence O. Hall. *Fuzzy Ants as a Clustering Concept.* In 22nd International Conference of the North American Fuzzy Information Processing Society, 2003. NAFIPS 2003., pages 227–232. IEEE, 2003. (Cited on page 37.)

[Karypis 1999] George Karypis, Eui-Hong (Sam) Han and Vipin Kumar. *Chameleon: Hierarchical Clustering Using Dynamic Modeling.* Computer, vol. 32, no. 8, pages 68–75, 1999. (Cited on page 33.)

[Kaufman 1990] Leonard Kaufman and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1990. A Wiley-Interscience publication. (Cited on pages 32 and 33.)

[Kennedy 1995] J. Kennedy and R. Eberhart. *Particle Swarm Optimization.* In IEEE International Conference on Neural Networks, 1995. Proceedings., volume 4, pages 1942–1948 vol.4, 1995. (Cited on page 37.)

[Kersten 1997] P.R. Kersten. *Implementation Issues in the Fuzzy C-medians Clustering Algorithm.* In Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, 1997., volume 2, pages 957–962, 1997. (Cited on page 34.)

[Kivijärvi 2003] Juha Kivijärvi, Pasi Fränti and Olli Nevalainen. *Self-Adaptive Genetic Algorithm for Clustering.* Journal of Heuristics, vol. 9, no. 2, pages 113–129, 2003. (Cited on page 36.)

[Klawonn 1998] Frank Klawonn and Annette Keller. *Fuzzy Clustering with Evolutionary Algorithms.* International Journal of Intelligent Systems, vol. 13, no. 10-11, pages 975–991, 1998. (Cited on page 36.)

[Kleinberg 2003] Jon M. Kleinberg. *An Impossibility Theorem for Clustering.* In S. Becker, S. Thrun and K. Obermayer, editeurs, Advances in Neural Information Processing Systems 15, pages 463–470. MIT Press, 2003. (Cited on page 31.)

[Kolen 2002] J. F. Kolen and T. Hutcheson. *Reducing the Time Complexity of the Fuzzy C-means Algorithm.* IEEE Transactions on Fuzzy Systems, vol. 10, no. 2, pages 263–267, 2002. (Cited on page 34.)

[Kolind 2009] Shannon H. Kolind, Burkhard Mädler, Stefan Fischer, David K.B. Li and Alex L. MacKay. *Myelin Water Imaging: Implementation and Development at 3.0 T and Comparison to 1.5 T Measurements.* Magnetic Resonance in Medicine, vol. 62, no. 1, pages 106–115, 2009. (Cited on pages 10, 11, 15 and 23.)

[Kolind 2012] Shannon Kolind, Lucy Matthews, Heidi Johansen-Berg, M. Isabel Leite, Steven C.R. Williams, Sean Deoni and Jackie Palace. *Myelin Water Imaging Reflects Clinical Variability in Multiple Sclerosis.* Neuroimage, vol. 60, no. 1, pages 263–270, 2012. (Cited on page 11.)

[Krishna 1999] K. Krishna and M.N. Murty. *Genetic K-means algorithm.* IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 3, pages 433–439, 1999. (Cited on page 36.)

[Krishnapuram 1992] R. Krishnapuram, O. Nasraoui and H. Frigui. *The fuzzy c spherical shells algorithm: A new approach.* IEEE Transactions on Neural Networks, vol. 3, no. 5, pages 663–671, 1992. (Cited on page 34.)

[Krishnapuram 1993] R. Krishnapuram and J.M. Keller. *A Possibilistic Approach to Clustering.* IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pages 98–110, 1993. (Cited on page 34.)

[Krishnapuram 1995] R. Krishnapuram, H. Frigui and O. Nasraoui. *Fuzzy and Possibilistic Shell Clustering Algorithms and Their Application to Boundary Detection and Surface Approximation. Part I and II.* IEEE Transactions on Fuzzy Systems, vol. 3, no. 1, pages 29–60, 1995. (Cited on page 34.)

[Krovi 1992] R. Krovi. *Genetic Algorithms for Clustering: a Preliminary Investigation.* In Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, 1992., volume 4, pages 540–544, 1992. (Cited on page 36.)

[Kuhlmann 2002] Tanja Kuhlmann, Gueanelle Lingfeld, Andreas Bitsch, Jana Schuchardt and Wolfgang Brück. *Acute Axonal Damage in Multiple Sclerosis is Most Extensive in Early Disease Stages and Decreases Over Time.* Brain, vol. 125, no. 10, pages 2202–2212, 2002. (Cited on pages 7 and 113.)

[Kuncheva 1997] Ludmila I. Kuncheva and James C. Bezdek. *Selection of Cluster Prototypes from Data by a Genetic Algorithm.* In Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing, pages 1683–1688, 1997. (Cited on page 36.)

[Kurtzke 1983] John F. Kurtzke. *Rating Neurologic Impairment in Multiple Sclerosis an Expanded Disability Status Scale (EDSS).* Neurology, vol. 33, no. 11, pages 1444–1444, 1983. (Cited on pages 6, 7, 8 and 141.)

[Laule 2004] C. Laule, I.M. Vavasour, G.R.W. Moore, J. Oger, D.K.B. Li, D.W. Paty and A.L. MacKay. *Water Content and Myelin Water Fraction in Multiple Sclerosis.* Journal of Neurology, vol. 251, no. 3, pages 284–293, 2004. (Cited on pages 10, 11, 15, 20 and 23.)

[Laule 2006] C. Laule, E. Leung, D.K.B. Li, A.L. Traboulsee, D.W. Paty, A.L. MacKay and G.R.W. Moore. *Myelin Water Imaging in Multiple Sclerosis: Quantitative Correlations with Histopathology.* Multiple Sclerosis Journal, vol. 12, no. 6, pages 747–753, 2006. (Cited on pages 10 and 14.)

[Laule 2007] Cornelia Laule, Irene M. Vavasour, Burkhard Mädler, Shannon H. Kolind, Sandra M. Sirrs, Elana E. Brief, Anthony L. Traboulsee, G.R. Moore, David K.B. Li and Alex L. MacKay. *MR Evidence of Long T_2 Water in Pathological White Matter.* Journal of Magnetic Resonance Imaging, vol. 26, no. 4, pages 1117–1121, 2007. (Cited on page 16.)

[Laule 2008] Cornelia Laule, Piotr Kozlowski, Esther Leung, David K.B. Li, Alex L. MacKay and G.R. Wayne Moore. *Myelin Water Imaging of Multiple Sclerosis at 7 T: Correlations with Histopathology.* Neuroimage, vol. 40, no. 4, pages 1575–1580, 2008. (Cited on page 10.)

[Laule 2011] Cornelia Laule, Irene M. Vavasour, Esther Leung, David K.B. Li, Piotr Kozlowski, Anthony L. Traboulsee, Joel Oger, Alex L. MacKay and G.R. Wayne Moore. *Pathological Basis of Diffusely Abnormal White Matter: Insights from Magnetic Resonance Imaging and Histology.* Multiple Sclerosis Journal, vol. 17, no. 2, pages 144–150, 2011. (Cited on page 10.)

[Lee 2011] Jongho Lee, Peter van Gelderen, Li-Wei Kuo, Hellmut Merkle, Afonso C. Silva and Jeff H. Duyn. *T_2^*-based Fiber Orientation Mapping.* Neuroimage, vol. 57, no. 1, pages 225–234, 2011. (Cited on page 22.)

[Lee 2017] Jongho Lee, Yoonho Nam, Joon Yul Choi, Eung Yeop Kim, Se-Hong Oh and Dong-Hyun Kim. *Mechanisms of T_2^* Anisotropy and Gradient Echo Myelin Water Imaging.* NMR in Biomedicine, vol. 30, no. 4, 2017. (Cited on pages 11, 12 and 22.)

[Lenz 2012] C. Lenz, M. Klarhöfer and K. Scheffler. *Feasibility of In Vivo Myelin Water Imaging Using 3D Multigradient-echo Pulse Sequences.* Magnetic Res-

onance in Medicine, vol. 68, no. 2, pages 523–528, 2012. (Cited on pages 10 and 11.)

[Li-Ping 2005] Zhang Li-Ping, Yu Huan-Jun and Hu Shang-Xu. *Optimal Choice of Parameters for Particle warm Optimization.* Journal of Zhejiang University Science A, vol. 6, no. 6, pages 528–534, 2005. (Cited on page 37.)

[Likas 2003] Aristidis Likas, Nikos Vlassis and Jakob J. Verbeek. *The Global K-Means Clustering Algorithm.* Pattern Recognition, vol. 36, no. 2, pages 451–461, 2003. (Cited on page 32.)

[Liu 1995] Jianzhuang Liu and Weixin Xie. *A genetics-based approach to fuzzy clustering.* In Proceedings of 1995 IEEE Int Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., volume 4, pages 2233–2240, 1995. (Cited on page 36.)

[Liu 2003] H. Liu, J. Li and M.A. Chapman. *Automated road extraction from satellite imagery using hybrid genetic algorithms and cluster analysis.* Journal of Environmental Informatics, vol. 1, no. 2, pages 40–47, 2003. (Cited on page 36.)

[Lloyd 1957] Stuart P. Lloyd. *Least Squares Quantization in PCM's Bell Telephone Labs.* Bell Telephone Laboratories Paper, 1957. (Cited on pages 32 and 92.)

[Lloyd 1982] Stuart P. Lloyd. *Least Squares Quantization in PCM.* IEEE Transactions on Information Theory, vol. 28, pages 129–137, 1982. (Cited on pages 32 and 92.)

[Lozano 1999] J.A. Lozano and P. Larrañaga. *Applying Genetic Algorithms to Search for the Best Hierarchical Clustering of a Dataset.* Pattern Recognition Letters, vol. 20, no. 9, pages 911–918, 1999. (Cited on page 36.)

[Lu 2004a] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng and Susan J. Brown. *FGKA: A Fast Genetic K-means Clustering Algorithm.* In Proceedings of the 2004 ACM Symposium on Applied Computing, SAC '04, pages 622–623, New York, NY, USA, 2004. ACM. (Cited on page 36.)

[Lu 2004b] Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng and Susan J. Brown. *Incremental Genetic K-means Algorithm and Its Application in Gene Expression Data Analysis.* BMC bioinformatics, vol. 5, no. 1, page 172, 2004. (Cited on page 36.)

[Lucasius 1993] Carlos B Lucasius, Adrie D Dane and Gerrit Kateman. *On k-medoid Clustering of Large Data Sets with the Aid of a Genetic Algorithm: Background, Feasibility and Comparison.* Analytica Chimica Acta, vol. 282, no. 3, pages 647–669, 1993. (Cited on page 36.)

[Lumer 1994] Erik D Lumer and Baldo Faieta. *Diversity and Adaptation in Populations of Clustering Ants.* In Proceedings of the third international conference on Simulation of adaptive behavior: from animals to animats 3, pages 501–508. MIT Press, 1994. (Cited on page 37.)

[Ma 2006] P.C.H. Ma, K.C.C. Chan, Xin Yao and D.K.Y. Chiu. *An Evolutionary Clustering Algorithm for Gene Expression Microarray Data Analysis.* IEEE Transactions on Evolutionary Computation, vol. 10, no. 3, pages 296–314, 2006. (Cited on page 36.)

[MacQueen 1967] J. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations.* In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297. University of California Press, 1967. (Cited on pages 32 and 34.)

[Man 1994] Y.H. Man and I. Gath. *Detection and Separation of Ring-shaped Clusters Using Fuzzy Clustering.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 8, pages 855–861, 1994. (Cited on page 34.)

[Manning 2008] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, New York, NY, USA, 2008. (Cited on page 59.)

[Mathur-De Vré 1984] R. Mathur-De Vré. *Biomedical Implications of the Relaxation Behaviour of Water Related to NMR Imaging.* The British Journal of Radiology, vol. 57, no. 683, pages 955–976, 1984. (Cited on pages 10 and 22.)

[Maulik 2000] Ujjwal Maulik and Sanghamitra Bandyopadhyay. *Genetic algorithm-based clustering technique.* Pattern recognition, vol. 33, no. 9, pages 1455–1465, 2000. (Cited on page 36.)

[Maulik 2003] U. Maulik and S. Bandyopadhyay. *Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification.* IEEE Transactions on Geoscience and Remote Sensing, vol. 41, no. 5, pages 1075–1081, 2003. (Cited on page 36.)

[Merz 2002] Peter Merz and Andreas Zell. *Clustering Gene Expression Profiles with Memetic Algorithms.* In Parallel Problem Solving from Nature-PPSN VII, pages 811–820. Springer, 2002. (Cited on page 36.)

[Meyers 2009] Sandra M. Meyers, Cornelia Laule, Irene M. Vavasour, Shannon H. Kolind, Burkhard Mädler, Roger Tam, Anthony L. Traboulsee, Jimmy Lee, David K.B. Li and Alex L. MacKay. *Reproducibility of Myelin Water Fraction Analysis: a Comparison of Region of Interest and Voxel-based Analysis Methods.* Magnetic Resonance Imaging, vol. 27, no. 8, pages 1096–1103, 2009. (Cited on page 10.)

[Minty 2009] Evan P. Minty, Thorarin A. Bjarnason, Cornelia Laule and Alex L. MacKay. *Myelin Water Measurement in the Spinal Cord.* Magnetic Resonance in Medicine, vol. 61, no. 4, pages 883–892, 2009. (Cited on page 10.)

[Monmarché 1999] Nicolas Monmarché, Mohamed Slimane and Gilles Venturini. *AntClass: Discovery of Clusters in Numeric Data by an Hybridization of an Ant Colony with the Kmeans Algorithm*, 1999. (Cited on page 37.)

[Murthy 1996] Chivukula A. Murthy and Nirmalya Chowdhury. *In Search of Optimal Clusters Using Genetic Algorithms.* Pattern Recognition Letters, vol. 17, no. 8, pages 825–832, 1996. (Cited on page 36.)

[Murty 1988] Katta G. Murty and Feng-Tien Yu. Linear complementarity, linear and nonlinear programming. Internet Edition, 1988. (Cited on page 15.)

[Naldi 2007] Murilo Coelho Naldi and André Carlos Ponce Leon Ferreira de Carvalho. *Clustering Using Genetic Algorithm Combining Validation Criteria.* In Proceedings of the 15th European Symposium on Artificial Neural Networks, pages 139–144, 2007. (Cited on page 36.)

[Nam 2015] Yoonho Nam, Jongho Lee, Dosik Hwang and Dong-Hyun Kim. *Improved Estimation of Myelin Water Fraction Using Complex Model Fitting.* NeuroImage, vol. 116, pages 214–221, 2015. (Cited on pages 11, 12 and 22.)

[Neeb 2006a] Heiko Neeb, Karl Zilles and N. Jon Shah. *Fully Automated Detection of Cerebral Water Content Changes: Study of Age- and Gender-related Patterns with Quantitative MRI.* Neuroimage, vol. 29, no. 3, pages 910–922, 2006. (Cited on page 18.)

[Neeb 2006b] Heiko Neeb, Karl Zilles and N. Jon Shah. *A new method for fast quantitative mapping of absolute water content in vivo.* Neuroimage, vol. 31, no. 3, pages 1156–1168, 2006. (Cited on page 23.)

[Neeb 2008] H. Neeb, V. Ermer, T. Stoecker and N.J. Shah. *Fast Quantitative Mapping of Absolute Water Content with Full Brain Coverage.* Neuroimage, vol. 42, no. 3, pages 1094–1109, 2008. (Cited on pages 2, 12, 13, 18, 20, 23 and 113.)

[Neeb 2012] Heiko Neeb, Jochen Schenk and Bernd Weber. *Multicentre Absolute Myelin Water Content Mapping: Development of a Whole Brain Atlas and Application to Low-grade Multiple Sclerosis.* NeuroImage: Clinical, vol. 1, no. 1, pages 121–130, 2012. (Cited on page 13.)

[Neeb 2014] Heiko Neeb, Andreas Boer, Detlef Gliedstein, Matthias Raspe and Jochen Schenk. *Predicting Multiple Sclerosis from Normal Appearing Brain Matter - Combination of Quantitative MRI Metrics with Supervised Learning.* In Proceedings of The World Congress on Engineering and Computer Science 2014, pages 502–507. WCECS 2014, 2014. (Cited on pages 13 and 114.)

[Neema 2007] Mohit Neema, James Stankiewicz, Ashish Arora, Zachary D. Guss and Rohit Bakshi. *MRI in Multiple Sclerosis: What's Inside the Toolbox?* Neurotherapeutics, vol. 4, no. 4, pages 602–617, 2007. (Cited on pages 9 and 12.)

[Nguyen 2012] Thanh D. Nguyen, Cynthia Wisnieff, Mitchell A. Cooper, Dushyant Kumar, Ashish Raj, Pascal Spincemaille, Yi Wang, Tim Vartanian and Susan A. Gauthier. *T2prep Three-dimensional Spiral Imaging with Efficient Whole Brain Coverage for Myelin Water Quantification at 1.5 Tesla.* Magnetic Resonance in Medicine, vol. 67, no. 3, pages 614–621, 2012. (Cited on page 10.)

[Nielsen 2011] Michael A. Nielsen and Isaac L. Chuang. Quantum computation and quantum information: 10th anniversary edition. Cambridge University Press, New York, NY, USA, 2011. (Cited on page 37.)

[Nugent 2010] Rebecca Nugent and Marina Meila. *An Overview of Clustering Applied to Molecular Biology.* In Heejung Bang, Xi Kathy Zhou, Heather L. van Epps and Madhu Mazumdar, editeurs, Statistical Methods in Molecular Biology, volume 620 of *Methods in Molecular Biology*, pages 369–404. Humana Press, 2010. (Cited on pages 34 and 35.)

[Oh 2007] Joonmi Oh, Eric T. Han, Michael C. Lee, Sarah J. Nelson and Daniel Pelletier. *Multislice Brain Myelin Water Fractions at 3T in Multiple Sclerosis.* Journal of Neuroimaging, vol. 17, no. 2, pages 156–163, 2007. (Cited on pages 10, 11, 15, 20 and 23.)

[Omran 2002] M. Omran, A. Salman and A.P. Engelbrecht. *Image Classification Using Particle Swarm Optimization.* In Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning, volume 1, pages 18–22. Singapore, 2002. (Cited on page 37.)

[Omran 2005] M. Omran, A.P. Engelbrecht and A. Salman. *Particle Swarm Optimization Method for Image Clustering.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no. 03, pages 297–321, 2005. (Cited on page 37.)

[Pakhira 2005] Malay K. Pakhira, Sanghamitra Bandyopadhyay and Ujjwal Maulik. *A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification.* Fuzzy Sets and Systems, vol. 155, no. 2, pages 191–214, 2005. (Cited on page 36.)

[Pan 2007] Shih-Ming Pan and Kuo-Sheng Cheng. *Evolution-Based Tabu Search Approach to Automatic Clustering.* IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, no. 5, pages 827–838, 2007. (Cited on page 36.)

[Park 2005] Han-Saem Park, Si-Ho Yoo and Sung-Bae Cho. *Evolutionary fuzzy clustering algorithm with knowledge-based evaluation and applications for gene expression profiling.* Journal of Computational and Theoretical Nanoscience, vol. 2, no. 4, pages 524–533, 2005. (Cited on page 36.)

[Peña 1999] J.M Peña, J.A Lozano and P Larrañaga. *An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm.* Pattern Recognition Letters, vol. 20, no. 10, pages 1027–1040, 1999. (Cited on pages 32 and 33.)

[Pedersen 2010] Magnus Erik Hvass Pedersen. *Good Parameters for Particle Swarm Optimization.* Hvass Lab., Copenhagen, Denmark, Tech. Rep. HL1001, 2010. (Cited on page 37.)

[Pelleg 2000] Dan Pelleg and Andrew W. Moore. *X-means: Extending K-means with Efficient Estimation of the Number of Clusters.* In Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. (Cited on page 34.)

[Peneva 2013] Anna Peneva. *Cluster Analysis with Quantitative MRI Data*, 2013. (Cited on page 24.)

[Polman 2005] Chris H. Polman, Stephen C. Reingold, Gilles Edan, Massimo Filippi, Hans-Peter Hartung, Ludwig Kappos, Fred D. Lublin, Luanne M. Metz, Henry F. McFarland, Paul W. O'Connor*et al. Diagnostic Criteria for Multiple Sclerosis: 2005 Revisions to the "McDonald Criteria".* Annals of Neurology, vol. 58, no. 6, pages 840–846, 2005. (Cited on page 9.)

[Polman 2011] Chris H. Polman, Stephen C. Reingold, Brenda Banwell, Michel Clanet, Jeffrey A. Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos*et al. Diagnostic Criteria for Multiple Sclerosis: 2010 revisions to the McDonald Criteria.* Annals of Neurology, vol. 69, no. 2, pages 292–302, 2011. (Cited on page 7.)

[Powers 2011] David Martin Ward Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* International Journal of Machine Learning Technology, vol. 2, no. 1, pages 37–63, 2011. (Cited on page 62.)

[Rabanal 2007] Pablo Rabanal, Ismael Rodríguez and Fernando Rubio. *Using River Formation Dynamics to Design Heuristic Algorithms.* In Unconventional Computation, pages 163–177. Springer, 2007. (Cited on page 37.)

[Ramos 2002] Vitorino Ramos, Fernando Muge and Pedro Pina. *Self-Organized Data and Image Retrieval as a Consequence of Inter-Dynamic Synergistic Relationships in Artificial Ant Colonies.* HIS, vol. 87, pages 500–512, 2002. (Cited on page 37.)

[Rashedi 2009] Esmat Rashedi, Hossein Nezamabadi-Pour and Saeid Saryazdi. *GSA: a gravitational search algorithm.* Information sciences, vol. 179, no. 13, pages 2232–2248, 2009. (Cited on page 37.)

[Rashedi 2010] Esmat Rashedi, Hossein Nezamabadi-Pour and Saeid Saryazdi. *BGSA: Binary Gravitational Search Algorithm.* Natural Computing, vol. 9, no. 3, pages 727–745, 2010. (Cited on page 37.)

[Sanjeev 2001] Arora Sanjeev and Ravi Kannan. *Learning Mixtures of Arbitrary Gaussians.* In Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing, STOC '01, pages 247–257, New York, NY, USA, 2001. ACM. (Cited on page 34.)

[Sati 2013] Pascal Sati, Peter van Gelderen, Afonso C. Silva, Daniel S. Reich, Hellmut Merkle, Jacco A. De Zwart and Jeff H. Duyn. *Micro-compartment Specific T_2^* Relaxation in the Brain.* Neuroimage, vol. 77, pages 268–278, 2013. (Cited on page 11.)

[Scheunders 1997] Paul Scheunders. *A Genetic C-means Clustering Algorithm Applied to Color Image Quantization.* Pattern Recognition, vol. 30, no. 6, pages 859–866, 1997. (Cited on page 36.)

[Sheng 2004] Weiguo Sheng and Xiaohui Liu. *A Hybrid Algorithm for K-medoid Clustering of Large Data Sets.* In Congress on Evolutionary Computation, 2004, volume 1, pages 77–82, 2004. (Cited on page 36.)

[Shi 1998] Yuhui Shi and Russell C. Eberhart. *Parameter Selection in Particle Swarm Optimization.* In V.W. Porto, N. Saravanan, D. Waagen and A.E. Eiben, editeurs, Evolutionary Programming VII, volume 1447 of *Lecture Notes in Computer Science*, pages 591–600. Springer Berlin Heidelberg, 1998. (Cited on page 37.)

[Sirrs 2007] Sandra M. Sirrs, Cornelia Laule, Burkhard Madler, Elana E. Brief, Sumia A. Tahir, Carole Bishop and Alex L. MacKay. *Normal-appearing White Matter in Patients with Phenylketonuria: Water Content, Myelin Water Fraction, and Metabolite Concentrations 1.* Radiology, vol. 242, no. 1, pages 236–243, 2007. (Cited on page 23.)

[Sneath 1957] Peter Henry Andrews Sneath. *The Application of Computers to Taxonomy.* Journal of General Microbiology, vol. 17, pages 201–226, 1957. (Cited on pages 33, 34 and 35.)

[Sørensen 1948] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biologiske Skrifter / Det Kongelige Danske Videnskabernes Selskab. I kommission hos E. Munksgaard, 1948. (Cited on page 33.)

[Steinhaus 1956] Hugo Steinhaus. *Sur la division des corp materiels en parties.* Bull. Acad. Polon. Sci, pages 801–804, 1956. (Cited on pages 32 and 92.)

[Taherkhani 2016] Mojtaba Taherkhani and Reza Safabakhsh. *A Novel Stability-based Adaptive Inertia Weight for Particle Swarm Optimization.* Applied Soft Computing, vol. 38, pages 281–295, 2016. (Cited on page 37.)

[Tan 2006] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to data mining. Pearson Addison Wesley, Boston, San Francisco (Paris), 2006. (Cited on pages 34 and 35.)

[Theodoridis 2008] S. Theodoridis and K. Koutroumbas. Pattern recognition, fourth edition. Academic Press, 4th édition, 2008. (Cited on pages 28, 30 and 32.)

[Tibshirani 2000] Robert Tibshirani, Guenther Walther and Trevor Hastie. *Estimating the number of clusters in a dataset via the Gap statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pages 411–423, 2000. (Cited on page 34.)

[Tintoré 2000] Mar Tintoré, Alex Rovira, Maria J Martínez, Jordi Rio, Pablo Díaz-Villoslada, Luis Brieva, Cecilia Borrás, Elisenda Grivé, Jaume Capellades and Xavier Montalban. *Isolated Demyelinating Syndromes: Comparison of Different MR Imaging Criteria to Predict Conversion to Clinically Definite Multiple Sclerosis.* American Journal of Neuroradiology, vol. 21, no. 4, pages 702–706, 2000. (Cited on page 9.)

[Tonkova 2010] Vyara Tonkova. *Erstellung von Quantitativen T_1, T_2^*- und Wassergehaltsatlanten aus 3T-MRT-Aufnahmen*, 2010. (Cited on pages 13 and 18.)

[Tonkova 2012a] Vyara Tonkova, Volker Arhelger, Jochen Schenk and Heiko Neeb. *Rapid Myelin Water Content Mapping on Clinical MR Systems.* Zeitschrift für Medizinische Physik, vol. 22, no. 2, pages 133–142, 2012. (Cited on pages 2 and 13.)

[Tonkova 2012b] Vyara Tonkova and Heiko Neeb. *Nuclear Potential Clustering as a New Tool to Detect Patterns in High Dimensional Datasets: Application to Quantitative MR Maps of Multiple Sclerosis.* In Proceedings of the 43rd Annual Meeting of the German Society for Medical Physics, 2012. (Cited on page 2.)

[Tonkova 2013] Vyara Tonkova, Dietrich Paulus and Heiko Neeb. *Nuclear Potential Clustering as a New Tool to Detect Patterns in High Dimensional Datasets.* In Journal of Physics: Conference Series, volume 410, page 012004. IOP Publishing, 2013. (Cited on page 2.)

[Trelea 2003] Ioan Cristian Trelea. *The Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection.* Information Processing Letters, vol. 85, no. 6, pages 317–325, 2003. (Cited on page 37.)

[Tsang 2006] Chi-Ho Tsang and Sam Kwong. *Ant Colony Clustering and Feature Extraction for Anomaly Intrusion Detection.* In Swarm Intelligence in Data Mining, pages 101–123. Springer, 2006. (Cited on page 37.)

[Tseng 2001] Lin Yu Tseng and Shiueng Bien Yang. *A Genetic Approach to the Automatic Clustering Problem.* Pattern Recognition, vol. 34, no. 2, pages 415–424, 2001. (Cited on page 36.)

[Van Den Bergh 2006] Frans Van Den Bergh. *An Analysis of Particle Swarm Optimizers.* PhD thesis, University of Pretoria, 2006. (Cited on page 37.)

[Van der Merwe 2003] D.W. Van der Merwe and Andries P. Engelbrecht. *Data Clustering Using Particle Swarm Optimization.* In The 2003 Congress on Evolutionary Computation, 2003. CEC'03., volume 1, pages 215–220. IEEE, 2003. (Cited on page 37.)

[Van Gelderen 2012] Peter Van Gelderen, Jacco A. De Zwart, Jongho Lee, Pascal Sati, Daniel S. Reich and Jeff H. Duyn. *Nonexponential T_2^* Decay in White Matter.* Magnetic Resonance in Medicine, vol. 67, no. 1, pages 110–117, 2012. (Cited on page 11.)

[Van Le 1995] T. Van Le. *Evolutionary Fuzzy Clustering.* In IEEE International Conference on Evolutionary Computation, 1995., volume 2, pages 753–758, 1995. (Cited on page 36.)

[Vavasour 2009] I.M. Vavasour, C. Laule, D.K.B. Li, J. Oger, G.R.W. Moore, A. Traboulsee and A.L. MacKay. *Longitudinal Changes in Myelin Water Fraction in Two MS Patients with Active Disease.* Journal of the Neurological Sciences, vol. 276, no. 1, pages 49–53, 2009. (Cited on pages 11, 15 and 23.)

[Vempala 2004] Santosh Vempala and Grant Wang. *A Spectral Algorithm for Learning Mixture Models.* J. Comput. Syst. Sci., vol. 68, no. 4, pages 841–860, 2004. (Cited on page 34.)

[Waltz 2006] Richard A Waltz, José Luis Morales, Jorge Nocedal and Dominique Orban. *An Interior Algorithm for Nonlinear Optimization That Combines Line Search and Trust Region Steps.* Mathematical Programming, vol. 107, no. 3, pages 391–408, 2006. (Cited on page 81.)

[Weinstein 2009] Marvin Weinstein and David Horn. *Dynamic Quantum Clustering: A Method for Visual Exploration of Structures in Data.* Physical Review E, vol. 80, page 066117, 2009. (Cited on page 38.)

[Weinstein 2013] Marvin Weinstein, F. Meirer, A. Hume, Ph. Sciau, G. Shaked, R. Hofstetter, Erez Persi, A. Mehta and David Horn. *Analyzing Big Data with Dynamic Quantum Clustering.* arXiv preprint arXiv:1310.2700, 2013. (Cited on page 38.)

[Woods 1954] Roger D. Woods and David S. Saxon. *Diffuse Surface Optical Model for Nucleon-Nuclei Scattering*. Physical Review, vol. 95, pages 577–578, 1954. (Cited on page 47.)

[Xiao 2003] Xiang Xiao, Ernst R. Dow, Russell Eberhart, Zina Ben Miled and Robert J. Oppelt. *Gene Clustering Using Self-organizing Maps and Particle Swarm Optimization*. In Parallel and Distributed Processing Symposium, 2003. Proceedings. International, pages 10–pp. IEEE, 2003. (Cited on page 37.)

[Xu 2005] Rui Xu and D.C. Wunsch. *Survey of Clustering Algorithms*. IEEE Transactions on Neural Networks, vol. 16, no. 3, pages 645–678, May 2005. (Cited on pages 30, 32, 33, 35 and 104.)

[Yuan 1995] B. Yuan, G.J. Klir and J.F. Swan-Stone. *Evolutionary fuzzy c-means clustering algorithm*. In Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int, volume 4, pages 2221–2226, 1995. (Cited on page 36.)

[Zadeh 1965] L.A. Zadeh. *Fuzzy sets*. Information and Control, vol. 8, no. 3, pages 338 – 353, 1965. (Cited on page 33.)

[Zhang 1996a] T. Zhang, R. Ramakrishnan and M. Livny. *BIRCH: an Efficient Data Clustering Method for Very Large Databases*. ACM SIGMOD Record, vol. 25, no. 2, pages 103–114, 1996. (Cited on page 131.)

[Zhang 1996b] Tian Zhang, Raghu Ramakrishnan and Miron Livny. *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD '96, pages 103–114, New York, NY, USA, 1996. ACM. (Cited on page 33.)

[Zhang 2004] Jiang-She Zhang and Yiu-Wing Leung. *Improved Possibilistic C-means Clustering Algorithms*. IEEE Transactions on Fuzzy Systems, vol. 12, no. 2, pages 209–217, 2004. (Cited on page 34.)