Multivariate Methods for Interpretable Analysis of Magnetic Resonance Spectroscopy Data in Brain Tumour Diagnosis

Albert Vilamala Muñoz

Computer Science Department

Universitat Politècnica de Catalunya

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)
in the subject of Artificial Intelligence

Under the supervision of

Dr. Alfredo Vellido Alcacena and
Dr. Lluís A. Belanche Muñoz

November 2015

Malignant tumours of the brain represent one of the most difficult to treat types of cancer due to the sensitive organ they affect. Clinical management of the pathology becomes even more intricate as the tumour mass increases due to proliferation, suggesting that an early and accurate diagnosis is vital for preventing it from its normal course of development. The standard clinical practise for diagnosis includes invasive techniques that might be harmful for the patient, a fact that has fostered intensive research towards the discovery of alternative non-invasive brain tissue measurement methods, such as nuclear magnetic resonance. One of its variants, magnetic resonance imaging, is already used in a regular basis to locate and bound the brain tumour; but a complementary variant, magnetic resonance spectroscopy, despite its higher spatial resolution and its capability to identify biochemical metabolites that might become biomarkers of tumour within a delimited area, lags behind in terms of clinical use, mainly due to its difficult interpretability. The interpretation of magnetic resonance spectra corresponding to brain tissue thus becomes an interesting field of research for automated methods of knowledge extraction such as machine learning, always understanding its secondary role behind human expert medical decision making. The current thesis aims at contributing to the state of the art in this domain by providing novel techniques for assistance of radiology experts, focusing on complex problems and delivering interpretable solutions. In this respect, an ensemble learning technique to accurately discriminate amongst the most aggressive brain tumours, namely glioblastomas and metastases, has been designed; moreover, a strategy to increase the stability of biomarker identification in the spectra by means of instance weighting is provided. From a different analytical perspective, a tool based on signal source separation, guided by tumour type-specific information has been developed to assess the existence of different tissues in the tumoural mass, quantifying their influence in the vicinity of tumoural areas. This development has led to the derivation of a probabilistic interpretation of some source separation techniques, which provide support for uncertainty handling and strategies for the estimation of the most accurate number of differentiated tissues within the analysed tumour volumes. The provided strategies should assist human experts through the use of automated decision support tools and by tackling interpretability and accuracy from different angles.

To my little Xènia, my wonderful wife Nataliya, my great parents Albert and Rafi; and my lovely sister Raquel.

Acknowledgements

First and foremost, I would like to thank my advisors, Alfredo Vellido and Lluís A. Belanche. They have always found a slot in their busy agendas whenever I needed some help, showing me the path to follow when I got lost in my research and providing new avenues to explore and discuss.

I would like to extend this gratitude to the rest of the Soft Computing research group (SOCO), especially to Ángela Nebot, Francisco Mugica, Enrique Romero and René Alquezar. They made me feel like we were a big family, sharing not only professional thoughts, but also personal events; an example of such are the annual gatherings at Can Nebot.

Next, I want to express my appreciation to Paulo Lisboa, who opened the door of the Liverpool John Moores University, allowing me to spend three fruitful months in his department. I had the chance to meet great people there: Terence Etchells, Héctor Ruiz, Simon Chambers and Vincent Kwasnica. The long conversations at both morning coffee and on Thursdays' evenings are memorable. I want to emphasize the role acquired by Ian Jarman during my visit there, becoming my mentor to ensure that I had a smooth integration to their culture. Many thanks, Ian!

Pursuing a doctorate is a long ride, often finding difficulties on the way that need to be overcome. All those shortcomings are better tackled if you are surrounded by the right people. In this respect, I have been lucky to count on my friends at the Ω-S1 floor. The somehow *long-term people* include Carles Creus, Eva Martínez, Maria Àngels Cerveró, Jesús Ojeda, Jorge Muñoz, Alessandra Tosi, Josep Lluís Berral, Javier de San Pedro, Alberto Moreno, Daniel Alonso, Ramon Xuriguera, Àlex Vidal, Adrià Gascón, Sergi Oliva, Solmaz Bagherpour, Álex Álvarez, Alberto Gutiérrez, Joel Ribeiro, Nikita Nikitin, Pedro Hermosilla, Isaac Besora, Andreu Mayo, Jaume Pujantell, Hendrik Molter and Laura Mascarell.

The last year of this thesis has been quite difficult: coupling job duties with writing the manuscript, together with family matters and pater-

nity might take part of your energy. Fortunately, I have been working in a great company, with great people conforming the Strategy, Business Analysis and Business Intelligence departments of Schibsted Spain. Specifically, I want to thank my managers Borja de Muller and Laura Lara for their patience regarding my never-ending thesis.

Understanding and developing the Bayesian part of this thesis has been a big challenge for me, which has finally been achieved with success. Part of this merit belongs to the help supplied by Jesús Cerquídes from Institut d'Investigaciò en Intelligència Artificial-CSIC and Mikkel Schmidt from the Technical University of Denmark, who switched on the light for me to see how to proceed on this topic. Also relevant, was the trip to the RecSys conference in Vienna with the Data Driven Solutions team, where a proper derivation was carried out thanks to the help of Javier Roldán and Daniel Abril. There are pictures capturing that moment!

Finally, I would like to thank my family, who are the ones who suffered me and my change of mood according to the results obtained in my research. Especially, to my wife Nataliya: this thesis would have not been possible without your help and patience; my daughter Xènia, who I want to apology for stealing some of her *play time with daddy*. Noteworthy has also been the unconditional support from my parents Albert and Rafi, as well as my sister Raquel, and my grandparents Manel, Maria, Juan and Rafaela. A special consideration towards my family in law Anatolii and Galyna, for always finding me a comfortable place to work.

Contents

List of Figures

List of Tables

List of Acronyms

*ac*2 Astrocytoma Grade II.

gbm Glioblastoma.

met Metastasis.

nom Normal cerebral tissue, white matter.

ACC Accuracy.

ACGT Advancing Clinico Genomic Trials on Cancer.

AI Artificial Intelligence.

ANN Artificial Neural Networks.

ASSIST Association Studies Assisted by Inference and Semantic Technologies.

AUC Area Under the ROC Curve.

AUH Area Under the Convex Hull of the ROC Curve.

BAC Balanced Accuracy.

BEL Breadth Ensemble Learning.

BER Balanced Error Rate.

BSD Bayesian Spectral Decomposition.

BSS Blind Source Separation.

CART Classification and Regression Trees.

CBTRUS Central Brain Tumor Registry of the United States.

CDP Centre Diagnòstic Pedralbes.

CGS Consensus Group Stable.

CNMF Convex Non-negative Matrix Factorisation.

CNS Central Nervous System.

COR Pearson Linear Correlation.

CSI Chemical-Shift Imaging.

CT Computerised Tomography.

CV Cross Validation.

DCNMF Discriminant Convex Non-negative Matrix Factorisation.

DGF Dense Group Finder.

DNA Deoxyribonucleic Acid.

DRAGS Dense Relevant Attribute Group Selector.

DT Decision Trees.

EM Expectation-Maximisation.

EMRS Expectation-Maximisation using Reconstructed Sources.

ER Error Rate.

ESPS Exact Structure Preservation Strategies.

FE Feature Extraction.

FN False Negative.

FP False Positive.

FPR False Positive Rate.

FSS Feature Subset Selection.

GABRMN Grup d'Aplicacions Biomèdiques de la Ressonància Magnètica Nuclear.

GRSI Grup de Recerca de Sistemes Intelligents.

IBIME Informática Biomédica.

ICA Independent Components Analysis.

ICT European Commission Information and Communication Technologies.

IDI Institut de Diagnòstic per la Imatge.

LIST OF ACRONYMS

INTERPRET International Network for Pattern Recognition of Tumours Using Magnetic Resonance.

ITACA Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas.

KI Kuncheva Index.

KKT Karush-Kuhn-Tucker.

LDA Linear Discriminant Analysis.

LOO Leave One Out.

LS-SVM Least Squares Support Vector Machines.

LTE Long Time of Echo.

MAP Maximum A Posteriori.

MBIW Margin Based Instance Weighting.

MCCS Model Component Combination Strategies.

MCMC Markov Chain Monte Carlo.

MDSS Medical Decision Support System.

MF Matrix Factorisation.

MH Metropolis Hastings.

ML Machine Learning.

MLPM Machine Learning for Personalized Medicine.

MMAS Multiple Model Aggregation Strategies.

MR Magnetic Resonance.

MRI Magnetic Resonance Imaging.

MRS Magnetic Resonance Spectroscopy.

MRSI Magnetic Resonance Spectroscopy Imaging.

MV Multi Voxel.

MVFS Margin Vector Feature Space.

NAA N-Acetyl Aspartate.

NB Naive Bayes.

NMF Non-negative Matrix Factorisation.

NMR Nuclear Magnetic Resonance.

NN Nearest Neighbours.

NPCR National Program of Cancer Registries.

PBCNS Primary Brain and Central Nervous System.

PC Principal Component.

PCA Principal Components Analysis.

PET Positron Emission Tomography.

PNET Primitive Neuroectodermal.

PPV Positive Predictive Value.

PRESS Point-Resolved Spectroscopy.

QDA Quadratic Discriminant Analysis.

RecSys Recommender Systems.

RF Random Forests.

RFE Recursive Feature Elimination.

RLIW Recursive Logistic Instance Weighting.

ROC Receiver Operating Characteristic.

ROI Region of Interest.

RS Reconstructed Sources.

RTICC Red Temática de Investigación Cooperativa en Centros de Cáncer.

SBG Sequential Backward Generation.

SEER Surveillance, Epidemiology and End Results.

SFG Sequential Forward Generation.

SFSA Sequential Feature Selection Algorithm.

SGHMS Saint George's Hospital Medical School.

SNMF Semi Non-negative Matrix Factorisation.

SOCO Soft Computing Research Group.

STE Short Time of Echo.

STEAM Stimulated Echo Acquisition Mode.

SV Single Voxel.

SVM Support Vector Machines.

TE Time of Echo.

TN True Negative.

TNR True Negative Rate.

TP True Positive.

TPR True Positive Rate.

UAB Universitat Autònoma de Barcelona.

UK United Kingdom.

UMCN University Nijmegen Medical Centre.

UPC Universitat Politècnica de Catalunya.

UPV Universitat Politècnica de València.

USA United States of America.

WHO World Health Organisation.

LIST OF ACRONYMS

Chapter 1

Introduction

According to a report published by the World Health Organisation (WHO) [1] in February of 2015, cancer is a leading cause of death worldwide. In 2012, 8.2 million people passed away due to this condition, lung cancer (accounting for 1.59 million deaths), liver cancer (754,000 deaths), stomach cancer (723,000 deaths), colorectal cancer (694,000 deaths), breast cancer (521,000 deaths) and esophageal cancer (400,000 deaths) being its most frequent types in terms of cause of decease. More importantly, far from diminishing, these numbers are expected to rise in the following years reaching up to a predicted 13.1 million deaths in 2030.

Although 70% of all deaths linked to cancer occur in low- and middle-income countries, mainly due to their difficulty to deliver proper treatment to their patients, high-income countries are also affected by this disease and poor prognosis cannot be avoided for certain types. Only in Catalonia, more than 33,700 new cancer cases were annually diagnosed during the period between 2003 and 2007 [2]. In fact, it is estimated that 50% of men and 33% of women will develop cancer at some point throughout their life. In 2004, cancer was the first cause of death for males (33.55% of total deaths) and the second one for women (22.02% of total deaths), just surpassed by deaths caused by pathologies of the circulatory system. In a study published in 2008 [3], the projections foresaw a stabilisation in the diagnosis and a decrease in its related mortality by 2015.

1. INTRODUCTION

Tumours of the Central Nervous System (CNS) and, particularly, brain tumours are an especially challenging type of cancer, given the poor prognosis associated to some of their subtypes. The Central Brain Tumor Registry of the United States (CBTRUS) [4] estimated the prevalence of this pathology to be 221.8 per $100,000$ inhabitants in 2010, meaning that around $688,000$ were living in the United States with a diagnosis of Primary Brain and Central Nervous System (PBCNS) tumour that year, of which more than 20% were malignant.

This is a relatively low prevalence, but, unfortunately, several types of brain tumour have a very poor prognosis associated. The Surveillance, Epidemiology, and End Results (SEER) program [4] estimated a five year relative survival rate of 32.6% for males and 35.3% for females, following diagnosis of a malignant PBCNS tumour, using data between years 1995 and 2011 in the USA.

Early and accurate diagnosis of tumour proliferation can decrease the mortality rates as well as improve the quality of life for these patients by means of providing the proper treatment in order to cure the disease or palliate its effects. This need for accurate diagnosis lays the foundations of the current thesis, which aims to provide semi-automated computer-based decision support for expert radiologists in whom ultimately the final diagnostic decision making resides.

The most reliable procedures doctors currently use for evaluating masses of uncontrolled cell proliferation in intracranial regions involve invasive techniques, such as the biopsy (the current gold standard in the field), which consists in extracting a sample from the tissue of interest and performing a histopathological study in a laboratory so as to provide accurate diagnosis and prognosis.

The application of invasive techniques is often harmful for the patient, who undergoes surgery with uncertainty about collateral damage that this clinical procedure might induce to the patient's cognitive abilities, with the non-negligible probability of getting severely impaired, depending on the region of the brain the tumour is located. This strong inconvenience has led biomedical engineers to design, and physicians to use over the last decades,

alternative non-invasive indirect measurement techniques to harmlessly inspect the affected mass.

Radiology can indeed play an important role in the discrimination between brain tumour types. The diagnoses of some types of tumour are not always obvious from conventional Magnetic Resonance Imaging (MRI), as images from different tumours may be too similar for discrimination. Further diagnostic support can be obtained from the so-called physiological Magnetic Resonance (MR) techniques. Most of them use the infiltrative pattern of growth of tumours to accomplish the diagnostic differentiation. These techniques include perfusion MR and diffusion MR. Alternative techniques include two-dimensional Turbo Spectroscopic Imaging information [5], Diffusion Tensor Imaging [6, 7] and multiple-voxel Magnetic Resonance Spectroscopy (MRS) with 2D Chemical-Shift Imaging (CSI) and peak amplitude ratios [8]. Recent studies have also resorted to Morphometric Analysis of MR Images [9].

Among the most matured techniques in radiological practise are those relying on the resonance of certain chemical nuclei present in human tissue under magnetic fields, the already mentioned MRI and its MRS counterpart.

Making sense of the complexity of the data that MRS yields is far from being a trivial matter, even for expert radiologists. This has led, in recent years, to the search for alternative answers from the field of pattern recognition and multivariate statistics. The interest in these fields is also related to the possibility of designing at least semi-automated, computer-based Medical Decision Support Systems (MDSS) to facilitate radiologists' task and ease the interpretation of results [10, 11]. The current thesis aims at contributing to the area by developing new techniques aligned to these needs.

Over the last decade, European-funded research has focused on the problem of automated diagnosis and prognosis for oncology. The European Commission Information and Communication Technologies (ICT) for Health Unit of the Information Society and Media Directorate General managed several international research projects in the medical ambit, funded under

the Sixth Framework program (FP6). Within the program's 4^{th} call, *Integrated biomedical information for better health*, several projects concerned cancer research. All these projects involved data analysis, and some of them realised it through Data Mining or Computational Intelligence methods, often related to Machine Learning (ML). They included Advancing Clinico Genomic Trials on Cancer (ACGT) [12], which aimed to fill-in the technological gaps of clinical trials for two pathologies: breast cancer and paediatric nephroblastoma, and used data mining tools and the R statistically-oriented programming language in a grid environment; ASSIST [13], which aimed to provide medical researchers of cervical cancer with an environment that will unify multiple patient record repositories; and the Computational Intelligence for Biopattern Analysis in Support of eHealthcare (Biopattern) [14], whose goal was to develop a pan-European, intelligent analysis of a citizen's bioprofile and to exploit this bioprofile to combat major diseases such as ovarian, breast and brain cancers, leukaemia and melanoma.

More recent European projects in the field, all part of FP7, include ML for Personalized Medicine (MLPM) [15], a Marie Curie Initial Training Network for the pre-doctoral training of scientists in research at the interface of ML and medicine; Epigene Informatics [16], a Marie Curie action to investigate ML approaches to epigenomic research; and Metoxia (Metastatic Tumours Facilitated by Hypoxic Tumour Micro-Environments), a project for the analysis of metastatic tumours facilitated by hypoxic tumour microenvironments that includes the development of a ML-based classifier of tumour hypoxia [17].

Research efforts at the European level have also specifically focused on the use of pattern recognition for the analysis of brain tumours from MRS data. An example of it is the International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) [18] project (2000-2002), whose main objective was to facilitate the use of MRS into clinical routine by first constructing a large European database of standardised brain tumour spectra and clinical data; and secondly, using this database to build a user-friendly computer program for spectral classification.

Another example is eTumour [19] (2004-2009), in which a web accessible MDSS for brain tumour diagnosis and prognosis was developed, incorporating *in vivo* and *ex vivo* genomic and metabolomic data.

Finally, the HealthAgents project [20] (2006-2008) sought to develop an agent-based distributed MDSS to assist in the early diagnosis and prognosis of brain tumours. Parallel to it, a distributed data-warehouse was built, becoming the world's largest database of clinical, histological and molecular phenotype data for brain tumours.

These three projects became a milestone in the research of tumour diagnosis using MRS data, providing researchers with a sizeable and standardised set of spectra, from which useful and actionable knowledge could be extracted.

In Spain, research in the area include the Red Temática de Investigación Cooperativa en Centros de Cancer (RTICC), specialised in bioinformatics, biostatistics and image-base diagnosis; The Grupo de Redes Neuronales at Universidad de Extremadura, who, together with Servicio Extremeño de Salud developed the MAMMODIAG project [21] for the support in the diagnosis of breast cancer; the Grup de Recerca de Sistemes Intel·ligents (GRSI) at Universitat Ramon Llull in Catalonia and their HRIMAC project [22], which uses Artificial Intelligence (AI) techniques for the support in the breast cancer diagnosis.

The Grupo de Informática Biomédica (IBIME) at Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA) located at the Universitat Politècnica de València (UPV) has also participated in several European research projects aiming at produce integrated software solutions for biomedical problems, gaining high expertise in AI tools applied to build MDSS.

The current thesis is linked to the research project *AIDTumour: Herramientas Basadas en Métodos de Inteligencia Artificial para el Apoyo a la Decisión en Oncología*, led by the Soft Computing (SOCO) research group at Universitat Politècnica de Catalunya (UPC) in Barcelona. Several related theses also became background for the current proposal. F. González-Navarro [23] developed novel feature selection techniques for cancer diag-

nosis from microarray gene expression as well as ^{1}H-MRS data of brain tumours. C.J. Arizmendi's thesis [24] involved the development of advanced signal processing techniques and their application to signal pre-processing for ^{1}H-MRS. In her PhD thesis, S. Ortega-Martorell [25] developed new interpretable feature selection and extraction techniques for aiding in single-voxel (SV) ^{1}H-MRS brain tumour diagnosis. Moreover, she also used multi-voxel (MV) ^{1}H-MRS for the delimitation of the tumour pathological area. All this research was included in a made-to-measure software tool to be used by physicians [26]. This tool was used as a starting point for the current thesis, which gave us direct insights on the real applicability of last developed techniques, at the same time as we contributed back to the system by including new functionality.

The present document provides a detailed and thorough account of our research on the improvement of the state of the art in multivariate data analysis techniques specifically designed for the accurate diagnosis of the most aggressive conditions in neuro-oncology, therefore aiming to make these techniques trustworthy and interpretable for their use by radiology medical experts.

1.1 Contributions

Despite the fact that much research has already been carried out in the field of Pattern Recognition and ML as applied to the analysis of MRS information in neuro-oncology problems [24, 25, 23], a number of problems still remain unsolved or have not been yet fully addressed. These are research challenges to which this thesis aims to contribute by improving state of the art techniques to aid in the diagnosis of brain tumours using SV-^{1}H-MRS data.

More precisely, all the work presented in this document has the ultimate goal of contributing to the following important areas in neuro-oncology: the first consists in making progress on the diagnostic tools when presented with aggressive brain tumours, under the assumption that specific biomarkers coinciding with specific frequencies within the MR spectrum exist. The

second broadens the object of study by focusing on the influence of different tissues present in the vicinity of most common brain tumoural areas.

In the following lines, we explicitly state the different objectives that have been pursued in the development of new techniques in the aforementioned subareas; the challenges that previous solutions were unable to overcome and, finally, the contributions that our novel techniques provide to the medical domain and to the community at large.

1.1.1 Discrimination between aggressive brain tumours using the biomarker paradigm

1.1.1.1 Study 1

Goal The goal of this research is to increase the discriminative power of current analytical tools in terms of their ability to discriminate between the most aggressive brain tumour types, namely *glioblastomas* and *metastases*.

Challenges The main difficulties that current techniques face when discriminating between these tumour types are summarised next:

- High intra-class dissimilarity, meaning that the recorded MR spectra of patients sharing the same tumour type pathology are very different one to another.
- High inter-class similarity, which can be explained as the great degree of resemblance that can be encountered between two spectra, each related to different tumour types.
- Few frequencies contribute to the discrimination: out of the large amount of measurements performed in an MR scanning, only a small quantity of metabolites might be related to the tumour type being investigated.

Limitations of current solutions

- From an application domain point of view, no attempt to use a mixture of learners able to capture the heterogeneity of spectra has been made.

- The technical reasons for such lack of applicability can be traced back to the fact that classical ensemble techniques strongly rely on the random selection of features (i.e., frequencies or metabolites in our domain), which lead to suboptimal solutions given the nature of the data being analysed.

Research contributions For the current domain of application, reaching the pursued goal is by itself a contribution. From an ML perspective, the contributions of the current study are:

- The development of a novel ensemble learning technique able to subdivide the input space according to the most adequate feature subsets, in which specialised classifiers are built, leading to the maximisation of the overall discrimination accuracy.
- The derivation of an embedded feature selection strategy specifically designed to address the few frequency contribution challenge.

1.1.1.2 Study 2

Goal The second goal of our research is to increase the interpretability of the results provided by our techniques by finding more reliable metabolites which attribute to tumour-type discriminative power.

Challenges The principal problems found when aiming at reaching this goal are related to the concept of algorithm instability, which is defined as the incapacity to provide similar solutions over executions when small variations in the input are present. Three factors contribute to this phenomenon:

- Small sample size: when the amount of patients to be studied is very low to obtain statistically significant results.
- High dimensionality; that is, when the number of spectral measurements is very large as compared to the sample size.
- Redundancy of frequencies is a property that arises when the same piece of information is provided by several different features.

Limitations of current solutions

- To the best of our knowledge, no previous attempt to deal with the current goal has been proposed in our domain.
- The need for a brand-new feature selection stability technique can be explained by analysing the available tools, in which resampling techniques are the prevalent solutions. They present a high computational cost, added to the already existing pitfalls that are implied when sampling from a low number of samples.
- Finally, the limitations found when studying current importance-weighting solutions were an invitation to contribute to this field.

Research contributions

- A strategy to weight instances according to their typicalness is defined.
- An algorithm able to select the most adequate features for a specific task maintaining certain degree of stability is designed.

1.1.2 Diagnosis of most common brain tumours using the mixture of tissues paradigm

1.1.2.1 Study 3

Goal Identify the most relevant tissue types in a voxel contributing, in varying degrees, to the measured MRS signal; by exploiting prior knowledge on the nature of the signal they generate, as well as tissue-specific properties.

Challenges In-depth analysis of the available data motivates a paradigm shift on the research in order to approach the aforementioned goal from a different perspective. Challenges we face in this new setting include:

- The measured signal in an MR scanning can not be attributed to a single phenomenon, but to the interrelation of multiple sources due to the coexistence of several tissue types within a voxel, or to interferences from neighbouring ones.

- The possibility that the measurements contain coherent negative values that need to be unavoidably dealt with.
- The necessity to assess the contribution of each source of signal to every frequency measurement.
- Appropriately incorporate tissue-specific knowledge to increase the technique's capabilities.

Limitations of current solutions

- Previous studies in the current domain of application have shown the potential of Non-Negative Matrix Factorisation (NMF) techniques to successfully address most of the challenges we introduced. However, no study has used the available tissue-specific information to better extract the signal generating sources and their contributions.
- Technical barriers in the form of lack of algorithms able to apply Convex NMF (CNMF) from a supervised perspective are behind the inability to incorporate tissue-specific knowledge in our domain.

Research contributions

- An algorithm able to extract relevant source information from SV-^{1}H-MRS data and their contribution to the final measured signal is derived.
- This algorithm is, moreover, able to deal with both positive and negative values present in the signal generating sources.
- Interpretable degree of contribution from each source is also obtained as a byproduct.
- The ratios among metabolite values are preserved.
- Quality of all the above is improved by including tissue-specific information.

1.1.2.2 Study 4

Goal Automatic determination of the most appropriate number of tissues making up the MR signal for each specific pathology is the primary object of this last study. On the way, a mechanism to evaluate the certainty on the predictions is also sought.

Challenges Care must be taken when carrying out this research, since, besides all the difficulties stated in the previous study, we must add to the list:

- The estimation of the most adequate number of relevant tissues usually becomes a tedious and time-consuming process.
- Confidence on the predictions is undermined by the small number of subjects available in our data sets.
- Another consequence of the small data size is the possible occurrence of the overfitting phenomenon in the learning process.

Limitations of current solutions

- To the best of our knowledge, no study provides a probabilistic description of NMF in the domain of brain tumour signal separation from SV-^{1}H-MRS data able to respect all the constraints imposed by these data.
- The main rationale explaining such failure lays in the fact that out-of-the-box probabilistic NMF solutions are not able to address the singularity of data being employed, such as the evidence of sources showing positive and negative values and the constraints stating that their contributions must be positive.

Research contributions

- The obtained probabilistic solution is able to deal with positive and negative values.

- An efficient automatic selection of most appropriate number of tissues explaining the majority of the obtained signal is derived.
- A measure of confidence on the prediction is readily available as a byproduct of the whole process.
- Automatic control of regularisation is supplied, relegating overfitting to a rare phenomenon.
- Prior domain knowledge on both sources and contributions can explicitly be used to improve the results.
- The impact of parameter initialisation in terms of the obtained solution is diminished, since local minima are avoided.

1.2 Overview of the thesis

This thesis is structured in 8 chapters, the remaining of which are organised as follows:

Chapter 2 gives a brief introduction to the neuro-oncology domain, presenting the most frequent tumoural pathologies associated with the brain, the regular-practise diagnostic tools and the most common forms of treatment. Then, the applicability of various Nuclear Magnetic Resonance products as non-invasive techniques for tumour diagnosis is shown. Finally, a characterisation of the analysed biomedical data sets is also provided.

Chapter 3 intends to provide a general outlook on the large variety of technical strategies that are addressed throughout the thesis. Specifically, we glance at the broad domain of ML, focusing our attention on the concept of Ensemble Learning, typical dimensionality reduction approaches and different forms of instability which learning algorithms might suffer from. The last technical block corresponds to a gentle and self-contained introduction to the Bayesian framework of inference. The chapter ends with some examples of ML applications in the domain of brain tumour diagnosis.

Chapter 4 consists in a more in-depth analysis of Ensemble Learning theory, reviewing the fundamental parts that any algorithm of this kind

must be composed of, as well as discussing the rationale behind putting this type of approaches in place for the current domain. This is followed by a thorough explanation of our *Breadth Ensemble Learning* algorithm, specifically tailored to deal with the difficult discrimination of the most aggressive brain tumour types. Its suitability is assessed and the predefined hypotheses validated.

Chapter 5 analyses the stability phenomenon of Feature Selection algorithms. More precisely, it starts by showing the nature of our current domain data which directly affects the stability of the learning algorithms being used, as well as the limitations of available solutions. Then, stability measures and contemporary feature selection algorithms are reviewed, together with previous attempts to correct instability in Feature Subset Selection (FSS). Next, our proposal, named *Recursive Logistic Instance Weighting*, is introduced and evaluated to match the initial hypotheses.

Chapter 6 dives into the source separation problem. In particular, NMF variants are presented as suitable techniques to identify the different tissue types coexisting in a voxel, as well as their contribution to the retrieved MR signal. *Discriminant Convex Non-Negative Matrix Factorisation* (DCNMF) is derived and validated as a supervisedly-improved version of the CNMF, the most prominent technique in our domain.

Chapter 7 shifts the point of view from classical approaches of NMF towards a Bayesian interpretation of these techniques, incorporating the added value that such framework provides to our domain of application. First part of the chapter is a journey from frequentist to Bayesian paradigms as applied to Matrix Factorisation. Thereafter, our contribution in the form of *Probabilistic CNMF* and *Bayesian Semi Non-negative Matrix Factorisation* (SNMF) is laid out and their applicability to the neuro-oncology domain corroborated.

Finally, **Chapter 8** summarises the progress made by this thesis in the neuro-oncology domain, providing some discussion on those aspects that require additional attention; it also states some concluding remarks and paves the way for further improvement in the data-driven diagnostic tools

for neuro-oncology area. Moreover, a list of publications emanating from the research carried out during this thesis is also supplied.

Chapter 2

Medical background and materials

The current chapter aims at providing some up-to-date foundations in the medical field of diagnosis and insights about treatment techniques regarding the unfortunate event of tumoural tissue proliferation in the brain.

We start it by summarising some notions about the composition of the brain and introducing the most prevalent tumour types of the CNS. Then, different techniques for tumour diagnosis are briefly presented, before introducing the standard forms of treatment. This is followed by some fundamentals about brain tissue information acquisition through Nuclear Magnetic Resonance Spectroscopy, which is the focus of this thesis, and its use as a diagnostic tool from its analysed output. The chapter ends with a review of this data output that will be the basis of the analyses reported in the experimental chapters of the thesis.

2.1 Some fundamentals of neuro-oncology

Human beings are composed of different types of tissues, each of them suited to the function it has to perform. They are, in turn, made up of small entities called cells. There are also distinct types of cells according to the task they are entrusted with within a living body.

Despite differences among cells, most of them repair themselves and reproduce in a similar way. The latter is accomplished by dividing themselves in a controlled manner (using, for instance, processes of mitosis or meiosis).

However, and for a variety of reasons, these processes may be corrupted and the cells can end up reproducing in an uncontrolled fashion, leading to the development of tumour pathologies. If the tumour does not spread into the surrounding tissues, we are facing a benign tumour. Otherwise, if the tumour invades surrounding tissues (i.e., proliferates), it becomes a malignant tumour.

In the specific case of brain tumours, we can differentiate between primary, which have their origin in the brain, and secondary, which, having originated in other parts of the body (e.g., lungs, kidneys, colon, etc.), spread to the brain.

The WHO has defined standards for diagnosing and managing the treatment of brain tumours worldwide. They have published a system of grading the malignancy of different brain tumours, known as the WHO grading of the CNS. In its last revision, dating from 2007 [27], they define four categories of malignancy according to multiple histologic features:

- Grade I: Lesions with low proliferative potential and the possibility of cure following surgical resection alone.
- Grade II: Neoplasms generally infiltrative in nature and, despite low-level proliferative activity, often recur, sometimes progressing towards higher levels of malignancy.
- Grade III: Lesions with histological evidence of malignancy.
- Grade IV: Cytologically malignant, mitotically active, necrosis-prone neoplasm typically associated with rapid pre- and postoperative disease evolution and a fatal outcome.

2.1.1 Some basics about the brain

The brain is a mass of soft tissue that controls the activity of the other organs of the body. It is protected by the bones that form the skull in the

outer part and by a three-layered protective envelope called the meninges in the inner part (Figure 2.1).

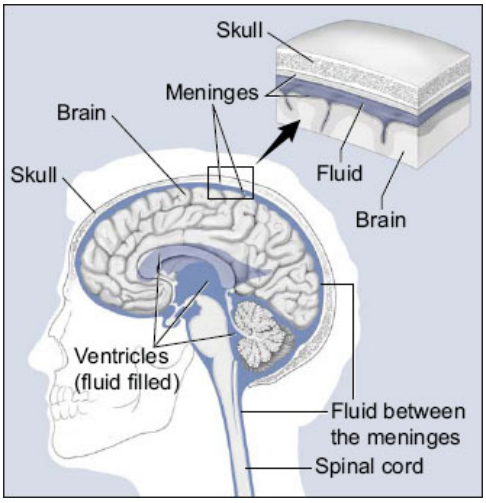


Figure 2.1: The brain and its surrounding structures - [28]

Three main structures make up the brain: the cerebrum, the cerebellum and the brainstem (Figure 2.2). The largest one is the cerebrum which is responsible of the high-level cognitive functions such as learning, memory, attention, sensory processing and motor control, amongst others. It is divided into two halves: the left hemisphere, that controls the movement of the right part of the body, and the right hemisphere which controls the opposite side. The cerebellum is in charge of balance, coordination and other complex semi-autonomous functions. The brainstem is the oldest part of the brain from an evolutionary point of view; it connects the brain with the spinal cord. Its tasks are of vital importance for maintaining the body alive; they include, for instance, the control of breathing, blood pressure, or blinking.

As in any other organ of the body, the brain tissue is made up of cells. Nearly 40 billion interconnected nerve cells or neurons form a complex network which conveys information back and forth in the form of electrical impulses and chemical signals. These neurons are fixed in place by the aid of other cells called glial. Different types of glial cells (e.g., astrocytes,

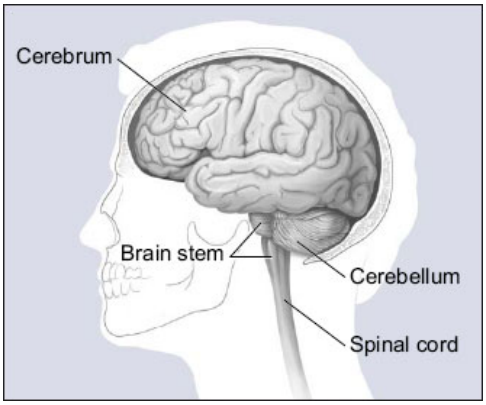


Figure 2.2: Main parts of the brain - [28]

oligodendrocytes or ependymocites) are often the origin of the most frequent tumour types.

2.1.2 Most common tumours of the Central Nervous System

The WHO recognises more than 120 different tumour types affecting the CNS. They are classified into seven categories according to the tissue that has originated the neoplasm [27]. For the sake of brevity, we just name the seven categories and introduce the most common ones. Details of the prevalence of each tumour type, according to a statistical study developed during the years 2006-2010, and the malignancy of each type according to WHO grading can be found in [29]. A fairly detailed distribution of brain tumours according to their histological origin can be seen in Figure 2.3.

Category 1 groups all those tumours originated in the *neuroepithelial tissue*. Also known as *astrocytic tumours* or *gliomas*, they constitute the most frequent tumour types present in adults, accounting for 31.2% of all primary CNS tumours. Among them, we can differentiate between *astrocytomas* (6.1% – grade I, II and III), *glioblastomas* (15.6% – grade IV), *oligodendrogliomas* (1.7% – grade III), *medulloblastomas* and *primitive neuroectodermal* (PNET) (1.2% – grade IV), and *ependymomas* (1.9% – grade II).

Category 2 contains those types whose source lie in the *cranial and paraspinal nerves* (8.1%). The most frequent tumours in this group are

benign *schwannomas* (grade I) and *nerve sheath tumours* (grade II, III and IV).

Category 3 deals with tumours affecting the *meninges*, *meningiomas* (35.8% – grade I) being a highly frequent type.

More rare tumour types can be found in categories 4 and 5 relative to the *haematopoietic system* (such as *lymphoma*, 2.1% – grade IV) and *germ cell tumours* (0.4% – grade IV), respectively.

Category 6 includes tumours in the *sellar region* (*pituicytoma*, 14.7% – grade II and *craniopharyngioma*, 0.8% – grade II).

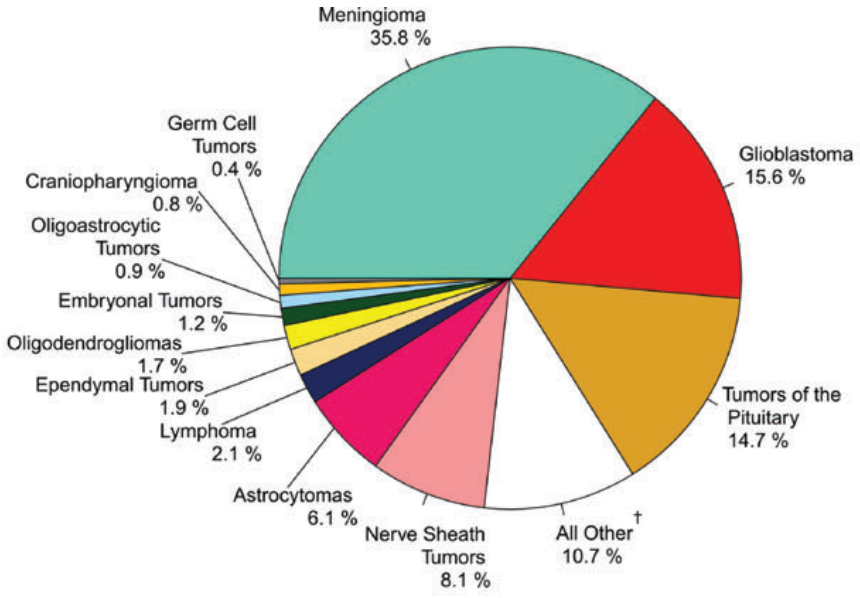


Figure 2.3: Distribution of Primary Brain and Central Nervous System tumours by histology - CBTRUS Statistical Report: NPCR and SEER Data from 2006-2010. N = 326,711 patients. [29]

Category 7 includes *metastatic* tumours, conforming the group known as secondary. This kind of tumours are the ones presenting the highest incidence rate; the breast, lung and melanoma cancers being the most likely sources to spread the tumour to the brain.

Neoplasms of the CNS can also be classified according to different criteria. Figure 2.4 shows the affectation of primary with respect to the region of the brain they raised from.

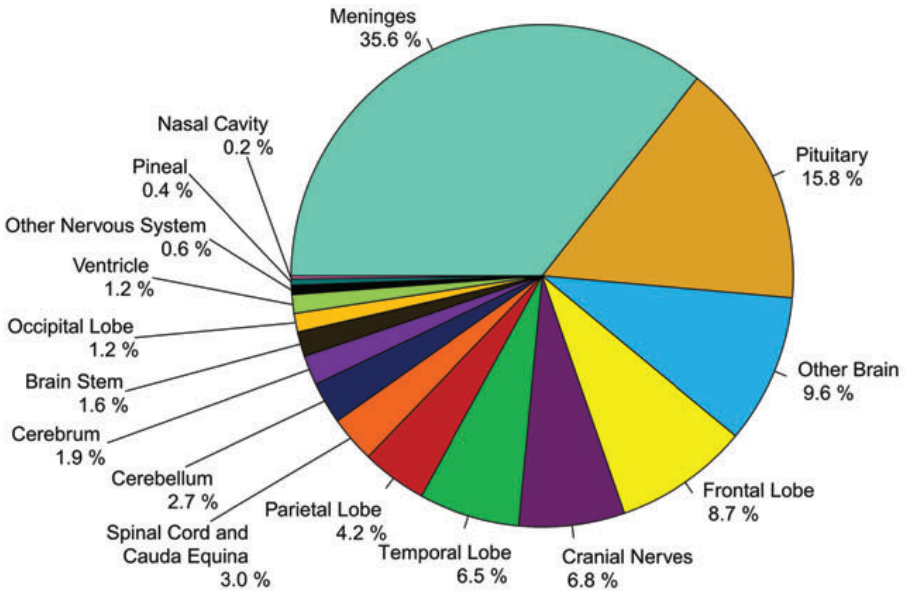


Figure 2.4: Distribution of Primary Brain and Central Nervous System tumours by brain region - CBTRUS Statistical Report: NPCR and SEER Data from 2006-2010. N = 326,711 patients. [29]

2.1.3 Tumour diagnosis

A tumour growing in the brain often increases the pressure within the skull, inducing a set of different effects. Among the most frequent are headaches, sickness, nausea or even seizures. A range of other symptoms depend on the affected part of the brain. However, pressure is not the only cause of symptoms, since for those tumours of invasive nature, the own damage of the tissue also contributes.

Once any of these symptoms is present, the patient should visit a specialist (a neurologist, an oncologist, or a radiologist), who will carry out tests, often using non-invasive techniques, to determine the presence or absence of a tumour, and, in the former case, its type and malignancy.

It is of vital importance to correctly assess the tumour's characteristics at this stage, because the treatment and prognosis of a tumour highly depends on them and varies according to its profile.

Biopsy The *gold standard* and most reliable test for the determination of the tumour type and level of malignancy is the biopsy. A biopsy is a surgical operation where a small piece of the tumour tissue is removed from the brain in order to be examined in a laboratory. The brain is accessed by performing a small hole in the skull with the purpose of introducing a fine needle that removes the targeted sample of tissue.

In certain cases (i.e., when the tumour is deep inside the brain) a specific kind of biopsy might be performed. In a guided biopsy (e.g., stereotactic biopsy, neuronavigation) the procedure is pretty similar to an ordinary biopsy with the difference that, in such cases, imaging technologies are used to help guiding the needle.

Despite the reliability of this test, it presents a blatant drawback, which is the non-negligible risk involved in the manipulation of such a sensitive organ as the brain is. Therefore, alternative non-invasive methods have been developed and are applied whenever possible for the same purpose.

Magnetic Resonance Imaging Magnetic Resonance Imaging (MRI) and Spectroscopy (MRS), details of which are presented in the next section, are non-invasive signal acquisition techniques based on the physical phenomenon of Nuclear Magnetic Resonance (NMR). MRI, as its name indicates, is an imaging technique used by expert radiologists to visualise the brain tissue in certain detail. It provides good spatial resolution, but no information concerning the tumour metabolism. MRS, instead, generates a signal in the time domain that is processed and transformed into the frequency domain. Its spatial information is less obvious to interpret, but it provides a metabolic *signature* of the analysed tissue. At best, both modalities (imaging and spectroscopy) can be used in parallel (MRSI) [30].

Computerised Tomography Computerised Tomography (CT) is a technique that uses X-rays flowing throughout the body with the objective of constructing 3D images of the tissues. Some of the radiation of the rays that pass through the body is absorbed differently by the tissues. The remaining signal that arrives to the electronic receptors is computer-processed so that

various cross-sectional images or slices of the inspected part of the body are generated. The superposition of several slices forms a 3D volume image, which is then analysed by the physician [31].

Positron Emission Tomography In a Positron Emission Tomography (PET) scan, a radioactive sample of glucose is injected into the patient's bloodstream. The compound eventually flows to the brain, carried by the blood. The accumulation of the liquid is detected by a scanner, monitorising the accumulation of this radiotracer and generating a computer-based image. Although this technique is not routinely used to diagnose brain tumours, it can be useful for the determination of their malignancy [31].

2.1.4 Brain tumour treatment

The prognosis of a tumour and the evolution of the patient will highly depend on the tumour type, its malignancy and the given treatment. Here, we very briefly introduce the most common treatments used to diminish the proliferation of brain tumours. Notice that each treatment does not exclude the others and several might be applied simultaneously or consecutively. For instance, it is not uncommon to apply radiotherapy after a craniotomy in those cases where a tumour could not be removed in full.

Craniotomy A craniotomy is a surgical operation that consists in opening the skull and removing the region affected by the tumour. Depending on the tumour location and spatial distribution, a complete removal may not be possible or advisable and only a partial resection is performed. It might be the case that the tumour is difficult to reach and accessible only through healthy tissue, which might be damaged in the process.

Chemotherapy Chemotherapy is a treatment in which specific drugs are given to the patient with the purpose of shrinking a tumour or slowing down its growth with the final aim of reducing its symptoms. It will rarely be effective for complete tumour removal. Chemotherapy may be delivered after surgery and might be complemented with radiotherapy. The form of

intake might be oral in form of pills, intravenously or by means of an implant introduced during surgery, which slowly releases the appropriate dosage of medicine into the body.

Radiotherapy Radiotherapy is a technique that uses high-energy rays aimed at destroying the targeted cancerous cells while not affecting the healthy ones. It is often used after surgery to kill cancerous cells that might have been left over; also for treating secondary brain tumours, or in recurrent primary tumours reappearing after surgery. It is administered by providing high dosage beams focusing the tumourous cells through several sessions, but can also be applied to the whole brain in smaller dosage to deal with secondary tumours. It can be delivered alone or together with chemotherapy.

2.2 Nuclear Magnetic Resonance in neuro-oncology

The physical phenomenon in which atom nuclei placed in strong magnetic fields absorb and emit electromagnetic energy is known as Nuclear Magnetic Resonance (NMR or MR for short). In medicine, this effect is used to extract, non-invasively, information from regions of the body that are difficult to reach, such as the brain. This information can be computer-processed to generate images or other types of signal and is investigated by experts in the area of neuroradiology.

MR scanners apply a uniform magnetic field to the body (in the region of interest, or ROI) in order to align the magnetic moment of many protons. Then, a radiofrequency pulse at a specific frequency is transmitted and its energy absorbed by the protons, which flip their spin magnets. This radiofrequency pulse is then switched off and the protons return to their normal state, releasing the previously absorbed energy to the environment. This remaining energy is collected by the receiver coils in order to quantify the nuclei involved in all this process. The most widely used nucleus in medical MR is the proton of the isotop hydrogen 1 (i.e., ^{1}H) due to its abundance in living tissue.

Because different tissues release the absorbed energy at different relaxation rates, it is thus possible to construct an inner picture of the explored body by identifying tissue types. In MRI, in order to obtain a visually contrasted image, there exist a number of acquisition parameters that can be tuned: for instance, the use of gradient magnetic fields to locate the source of the signal, or the use of different echo times.

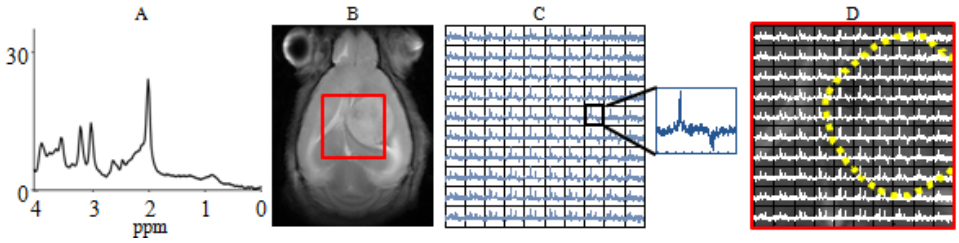


Figure 2.5: Nuclear Magnetic Resonance variants - Signals obtained with different MR modalities [25]: A) Single-voxel MRS; B)MRI; C) Multi-voxel MRS; D) Multi-voxel-MRS with imaging information superimposed.

In MRS, efforts often focus on a specific small volume ROI, called voxel (e.g., $\sim$1 cm^3), from which signal in the time domain, subsequently transformed into the frequency domain, is extracted (single-voxel proton MRS, or SV-^{1}H-MRS). The resonance frequency of several metabolites of interest is well documented, so that SV-^{1}H-MRS provides a metabolic signature of the explored tissue.

Note that, while MRI provides a space-related morphologic characterisation of tissues in the explored body, MRS provides localised biochemical information. MRSI, as a combined extension of MRS and MRI, was developed with the purpose of skimming the best of both techniques, taking advantage of the complementary information they provide. This technique provides spatially-located biochemical knowledge by performing several SV-MRS and plotting them in a grid-like fashion over an MRI. This way, multi-voxel (MV) information is made available to the radiologist, obtaining not only a picture of the inner tissues, but also information about their biological composition and metabolical behaviour.

2.2.1 Magnetic Resonance Spectroscopy in neuro-oncology

MRS can be extremely valuable when applied to the brain tumour diagnosis, since the increment or decrement of certain metabolites involved in tumoural tissues can be observed in an MRS and compared to normal tissue.

Among the parameters that must be set to perform an MRS scan, the time of echo (TE) is paramount. This is the elapsed time between the moment in which the radiofrequency pulse is switched off and the data acquisition starts. Usually, this time varies in the range of 18 and 288 ms in *in vivo* ^{1}H-MRS, and is characterised as either short time of echo (STE) TE $\leq$ 45 ms, or long time of echo (LTE) otherwise.

Scans at STE (usually 20-35 ms) are fast and robust and provide spectra with good resolution for certain metabolites. However, they present several overlapping peaks and are prone to include noisy artefacts. On the other hand, LTE spectra (around 135 ms) may not show T2 resonances, with the consequent loss of information, but they present less baseline distortion and frequently become easier to analyse.

Some of the most relevant metabolites in the analysis of human brain tumours are described in some detail next [32] (Figure 2.6):

N-Acetyl Aspartate (NAA, 2.01 *ppm* - parts per million) is the highest MRS peak in normal brain tissue. Although the exact role of this metabolite is unknown, it is usually interpreted as a neuronal marker. Being a 35% more present in grey and white matter than in the thalamus, with a proportion of 1.5 in grey matter with respect to white matter, the reduction of NAA means a reduction of the number of neurons in that region. It thus becomes a clear sign of dysfunction or death of neurons.

Creatine (Cr, 3.02 *ppm*) is understood to be an indicator of energy metabolism. It can be used as a reliable marker of cellular integrity. Usually, Cr is assumed to be quite stable in tumoural and non-tumoural tissues, which makes it a good candidate for the calculation of ratios with respect to other metabolites. A decrease of Cr can be found in brain lesions lacking kinase, as meningiomas, limphomas or metastatic brain tumours, as well as in aggressive tumours and hypoxic tissues.

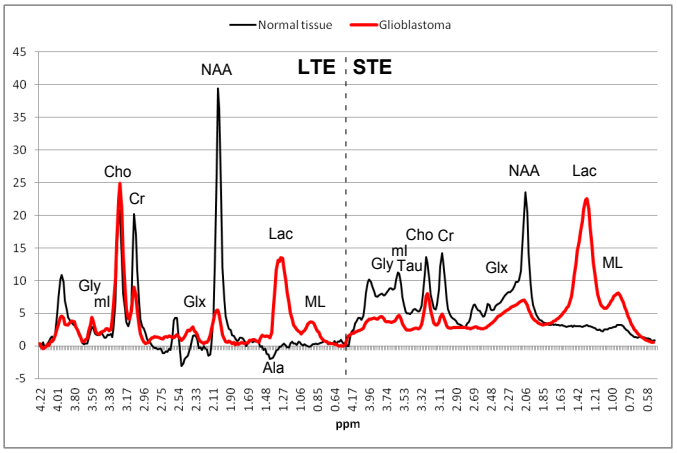


Figure 2.6: Main metabolites present in ^{1}H-MR spectra of the brain - Example mean spectrum of brain normal tissue (black) and glioblastoma (red) from a real MRS database, shown with the tags of the main tissue metabolites. The plot represents data acquired at LTE and STE, on the left and right part respectively (divided by a vertical line). Y-axes represent unit-free metabolite concentrations and X-axes represent frequency as measured in parts per million (*ppm*). Notice that not all the mentioned metabolites show a peak signal in this plot, given that their presence depends on the pathology.

Choline (Cho, 3.20 *ppm*) is a metabolic marker of membrane synthesis, density and integrity, which is found in higher concentration in glial cells than in neurons. Elevated concentrations of Cho may be associated with cell proliferation, hence generally increasing in tumoural areas, showing a correlation with malignancy. High Cho values are present in high-grade gliomas and glioblastomas, but also in infarction and inflammation. On the contrary, necrotic regions show low Cho signal.

Glutamine and *glutamate* (Glx, 2.05 - 2.46 *ppm*) are two metabolites that can be detected along the specified range, and they are usually considered together as Glx. They are found in neurons and astrocytes. While glutamate is a neurotransmitter, they both carry out detoxification and regulation tasks of neurotransmitters. High values of Glx might represent toxicity of the brain as well as indicate an altered energy metabolism, involving partial oxidation of glutamine. Elevated concentrations of Glx are often present in meningiomas.

Lactate (Lac, 1.31 *ppm*) appears as an inverted (usually negative) peak under the baseline in signal acquired at LTE and above that line in STE in MRS performed at brain regions in abnormal state. Lac does not show up in normal MRS. An increase in Lac might be due to a variety of conditions (e.g., hypoxia, ischemia, reduced oxygen supply, accelerated glycosis, inflammation, etc.), but it usually indicates a failure in the normal aerobic oxidation mechanism (meaning that oxygen may not be flowing to the analysed area through the vascular system). High-grade malignant tumours often generate high Lac peaks.

myo-Inositol (mI, 3.26 and 3.53 *ppm*) is a carbohydrate that is absent in neurons, synthesised in glial cells, hence being a glial marker. An increase of mI indicates a glial proliferation that could be caused by inflammation. Astrocytomas and low-grade gliomas are usually associated with an increase of mI.

Glycine (Gly, 3.55 *ppm*) is an amino acid found in high concentrations in astrocytomas and absent in meningiomas. Recent studies show it as a promising biomarker of malignancy in paediatric brain tumours [33].

Taurine (Tau, 3.42 *ppm*) is an organic acid that can only be observed at STE. It is difficult to measure because its peak overlaps with those of mI and Cho. It is routinely used as a biomarker for paediatric medulloblastoma and for measuring apoptosis in gliomas.

Alanine (Ala, 1.47 *ppm*) is an amino acid that can be found as an inverted peak in LTE in some meningiomas and pyosenic abscesses, but undetectable in normal brain. Its function is uncertain.

Mobile Lipids (ML, 1.3 and 0.9 *ppm*) are the major components of the brain, although no significant peak intensities of these components are found in normal MR spectra. Apparently, these lipids come from cell membrane during the ongoing metabolic changes associated with programmed apoptosis. The appearance of these peaks is usually associated with necrosis and hypoxia, which is often the case in high-grade tumours and metastases.

2.3 Biomedical data sets

The methods designed throughout this PhD thesis are assessed using, when appropriate, both artificial and real data. The real data are SV-^{1}H-MRS data acquired *in vivo* from brain tumour patients; and microarray gene expression for different diseases. For that, we have access to different data sources, whose characteristics are described next.

The INTERPRET ^{1}H-MRS database The database built as part of the INTERPRET (The Multi-Centre International Database of MR Spectra from Brain Tumours) European research project [18] is the main source of real data for this thesis. The creation of this database was coordinated by the Grup d'Aplicacions Biomèdiques de la Ressonància Magnètica Nuclear (GABRMN) at Universitat Autònoma de Barcelona (UAB, Barcelona - Spain) and was gathered from four different international institutions: Centre Diagnòstic Pedralbes (CDP, Barcelona - Spain), Institut de Diagnòstic per la Imatge (IDI, Barcelona - Spain), Saint George's Hospital Medical School (SGHMS, London - UK) and University Nijmegen Medical Centre (UMCN, Nijmegen - The Netherlands).

The data are SV-^{1}H-MRS acquired using Point-Resolved Spectroscopy (PRESS) and Stimulated Echo Acquisition Mode (STEAM) sequences at both STE (30 - 32 ms) and LTE (135 - 136 ms), including 512 spectral frequencies. The collected samples were validated and included in the database only if they complied with the following criteria:

- The voxel had to be positioned on the nodular part of the tumoural mass, avoiding cystic, oedematous or contralateral areas. In the case of normal volunteers, the voxel had to be positioned in a normal white matter region.
- The voxel had to be positioned in an area validated as the place where the biopsy or tumour resection was performed.
- The short echo spectrum from the validated voxel should not have been discarded because of acquisition artefacts, or for other reasons.

- Histopathological diagnosis had to be agreed among a committee of neuropathologists.

Table 2.1: Content of the INTERPRET database

Tumour type	STE	LTE
Astrocytomas grade II	22	20
Astrocytomas grade III	7	6
Brain abscesses	8	8
Glioblastomas	86	78
Haemangioblastomas	5	3
Lymphomas	10	9
Metastases	38	31
Meningiomas	58	55
Normal cerebral tissue, white matter	22	15
Oligoastrocytomas	6	6
Oligodendrogliomas	7	5
Pilocytic astrocytomas	3	3
Primitive neuroectodermal tumours and medulloblastomas	9	9
Rare tumours	19	18
Schwannomas	4	2

This table contains a list of the available tumour types and their number of cases acquired at STE and LTE [24].

The final database contains MR spectra from 266 patients at LTE and 304 at STE. The spectra were labelled according to the WHO system for diagnosing brain tumours determined by histopathological analysis of biopsy. Some of them might also contain MR images with the selected voxel to perform the MRS explicitly marked as well as a detailed patient anonymous profile.

The exact number of cases per tumour pathology is described in Table 2.1.

eTumour ^{1}H-MRS data set Further data that were made available to this thesis were obtained as part of the European Union-funded eTumour [19] research project .

These data were acquired, amongst others, from three clinical centres in the Barcelona metropolitan area: CETIR-CDP (Centre Diagnòstic Pe-

dralbes, Unitat Esplugues, Esplugues del Llobregat), Corporació Sanitària IAT (Institut d'Alta Tecnología, Barcelona) and IDI-Badalona (Institut de Diagnòstic per la Imatge, Unitat Badalona, Badalona). The data set consists in ^{1}H-MRS data (both at STE and LTE) from 10 patients affected by *glioblastoma* brain tumours and 30 records from patients diagnosed as having a brain *metastasis*. Similar acquisition conditions and preprocessing as previous INTERPRET project was required, a fact that makes these data directly comparable to those from INTERPRET.

Microarray gene expression database A widely-used collection of microarray gene expression data sets presenting a variety of diseases is used in this thesis. In particular, each one of them shows the expression levels of a large number of genes corresponding to different individuals (i.e., affected patients and controls). The high feature dimensionality (genes) with respect to the number of samples (patients) makes these data sets suitable to validate our models.

Table 2.2: Microarray gene expression database

Dataset	patients	genes
Colon cancer [34]	62	2,000
Leukaemia [35]	72	7,129
Prostate cancer [36]	102	6,034
Lung cancer [37]	181	5,000
Breast cancer [38]	97	5,000
Melanoma [39]	70	5,000
Parkinson [40]	105	5,000

This table contains information regarding the size of microarray gene expression data sets after pre-processing.

In the case of Prostate cancer data set, a preprocessing similar to the one in [41] is performed: it consists in fixing a valid range of values for each gene to lay between $[10, 16000]$. Any value out of this interval is set to its closest limit. Subsequently, genes presenting low variability ($max/min < 5$ or $max - min < 50$) are removed.

For the Lung, Breast and Melanoma cancer data sets, as well as for the Parkinson data set, a standard t-test was applied, retaining the $5,000$ top genes [42]. No pre-processing was applied to the Colon and Leukaemia cancer data sets. Table 2.2 summarises the properties of each data set.

Chapter 3

Technical background

In this chapter, we summarily present some general technical concepts of relevance to the thesis. We start by providing an introductory overview of Machine Learning (ML) approaches and we do follow this by self-contained descriptions of ML techniques and problems of relevance to our work, including ensemble learning, dimensionality reduction, algorithmic stability and Bayesian inference. The chapter concludes with a brief review of the state of the art in the application of pattern recognition and ML techniques to neuro-oncology problems.

3.1 Machine Learning

Machine Learning, a field of research under the *umbrella* concept of Artificial Intelligence, aims at developing new algorithms able to *learn* (model) an unknown function f from a set of observed data. By running an ML algorithm on the available data, a model is *trained* for a specific task (that is, f is learnt). The ultimate goal of this trained model is to be capable of predicting realistic outcomes for unseen data.

Depending on the task to be performed, diverse approaches can be adopted, giving rise to different subfields within ML. In this section, we summarily review the two categories traditionally considered as the most relevant: *supervised* and *unsupervised* learning.

3.1.1 Supervised learning

Given a data set D containing N pairs $\{(\mathbf{x}_1, t_1), ..., (\mathbf{x}_N, t_N)\}$, where $\mathbf{x}_n \in X$ is a multivariate data point and $t_n \in T$ its corresponding so-called label (for instance, class information representing membership in classification tasks), a supervised algorithm attempts to learn a function $f : X \to T$ such that a label $\hat{t}_n \in T$ for new unlabeled data $\mathbf{y}_n \in X$ is *coherently* inferred: $\hat{t}_n \leftarrow f(\mathbf{y}_n)$, such that $\hat{t}_n$ is the most probable realisation for $f(\mathbf{y}_n)$.

Classic learning algorithms for this setting [43] include, but are not limited to, Nearest Neighbour (NN), which assigns $\mathbf{y}_n$ the label $\hat{t}_n$ of its most similar instance from D; Decision Trees (DT), that build a tree-like hierarchical model (according to D) to be used as a flow-chart guiding the inference from the root to the leaves, evaluating the appropriate attribute at each level until the proper label is finally assigned in the resulting leaf; Linear Discriminant Analysis (LDA), that aims at finding a geometrical representation that maximises the distance between the classes' averages while minimising the variance between instances of the same class; Logistic Regression, which models the probability of $\hat{t}_n$ for $\mathbf{y}_n$ by fitting a logistic function to D; Support Vector Machines (SVM), which searches for the hyperplane that maximises the separation between class boundaries; and Artificial Neural Networks (ANN), a biologically-inspired technique that mimics the functionality of the brain by constructing a network of artificial neurons capable to learn and infer on given data.

3.1.2 Unsupervised learning

In the unsupervised framework, the data set D lacks any information regarding class assignment. In this case, the goal is finding the underlying structure of the data. Among the specific subtasks that have been widely pursued among the research community, we pay special attention to Clustering, which consists in inferring groupings of similar instances; and Blind Source Separation (BSS), which attempts to find the underlying hidden signals from the observed data, a noisy mixture of which conforms each of our $\mathbf{x}_n$.

Well-established algorithms for the former include, amongst others, K-means [44], where an iterative process assigns instances to each of the k groupings according to the distance to its group prototype; Hierarchical Clustering [43], that assumes that the grouping structure of the data operates at different levels of detail and strives to obtain a hierarchy of groupings by merging similar elements; and Self-Organising Maps [45].

BSS usually borrows procedures from Feature Extraction, a form of dimensionality reduction that will be discussed in some detail in Section 3.3.2.

3.1.3 Assessing predictive capability

In order to quantitatively assess the modelled f function, different measures have traditionally been used to evaluate the predictions on data that were not present during the training. Here, we review some of these measures that will be employed to determine the correctness of the new models generated in this thesis.

In a typical binary classification setting under the supervised learning paradigm described in Section 3.1.1, where $\hat{t}_n \in \{0, 1\}$ is the outcome of each prediction $f(\mathbf{y}_n)$ for a set $\{\mathbf{y}_n\}_{n=1}^{N}$, representing whether instance $\mathbf{y}_n$ belongs to the positive class ($\hat{t}_n = 1$) and t_n is the real outcome, we say the prediction falls within one of the following categories:

- True Positive:

$$\mathrm{TP}_n = \mathrm{TP}(\hat{t}_n, t_n) = \begin{cases} 1 & \text{if } \hat{t}_n = 1 \;\&\; t_n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- False Positive:

$$\mathrm{FP}_n = \mathrm{FP}(\hat{t}_n, t_n) = \begin{cases} 1 & \text{if } \hat{t}_n = 1 \;\&\; t_n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

- True Negative:

$$\mathrm{TN}_n = \mathrm{TN}(\hat{t}_n, t_n) = \begin{cases} 1 & \text{if } \hat{t}_n = 0 \;\&\; t_n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

- False Negative:

$$\mathrm{FN}_n = \mathrm{FN}(\hat{t}_n, t_n) = \begin{cases} 1 & \text{if } \hat{t}_n = 0 \;\&\; t_n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Insights on the behaviour of the evaluated model can be obtained by accumulating the number of predictions falling in each category. For instance, we define the *precision* (a.k.a Positive Predictive Value - PPV) as the fraction of correct positive predictions out of all instances predicted to belong to the positive class. That is:

$$\text{PPV} = \frac{\sum_{n=1}^{N} \text{TP}_n}{\sum_{n=1}^{N} (\text{TP}_n + \text{FP}_n)}.$$

Similarly, *sensitivity* (a.k.a recall or True Positive Rate - TPR) is described as the ratio of correct positive predictions out of all instances really belonging to the positive class. In symbols:

$$\text{TPR} = \frac{\sum_{n=1}^{N} \text{TP}_n}{\sum_{n=1}^{N} (\text{TP}_n + \text{FN}_n)}.$$

Contrarily, the *specificity* (a.k.a True Negative Rate - TNR) measures the proportion of negative predictions correctly classified as such out of all real instances belonging to the negative class, which can be derived as:

$$\text{TNR} = \frac{\sum_{n=1}^{N} \text{TN}_n}{\sum_{n=1}^{N} (\text{TN}_n + \text{FP}_n)}.$$

Last measure of this kind being of interest in this thesis is the *fallout* (a.k.a False Positive Rate - FPR), which is defined as the complementary of the *specificity* by calculating the incorrectly predicted positive instances out of all the real instances belonging to the negative class. Its formula is:

$$\text{FPR} = \frac{\sum_{n=1}^{N} \text{FP}_n}{\sum_{n=1}^{N} (\text{TN}_n + \text{FP}_n)}.$$

While any of the above measures can be directly employed to evaluate the performance of a model from different angles, it is quite common to summarise the overall performance in a single measure. In this respect, we encounter the widely used Accuracy (ACC) and its complementary Error Rate (ER), which can be formulated as:

$$\text{ACC} = 1 - \text{ER} = \frac{\sum_{n=1}^{N} (\text{TP}_n + \text{TN}_n)}{N}.$$

Despite being frequently used, there exists a major shortcoming in using this measure when the dataset employed to validate the model is highly unbalanced (i.e., the proportion of instances belonging to one class is much higher than the other), leading to a false appearance of good performance (e.g., a naive model always predicting the positive class achieves a 0.99 accuracy in a dataset made up of 99 positive instances and only 1 negative instance). To overcome such phenomenon, the Balanced Accuracy (BAC) and its Balanced Error Rate counterpart can be defined as:

$$\mathrm{BAC} = 1 - \mathrm{BER} = 0.5 \times \mathrm{TPR} + 0.5 \times \mathrm{TNR}.$$

Another measure of interest is the F-measure (F), which corresponds to the harmonic mean of precision and recall and is defined as:

$$\mathrm{F} = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}.$$

When classifications are based on a continuous random variable, we can assess the probability of an instance belonging to a class as a function of a decision threshold τ. Picking the appropriate τ that leads to the best classification accuracy can be achieved by drawing a Receiver Operating Characteristic (ROC) curve by plotting the TPR (y-axis) as a function of the FPR (x-axis). The value of τ corresponding to the point at the top-left corner of the plot will be the best choice. In certain cases it can be of interest to summarise the predictive accuracy of a model regardless of τ in a single score by calculating the Area Under the ROC Curve (AUC) [46], despite some recent controversies on using this measure to assess classification models [47, 48]:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\tau) \times \mathrm{FPR}(\tau) \mathrm{d}\tau$$

In certain cases, the AUC is underestimated due to the procedure used to approximate the integral. When this happens, an optimistic estimate evaluating the Area Under the Convex Hull of the ROC curve (AUH) can be of interest. A thorough description of a fast algorithm to calculate it can be found in [49].

Apart from the measures introduced in this section to assess the predictive capability of the generated models, there exist other formulae that will

be used in this thesis to evaluate other aspects of the ML experiments, such as the effective *stability* in feature selection algorithms or the *correlation* between inferred and real data. They will be thoroughly explained in their respective chapters, when required.

3.2 Ensemble learning

The different techniques mentioned in the previous section may be suitable for a broad range of problems. Nonetheless, there are some situations (i.e., when f is not smooth) in which single learners are not able to properly capture the properties of the function.

To address this limitation, the ML community borrowed a concept from the fields of psychology and social sciences, known as the *wisdom of the crowd*, and applied it to its domain. As on a trial, where a fair verdict might come from a popular jury, even though each one of the individuals may have a different background and does not need to be an expert in the domain, it seems plausible to use an algorithmic analogy of this approach in which a combination of different learners are used to obtain a final single classification or regression decision.

Within the ML community, this analogous concept is known as Ensemble learning and is instantiated in the form of an *ensemble of classifiers* or a *committee*. Its purpose is to improve the prediction accuracy of the single models by aggregating, in different ways, their individual outputs.

All ensemble techniques share the same overall structure (Figure 3.1):

- A set of different *base learners* that conform the committee.
- An *aggregation strategy*.
- A process responsible of generating *diversity*.

Behind the intuition of why ensembles work, there is a sound statistical theory, known as the bias-variance decomposition [50], which helps in explaining this phenomenon. It states that the error of any model can be split into three different terms:

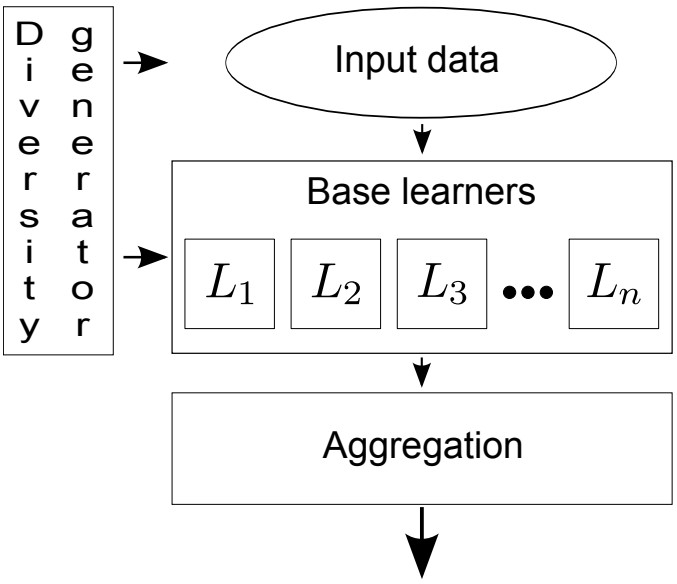


Figure 3.1: General ensemble structure - General ensembles consist of three main modules: the base learners, the aggregation strategy and the diversity generator.

- *Bias*: a quantity measuring the difference between the models' average guess and the real hypothesis.
- *Variance*: a quantity measuring the spread of individual models with respect to the average guess.
- *Intrinsic noise*: a quantity measuring the minimum achievable loss of the model, known as the Bayes error.

Notice that bias usually increases when a model has insufficient flexibility to model the data adequately. Conversely, when increasing model flexibility in an attempt to decrease bias, sampling variance is increased. Therefore, in any process of prediction, error minimisation can be considered as a trade-off between bias and variance.

In the case of ensemble learning, its purpose is to increase prediction accuracy by reducing either bias, variance or both components of this equation by means of aggregating multiple models. A graphical representation of the bias and variance decomposition is shown in Figure 3.2.

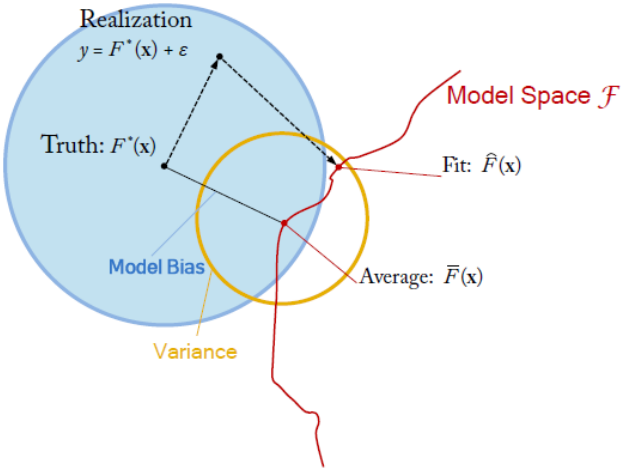


Figure 3.2: A representation of bias and variance decomposition - Model bias is represented as the distance between the true function $F^*(x)$ and the models' average guess $\bar{F}(x)$. Variance is shown as the spread of the different models $\hat{F}(x)$ around their average $\bar{F}(x)$ [51].

3.2.1 Classical ensembles

A handful of ensemble techniques have been proposed over the last decades. Here, we introduce some of them; they constitute the core of approaches worth describing in some detail.

3.2.1.1 Bagging

Bagging (Bootstrap aggregating) [52] is an algorithm to create an ensemble of classifiers that uses bootstrap samples to generate the diversity among its base learners. The procedure consists in uniformly sampling M instances from the training set with replacement for each of the L classifiers. For a large value of M, the ratio of different instances in each bootstrap sample is $1 - \frac{1}{e}$, which corresponds to a 63.2%. This value justifies the use of unstable learning algorithms (e.g., ANN or decision trees) able to predict differently with this limited diversity. Finally, the aggregation of the predictions provided by the classifiers is performed by a simple majority voting. This algorithm aims at reducing variance to improve accuracy, given that it is composed of unstable classifiers, which are known to produce high variance.

3.2.1.2 Boosting

Adaboost [53] is the most representative algorithm using the *boosting* strategy. It creates an ensemble of classifiers by iteratively sampling from the training set using different probabilities to pick each instance, in order to construct every base learner.

That is, it starts by constructing the first learner using a sample from the training set, where each instance has probability $1/N$ to be chosen. A classifier is built using this sample and the prediction for each instance is computed. Then, the probability to pick each instance is modified according to the errors obtained, giving higher probability to those instances that have been misclassified (i.e., harder instances to predict). The next classifier is created by sampling from the training set using the new probabilities. This process is carried out until all classifiers are created.

Moreover, each classifier also calculates its weight within the ensemble by computing its generalization error. This value is taken into account in the aggregation phase, where the ensemble prediction is obtained by weighted majority voting.

Its success can be devoted to a reduction in variance by averaging different hypotheses; however, the effect of forcing the weak learner to concentrate in different instance space also contributes to a decrease in bias.

3.2.1.3 Random Subspace

The Random Subspace method [54] is a strategy originally devised to construct ensembles of classifiers based on decision trees, although it can be used with other learning methods.

Let L be the number of base classifiers that conform the ensemble, D the number of features of the data and $d << D$ the number of features to be used by each classifier. A selection of d different features out of D is performed and the data is projected into the d-dimensional subspace, where a tree learner is grown. The ensemble is constructed by applying this procedure L times, leading to the construction of the required number of base classifiers. In the prediction phase, their outputs are aggregated using

one of the many variety of techniques already commented in order to obtain the ensemble prediction. Success of Random Subspace method resides in diminishing variance yet the explanation of why this happens is far from obvious.

3.2.1.4 Random Forest

Random Forest [55] is a technique used to build a committee of experts made of tree classifiers. It borrows ideas from both Bagging and Random Subspace with the purpose of devising a technique able to make the most of both strengths.

Let N be the number of instances, D the number of features and L the number of classifiers conforming the ensemble. For every learner l, a bootstrap sample of N instances (with replacement) is retrieved. Then a tree classifier is built using these samples where at each node of the tree, a subset of features $d << D$ is randomly selected. The best split for the d features is kept as the decision made at this node.

This process is repeated at every node for each learner until achieving a whole ensemble made of unpruned trees. The final ensemble prediction is aggregated using usual strategies.

According to Breiman's experiments [55], results suggest that Random Forest acts as a bias reducer, yet the explanation on why this happens is not trivial.

3.3 Dimensionality reduction

A recurrent challenge when classifying high-dimensional data is the *curse of dimensionality* [56], which hinders the process of knowledge extraction from data using ML techniques. Among the drawbacks it generates, the most frequent is the inability of many pattern recognition techniques, which perform very well in a low-dimensional space, to maintain their accuracy and robustness when dimensionality increases due to data sparsity; or the instability that uninformative or misleading features can generate in those techniques for the task at hand.

In certain application domains, as in the one this thesis deals with, keeping data dimensionality low is crucial for the sake of achieving visualisation and interpretability, this being an often mandatory requirement for knowledge extraction.

For the reasons explained above, as well as others, many techniques to reduce the dimensionality of data have specifically been designed for either classification purposes (i.e., supervised methods which take into account the class label of every instance) or for general purpose (i.e., unsupervised methods which use correlations among features to rank their importance). In this section, we revise the most commonly used.

3.3.1 The feature selection problem

Feature subset selection (FSS) in a set Y of size D is commonly seen as a *search problem* where the search space is the power set of Y, $\mathcal{P}(Y)$ [57]. Without loss of generality, we assume that the evaluation measure $\mathcal{L} : \mathcal{P}(Y) \rightarrow \mathbb{R}^+ \cup \{0\}$ is to be maximised. The criterion $\mathcal{L}$ may be problem-independent or may depend on the classifier that will be used to solve a classification problem. In any case, we will refer to $\mathcal{L}(X)$ as the *usefulness* of feature subset X.

Let $\mathcal{L}$ be an evaluation measure to be optimised (say, to maximise). The selection of a feature subset can be carried out under two premises:

- Find $X^* \subset Y$, such that:

$$X^* = \underset{X \in \mathcal{P}(Y)}{\arg\max}\, \mathcal{L}(X) \tag{3.1}$$

- Set a real value $\mathcal{L}_{min}$, that is, the minimum $\mathcal{L}$ that is going to be accepted. Find the $X_K \subseteq Y$ with smaller K such that $\mathcal{L}(X_K) \geq \mathcal{L}_{min}$. Alternatively, given $\epsilon > 0$, find the $X_K \subseteq Y$ with smaller K, such that $|\mathcal{L}(X_K) - \mathcal{L}(Y)| < \epsilon \mathcal{L}(Y)$.

Notice that, with this definition, the optimal subset of features always exists but is not necessarily unique. Also noteworthy is the fact that, denoting

by X^* one of the optimal solutions, either of $\mathcal{L}(X^*) > \mathcal{L}(Y)$, $\mathcal{L}(X^*) = \mathcal{L}(Y)$, $\mathcal{L}(X^*) < \mathcal{L}(Y)$ may occur.

Ideally, feature selection methods search through all the subsets of features and try to find the best one. It is clear though, that if we had to test all possible subsets of features using either of the methods, we would be faced by a combinatorial explosion of possibilities. If our initial set of features is Y and $|Y| = D$, the number of evaluations we would have to do would be equal to the cardinality of the power set of Y: $|\mathcal{P}(Y)| = 2^D$. A *complete* search (as with the Branch and Bound method), is a feasible procedure to guarantee the finding of an optimal subset; this method also requires the monotonicity of the inducer evaluation. This implies that when a feature is added to the current subset, the value of the criterion or evaluation function does not decrease. In most practical applications, this approach is computationally prohibitive and the mainstream of research on FSS has thus been directed to *sequential* suboptimal search methods.

A Sequential Feature Selection Algorithm (SFSA) is a polynomial-time computational solution that is motivated by a certain definition of *usefulness*. An important family of SFSAs perform an explicit search in the space of subsets by iteratively adding and/or removing features one at a time until some stop condition is met. These methods typically share the same basic steps:

1. The *subset generation* to produce candidate subsets for evaluation
2. The *evaluation criterion* providing the usefulness of each subset
3. The *stopping criterion* to decide when to stop

Looking at the evaluation criterion, [58] divided the feature selection methods into two main approaches: *filter* methods and *wrapper* methods. These two families of methods only differ in the way they evaluate the candidate sets of features. A third group of methods called *embedded* methods are a more recent approach to feature selection where the selection process is done implicitly as part of the classifier design.

3.3.1.1 Filters

These methods use a problem independent criterion. The basic idea of these methods is to select the features according to some prior knowledge of the data. For example, selection of features based on the *conditional probability* that an instance is a member of a certain class given the value of its features [59]. Another criterion commonly used by filter methods is the *correlation* of a feature with the class (i.e., selecting features with high correlation [60]). A well known family of filter algorithms is Relief [61], which estimates the usefulness of features according to how well their values distinguish between the instances of the same and different classes that are near to each other.

3.3.1.2 Wrappers

These methods suggest a set of features that is then supplied to a classifier, which uses it to classify the training data and returns the classification accuracy or some other measure thereof [62]. The search is guided by the classifier used as a *black box* (i.e., the feature selection process does not depend on how the classifier works). It is suggested in the literature that wrapper methods, although they tend to overfit, perform better than filters [58, 62] because using the classifier error rate used as the evaluation criterion catches the structure and properties of the classifier better. Among the proposed algorithms for attacking this problem, we find Sequential Forward Generation (SFG) and Sequential Backward Generation (SBG), the Plus l - Take Away r or PTA(l, r) proposed by Stearns [63], or the Floating Search methods [64]. They both introduce methods for the generation of the sets of features by combining steps of SFG with steps of SBG, but keep using a certain $\mathcal{L}(X)$ as evaluation criterion.

3.3.1.3 Embedded methods

The idea here is to optimise the evaluation criterion $\mathcal{L}(\cdot)$ directly and to perform feature selection as part of the classifier training. This mechanism can be found in algorithms like SVM [65], Adaboost [53], or Classification and Regression Trees (CART) [66].

Filter measures (as probabilistic *separability* measures) do not induce the same preference order as would be obtained by comparing classification *error rates*. This is due to the fact that error rates capture not only class separability but any structural error imposed by the form of the classifier. As the second aspect is not reflected in FSS based exclusively on filter measures, the resulting features may perform poorly when applied as the input of the classifier. Therefore, the legitimate way of evaluating feature subsets must be through the error rate of the classifier being designed [62].

3.3.2 Feature extraction

In Feature Extraction (FE), we aim to find a new set of features that are a combination of the original D observed data dimensions. The number of these extracted features is often lower than D, thus achieving dimensionality reduction. This approach works on the hypothesis that the observed features are not the true variables from the source, but a noisy combination of the real hidden, unobserved, or latent variables generated by the underlying process that created the data. Some linear FE techniques are described next.

3.3.2.1 Principal Components Analysis

Principal Components Analysis (PCA) [67] is an unsupervised FE technique that projects the data into a new orthogonal D-dimensional space in such a way that the variance in each linearly uncorrelated dimension is maximised. The projection is performed by defining the first Principal Component (PC or dimension) as the axis along which the data accounts for most of the variability. The rest of the components are defined sequentially as the orthogonal axes that explain the remaining variance in decreasing order.

In other words, the PCs are the eigenvectors of the covariance matrix ranked in order of importance according to their eigenvalues.

The reduction in dimensionality is often achieved by using only those PCs that explain a given (reasonably high) amount of the data variability (e.g., using the K PCs accounting for 75% of data variance).

3.3.2.2 Independent Components Analysis

Independent Components Analysis (ICA) [68] is a statistical BSS method that seeks to decompose a data set into independent subparts. In this model, it is hypothesized that the observed data $\mathbf{V}$ (matrix of D features by N instances) is a product of combining the non-Gaussian, mutually independent latent variables $\mathbf{W}$ (D by K) with the mixing matrix $\mathbf{H}$ (K by N). Hence, algorithms implementing ICA find the proper combination of $\mathbf{V} \approx \mathbf{WH}$ such that the statistical independence among the estimated components is maximised.

3.3.2.3 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) [69] is an alternative technique for matrix factorisation ($\mathbf{V} \approx \mathbf{WH}$), similar to ICA, but with the difference that it imposes the constraint that all matrices $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ must be non-negative (i.e., all elements must be equal to or greater than 0). Moreover, in its initial formulation, the goal of NMF is to minimise the divergence between $\mathbf{V}$ and $\mathbf{WH}$.

3.4 Algorithmic stability

An important aspect to address when designing a learning algorithm, besides its capability to predict accurately, is its stability. Stability is defined here as the robustness of an algorithm to possible perturbations to its inputs. That is, if small changes in the inputs lead the models learnt at different runs of the algorithm to provide completely different outputs, we deem the learning algorithm as unstable; otherwise, the learning algorithm is considered to be stable.

One of the biggest impacts of unstable learning algorithms is on the trust that domain experts can place to the model. Such experts are bound not to trust a model that makes different decisions at each execution, despite consistently providing high prediction accuracy.

Traditionally, emphasis has been placed in analysing and improving *stability in learning algorithms*. This idea was first introduced in [70], who not

only provided a formal definition for that concept, but also a figure of merit to quantify stability based on the level of agreement between the outputs of two perturbed learnt models.

A major contribution in the field was published in [71], where stability of learning algorithms was linked to the concept of generalisation error. The study concluded that stable models tend to generalise better than unstable ones.

3.4.1 Stability of feature selection

As stated in Section 3.3.1, FSS is employed in many domains to aid learning algorithms improve their performance in high-dimensional data spaces. Most of the efforts in this field have been made in the design of algorithms that are able to obtain a minimum subset of features that best captures the properties of data for the ultimate goal of building a classification or regression model.

Special attention has been paid to identifying and reducing the redundancy among the selected features. Redundant features increase data dimensionality while not providing new information for the task at hand. Therefore, it is common practise to keep any of the relevant redundant features while discarding the rest. Notice that such approach might end up damaging the *stability of feature selection*. In other words, if several redundant features are equally relevant and probable, different runs of a FSS algorithm might select different features to explain the same phenomenon, which translates into a decrease in stability.

Instability of FSS is especially harmful in knowledge discovery, whose main purpose is to identify those features that best explain the differences between groups of samples. An example can be found in the field of biological sciences (particularly in the *-omics* sciences), where FSS techniques are used to obtain a set of potential biologically relevant feature candidates (a.k.a. biomarkers) that must be further validated in costly biological settings.

Another source of instability of FSS arises when the number of data instances is very small as compared to the dimensionality of data. In such

cases, very different subsets of features might be equally good for explaining the target concept.

Despite being identified as an important issue, stability of FSS has only scarcely been addressed in ML research. Most of the existing studies deal with providing measures for assessing the stability of FSS [72, 73, 74], and only recently, a few studies have proposed strategies for actively increasing the stability of FSS algorithms while maintaining their predictive capability [75, 76, 42]. Notice at this point that nobody is likely to be interested in a highly stable FSS algorithm at the cost of major decrease in prediction accuracy.

3.5 Bayesian inference

Up to this point, we have been using the so-called *frequentist* approach to explain our methods, which is just one of the two main avenues of ML in a statistical setting. The philosophy of frequentists explains the world to be made up of a set of fixed (either known or unknown) phenomena. This translates in assuming that data are repeatable (i.e., sampling is infinite) and the parameters defining the underlying process that generates them are fixed (although unknown). Moreover, there is no information to be used prior to the model specification; hence, inference is based only on processed data.

The other side of the coin is known as the *Bayesian* paradigm. This setting views the world probabilistically, meaning that unknown quantities (i.e., model parameters) are defined to be an instantiation of a random variable from a probability distribution. It also assumes all the available data to be fixed and contained in the sample realisation. It is therefore important to account for prior information of interest.

This alternative method uses Bayes' rule for updating a probability estimate of a predefined hypothesis as new evidence (i.e., in form of new data) is observed:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)},$$

where H stands for hypothesis and E for evidence. In this equation, $P(H)$ is the so-called *prior probability*, which models the previous belief on the hypothesis before any evidence has been observed; $P(E|H)$ corresponds to the *likelihood* of the model, and accounts for the probability that evidence E occurred under hypothesis H; finally, $P(E)$, usually known as *marginal likelihood* or *model evidence*, acts as a normalising constant ensuring that probability integrates to 1. The resulting $P(H|E)$, known as *posterior probability*, is the probability that hypothesis H occurs after observing evidence E.

The Bayesian paradigm is not only a probabilistic interpretation of classical frequentist methods; it also provides a number of advantages mainly due to the power of the marginalisation, allowing to integrate all *nuisance variables* out instead of estimating them [77]. Next, we present some of the advantages of using a Bayesian approach.

The first benefit is that Bayesian inference is based on solid statistical theory, providing a reliable tool that experts can employ confidently.

Secondly, it provides a full probability model, meaning that the output of a model is not only a sharp decision, but a probability. For instance, in the context of classification, this probability can express the degree of membership for a given class prediction.

Moreover, the unavoidable uncertainty of predictions is straightforwardly dealt with by means of providing credible intervals to the solution.

Any prior domain knowledge can be easily incorporated into the model using prior probabilities. This is especially interesting in those cases where few data are available (e.g., in medical contexts). Continuing with small sample size-related problems, the Bayesian approach allows avoiding resampling strategies such as cross-validation that further reduce these small data sets in the learning phase of the model.

Another advantage is the capacity to automatically update a given model as new data are obtained. That is, the previous model can be used as prior information for building the new model using the new obtained data. This means that there is no need to fully retrain the model, but we can use an automatic update instead.

Further gains include the property of automatically avoiding overfitting due to the *integrating out* operation on the parameters. Additionally, model complexity is regulated (as an implementation of Ockham's razor), given that over-complex models are penalised through assignment of lower posterior probabilities.

The final advantage we want to highlight is the ability of Bayesian inference to provide a framework for model selection, using Bayes' factor [78] to compare among different models.

Nevertheless, there is a main drawback that has to be taken into account when designing Bayesian models in real scenarios, related to their computational complexity. Some of the operations involved in the resolution of Bayes' rule (e.g., determining the marginal likelihood) involve the computation of difficult integrals which often cannot be analytically solved. Different approaches based on approximations have been proposed to overcome this limitation, giving rise to active fields of research. They can be broadly grouped into stochastic and deterministic solutions [79]: the first uses sampling techniques by means of Monte Carlo methods (e.g., rejection sampling, importance sampling) or the more sophisticated Markov Chain Monte Carlo, including the Metropolis Hastings algorithm or Gibbs sampling. The second group is based on analytical approximations to the posterior distribution, including Variational Inference and Expectation Propagation.

To sum up, the Bayesian paradigm shows a wide range of useful properties for data modelling, but they come at the price of complex derivations, or high computational time. This is the reason why in certain practical settings, the frequentist approach (based on efficient optimisation) is just good enough.

3.6 Application of Machine Learning and Pattern Recognition to the diagnosis of brain tumours

In current medical practise, and unless absolutely necessary, the diagnosis and prognosis of human brain tumours are carried out on the basis of infor-

mation obtained through non-invasive techniques such as those described in the previous chapter.

The availability of this information in electronic format requires computer-based processing. As a result, it becomes suitable for analysis using pattern recognition methods stemming from the fields of statistics, computational intelligence and ML [11].

As previously mentioned, MRS is a promising data acquisition technique that provides the expert with detailed local information about the metabolites present in the analysed tissue. However, the interpretation of MRS requires the expertise of specialised radiologists, which do not abound in the field. Therefore, research on pattern recognition in general and ML in particular has emerged over the last two decades with the goal of providing analytic support for diagnostic and prognostic decision making in neuro-oncology.

Back in 1992, a review on the available literature about *in vivo* MRS of human cancers, conducted by W. Negendank [80], identified the potential of certain metabolites to become prognostic indices for different brain tumours and laid out the main foundations of research in this problem, emphasizing the need of improving diagnostic specificity and the use of statistical analysis of multiple spectral features (multivariate statistical analysis).

In 1996, Preul *et al.*[81] proposed, for the first time on record, the use of pattern recognition techniques for the classification of brain tumours on the basis of MRS data. They employed spectra retrieved at LTE to differentiate between different grades of astrocytoma (II, III and IV), meningiomas, metastases and non-tumoural tissue. LDA classification was performed using six well-known selected metabolites (Choline, Creatine, N-Acetyl Aspartate, Alanine, Lactate and Lipids), achieving up to a 99% success rate.

Hagberg [82] performed a thorough review on the most successful techniques used to date regarding FSS and classification applied to the problem of brain tumour diagnosis. Among them, PCA, LDA and Optimal Discriminant Vector, together with peak integration and intensities, were used for FSS. As for classifiers, the most employed were LDA and ANN. Results varied, and the classification settings were too different to conclude that any

technique showed clear advantages over the others. Since then, the relative merits of different techniques and approaches to tackle these problems have been discussed in some detail [83, 84, 85, 86].

De Edelenyi [87] introduced the concept of nosologic images to deal with the problem of heterogeneity within a tumour. The problem is clear: depending on the voxel of choice, tissue within it can be tumoural, non-tumoural, or a mixture of both, which has an impact on the resulting spectroscopic pattern. In nosologic images, we move from a SV measurement to multiple ones (MV), overlaying this information with a corresponding tumour image. Then, the spectrum of each voxel forming the tumour is classified as belonging to one of the histopathological classes and a colour is assigned to each of the classes, conforming a spatially-informative image.

Much work has been carried out specifically using data from the European INTERPRET project, the multi-centre study at the origin of some of the data sets analysed in this thesis. In 2003, Tate and colleagues [88] classified cases retrieved from three different centres using different acquisition protocols. The samples were split into three groups (meningiomas, low grade astrocytomas and aggressive tumours). Accuracies up to 92% were achieved using simple LDA classifiers.

Opstad [89] used the LCModel approach [90] followed by LDA to differentiate among astrocytoma grade II, astrocytoma grade III, glioblastoma, metastasis and meningioma. The results obtained reached an accuracy of 94% when differentiating high-grade gliomas, astrocytoma grade II and meningiomas (a reasonably easy problem), and 82% when astrocytoma grade III were also included. For the classification of glioblastomas from metastases, an inherently difficult problem, the score was 70%.

In 2004, Opstad [91] also analysed the problem of differentiating between metastases and glioblastomas. A sample of only 23 glioblastomas and 24 metastases was used. It was concluded that the Lipid and Macromolecule signals might be useful for such discrimination. Values of 80% sensitivity and 80% specificity were achieved.

In the doctoral thesis carried out by L. Lukas [92], discrimination between glioblastomas, meningiomas, metastases and astrocytomas was at-

tempted using kernel methods. Specially successful was the use of linear Least-Squares Support Vector Machines (LS-SVM), where an AUC bigger than 0.90 for LTE and over 0.95 for STE was achieved in the classification of these tumour types, with the only exception of the discrimination of glioblastomas from metastases.

Simonetti et al. [93] coupled the information provided by multi-voxel MRS and different MRI products to perform classification of tumour grade at every voxel. More precisely, they constructed a feature space by using 7 features from MRS (by means of PCA or peak integration) and 4 features (image variables) from T1- and T2-weighted image, proton density and Gadolinium-enhanced image respectively. They also provided a probability value that assesses the confidence of the prediction in each one of the voxels. A voxel might be left unclassified if its confidence was not good enough.

In [94], STE and LTE spectra were combined in an attempt to improve the discrimination between tumour types. Multiple feature selection and extraction techniques were applied, such as the sequential selection algorithm, Relief-F and PCA. LS-SVM and LDA were used for the classification task. Significant differences among performance estimations were obtained when using short, long or both TE together. More recently, Vellido et al. [95] also used the concatenation of data from both times of echo to discriminate between glioblastomas and metastases using a Single Layer Perceptron. An AUC of 0.86 was obtained using only a subset of 5 features which were automatically determined by the system. For further reviews on the application of Machine Learning and Pattern Recognition to the diagnosis of brain tumours and to cancer in general, see, for instance, [11] and [96].

Chapter 4

Ensemble learning

In Chapter 2, we described the different tumour types of the CNS, the tissues they affect, their varying degree of malignancy and some of the existing treatment techniques for the patient. Among the tumour pathologies that can be found in the brain, *glioblastomas* (*gbm*) and *metastases* (*met*) are especially sensitive due to their poor prognosis. There exists a real need to accurately differentiate these two types of tumours because the required treatment is completely different depending on the pathology. Given the difficulty to interpret indirect measurements obtained with non-invasive techniques (e.g., SV-^{1}H-MRS), reliable automatic analysis of the spectra becomes a big challenge. Up to date, most published research has failed in this task of discriminating them with acceptable success, mainly due to the similar MRS profiles that the two types present.

In this chapter, we will provide tools to overcome the limitations of previous studies by presenting a new ensemble-based technique that is able to obtain state of the art accuracy results, assessed on the established INTERPRET and eTumour data sets. We proceed by first motivating the reader with the known issues that lead to the failure of current techniques and the hypotheses on how to tackle them. Next, we provide an overview of the ensemble learning field and the different strategies used to develop each of the basic components conforming an ensemble architecture. Afterwards, we explain the workings of our proposed novel technique for the current specific

problem, followed by an empirical evaluation proving its suitability, before wrapping up the chapter with some conclusions.

4.1 Motivation

Despite many attempts to design robust models to accurately diagnose whether a patient's tumour belongs to the *gbm* or *met* type, truth is that very few of them have achieved acceptable results (Section 3.6). We conjecture that solutions based on classical single classifiers are unlikely to properly accomplish their task due to the heterogeneity on the spectral signature that these types of tumour show:

- High intra-class dissimilarity: two different tumours of the same type might present very different MRS spectra.
- High inter-class similarity: two tumours of different type might be described by very similar spectra.

We assume that algorithms able to subdivide the input space, searching for similar patterns within tumour subtypes, might be required. In that sense, ensemble techniques emerge as natural candidates to deal with this hypothesis.

However, we also assume that most of the features in a spectrum are of little relevance for the discriminating problem we face. Coupling the previous statement with the fact that most of the current cutting-edge ensemble techniques rely on random selection of features (see Section 3.2.1), makes us postulate that they will be of little help to fulfil the commended job.

Furthermore, by joining the facts of high data heterogeneity and low number of relevant features to explain the discrimination, we think that sub-grouping might be better accomplished by projecting the data into the space spanned by a subset of features instead of a subset of instances.

Hence, we hypothesize that a solution able to succeed in the current task must:

1. Present an ensemble-like structure capable of subdividing the input space, with a base learner specialised in each subdivision.

2. Each subdivision is obtained by projecting the data into the space spanned by different subsets of features.

3. An embedded wise feature selection strategy must be considered, where non-relevant features are dropped and the dimensionality is kept low.

4.2 State of the art

Chapter 3 contains a very brief introduction to the architecture of an ensemble (Figure 3.1), naming its basic components and presenting a list of the most successful solutions. Here we present a variety of traditional proposals to implement each component.

4.2.1 Base learners

The core component of any ensemble is the set of different classifiers that conform it. To fulfil their purpose, classical supervised learning methods, such as those introduced in Section 3.1.1, can be used.

Special attention must be paid to those techniques presenting high instability, meaning that a small manipulation of the learning process may end up generating completely different classification rules. ANN and DT are two examples of commonly used classifiers in ensemble learning.

Another desirable property is that each classifier must perform better than random guess, and their errors must be produced independently. This kind of learners, also known as *weak* classifiers, are the ones preferred when building ensembles. On the contrary, if their individual accuracies are under the random choice threshold, their combined outputs lead the ensemble to increase its error [97].

We also argue that the use of learners that are able to output not only a crisp classification label, but a class-conditional probability, as happens with probabilistic classifiers, should be considered; since they provide richer information than their crisp counterparts to the aggregation of outputs.

4.2.2 Aggregation strategy

Building a module that is able to retrieve the individual outputs delivered by the base learners and wisely combine them to obtain a single ensemble decision, making the most of the underlying knowledge implicit in each individual decision and the synergies among them, is of crucial importance in the design of an ensemble.

There is agreement on the fact that any aggregation technique falls within one of the two categories: either *selection* or *fusion* [98]. In *fusion* all the outputs provided by the learners contribute, in some way or another, to the final ensemble decision. When the outputs of the base classifiers are discrete (i.e., 1 or 0 whether an instance is predicted as belonging to the true class or not), *majority voting* is a frequently employed strategy. A more sophisticated technique might also compute the confidence on each classifier and use it to calculate a *weighted majority voting*.

Whenever the base learners provide continuous values (e.g., when class-conditional probabilities are outputted), aggregation methods include simple algebraic functions, such as calculating the *average* or *median*; the *weighted average* (i.e., a continuous version of weighted voting), or more elaborated techniques such as *fuzzy integral* [99], which calculates the strength or confidence of every possible subgroup of classifiers by means of a fuzzy measure to properly combine the outputs.

In the *selection* schema, only one of the base learners is used to provide the final ensemble decision. In that case a winner-takes-all strategy can be used, where the most confident learner is the one whose output is taken into account. The assumption behind it is that each learner specialises in a specific subspace, becoming an expert in this neighbourhood.

A third approach entails using a hybrid between the two strategies explained above, for instance by selecting a subset of classifiers where its outputs are combined.

Following a different classification criterion, literature splits aggregation strategies into *static* and *dynamic*, depending on the way a decision is made. Static decisions occur when they are performed using the whole training set,

without taking into account the current instance to be classified. An example of this type is *stacked generalization* [100], a technique that aggregates the base learners' outputs by using a meta-classifier that uses them as its inputs.

On the other hand, dynamic decisions are made when the local characteristics regarding the instance to be classified are taken into consideration to influence on the base learners in charge of predicting the current instance. Examples of this kind include *dynamic selection* and *dynamic voting* [101]. More precisely, they use cross-validation in the learning phase of the algorithm to assess the goodness of the prediction provided by each classifier to every instance in the training set and store this information in a table-like structure. In the prediction phase, the most *similar* case in the table is retrieved for every instance to be predicted, and its information taken into account to select the most suitable classifier (in the case of selection), or to ponder them accordingly (in the case of voting).

4.2.3 Diversity

It is sensible to say that no gain is achieved by an ensemble, as compared to a single classifier, if all the base learners return the same outputs. Therefore, there is a *common sense property* that we want our ensembles to fulfil: we want the base classifiers to disagree among them. This concept is known in the ensemble community as *diversity*.

Among the variety of strategies that has been used to ensure the existence of diversity, we present three different approaches in this section, showing examples of algorithms exploiting each of them.

The first and most intuitive is to influence directly on the base learners. This might be accomplished by using different groups of classifiers (e.g., an ensemble composed of three base learners: one ANN, one DT and one LDA), or by using different parameters (e.g., different initial conditions, or introducing other sources of randomness to the learners).

The second strategy consists in altering the training set in the learning phase of each base classifier by sampling differently from the training set. A different distribution of the instances is used for each learner leading the ensemble components to diverge. Examples of this strategy are Bagging

and Boosting algorithms (Section 3.2.1), or the cross-validation partitioning [102], where the data set is split into K folds and each classifier L_i is trained using all folds except the i-th. This i-th fold is then used as a validation set where model parameters are assessed.

The third strategy aims also at manipulating the training set, but in this case, instead of sampling different instances per learner, we focus on using different subsets of features. Each learner is specialised in a particular input subspace where certain instances are easy to classify. When the information is spread uniformly among all the features, the Random Subspace Method [54] is a good choice. It constructs the base learners by pseudo-randomly choosing the relevant features. Another study [103] used a genetic algorithm to search the best features while explicitly calculating a trade-off between accuracy and diversity. A further algorithm to be considered in this group is Input Decimation [104]. This strategy proposes selecting the features for each learner depending on the correlation with the class label while reducing the error correlation among base learners.

The importance of generating diversity has been emphasized in this section; however, we cannot forget that the final purpose of building a committee of learners is to increase their predictive ability by means of using an ensemble structure. Therefore, a good trade-off between diversity and accuracy should be actively sought for the success of ensemble classification.

4.3 Breadth Ensemble Learning

The proposed solution to overcome the limitations discussed in Section 4.1 is presented here under the name of *Breadth Ensemble Learning* (BEL). Its basic structure, following the classical ensemble architecture previously mentioned, can be seen in Figure 4.1 and includes: a *diversity generator* module, called *feature search* since this module not only provides diversity, but also divides the problem in sub-partitions of the input space; an *ensemble induction* module containing the set of *base learners*; and the *aggregation strategy*.

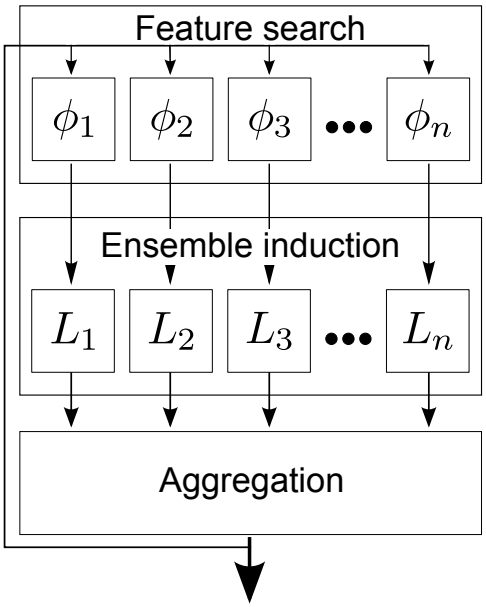


Figure 4.1: Breadth Ensemble Learning structure - The system consists of 3 main modules: the *feature search* (composed of different subsets of features ϕ_i), the *ensemble induction* (made of several base classifiers L_i) and the *aggregation strategy*.

We first introduce each component in turn; their relationships will then be explained to show the functioning of its workflow.

4.3.1 Base learners

The current component is built using a set of classifiers, these being of any type from the available palette of techniques, or a combination thereof. Nevertheless, some requirements led the choice of strategies. First, given that we wanted to employ learners that were shown to be effective in the domain under investigation, the *feature selection* module was kept to be the only responsible for generating diversity; hence all the chosen learners in the ensemble solution were of the same type and parameters.

Secondly, with the purpose of easing the *aggregation* phase, our predictors were required to have the ability to provide soft decisions in form of posterior probabilities, so as to obtain not only a crisp class label, but a qualitative measure of the prediction; therefore, probabilistic solutions

became the first choice. More precisely, several classifiers of this type of various degrees of complexity were used: the arguably weak-learner Naive Bayes (NB [56]); the preferred single learner in the domain [105], LDA; and the more complex Quadratic Discriminant Analysis (QDA [106]), both using their probabilistic interpretation [56].

Due to its success in many fields, state of the art SVM was also employed. Specifically, the linear LS-SVM [107] version with the add-on proposed by Platt [108], where distances between support vectors and instances are fitted into a sigmoid function to be used as an approximation to posterior probabilities.

The last learning algorithm of interest, given that we are developing a solution within an ensemble context, is the weak DT. Specifically CART [66] was used, and the posterior probabilities were computed as the prevalences in the final nodes.

4.3.2 Aggregation strategy

The main condition that the aggregation strategy was expected to fulfil was simplicity, since we wanted to focus our attention in the *feature selection* component; another condition was to take advantage of the fact that base learners output continuous values in form of probability, which can be interpreted as the degree of membership to the positive class that the current instance presents. Consequently, a static fusion aggregation strategy was chosen; more specifically, the arithmetic mean [98] was employed to combine individual outputs from each base classifier into a single ensemble prediction. This simple measure provides a global ensemble decision which is also a probability, expressing the degree of membership that the whole committee assigns to the current prediction.

4.3.3 Diversity by feature selection

The feature selection component is very important in the proposed system, since all the diversity for the ensemble's success is generated there. Also, the appropriateness of the input subspace projection directly depends on it.

Assuming that the component is made up of N FSS modules: each module generating a subset of features per base learner, we decided to implement a Sequential Forward Generation algorithm (SFG) (see Section 3.3) to obtain the proper subset of features for a given learner. Following the steps introduced for this kind of methods:

1. *Subset generation* was implemented by defining three operands: adding one feature to the current subset of features; removing one feature from the subset; or leaving the subset unchanged.

2. *Evaluation criterion* was defined as the prediction ability (measured by a chosen metric) of the whole ensemble on a validation set. Notice that, by using this approach, the proposed SFG falls within the *wrapper* category.

3. *Stopping criterion* was set to flag when no increase in ensemble prediction ability occurred within two consecutive iterations.

Results should be validated by measuring ensemble prediction ability on a test set; and, in our specific domain of application, by comparing the selected features with radiologists' knowledge on relevant MRS frequencies.

It is important to emphasize some singularities of the algorithm, given the fact that it is embedded into an ensemble structure: first, at each iteration, given a specific module, one operation is performed per feature, but only the operation that maximises the overall ensemble performance is kept; secondly, all modules are updated by one feature in turn at every iteration, leading to a construction of the ensemble in breadth.

A justification for the functioning explained above is as follows: on the one hand, a *forward* feature selection method was chosen due to the low number of relevant features present in the MRS data for the current discriminative task. Another reason for this decision was that, given the high number of features and small sample size that often occur in the current domain of application, LDA and QDA need to invert covariance matrices which turn to be singular in this setting. On the other hand, approaching the feature subset selection on a *breadth* basis is explained as the result of

allowing each base learner to select the preferred dimension such that, when added to its current dimensionality, the resulting data projection aids certain subgroups of data to be better classified, measured as an increment in the overall ensemble performance.

4.3.4 Algorithm's workflow

In order to better understand the BEL algorithm, the procedure is shown here and some advice on implementation issues is provided.

Let Θ be the full set of features and let N denote the number of base classifiers (which is constant). We denote by $L_i(\phi)$ the i-th base classifier developed using the feature subset ϕ. The ensemble at time (iteration) t can then be expressed as $\mathcal{L}(t) = \{L_1(\phi_1(t)), \ldots, L_N(\phi_N(t))\}$, where $\phi_i(t) \subseteq \Theta$.

The algorithm starts by assigning one feature to each module in the feature search component. Notice that, given that the feature updating works in batch mode (i.e., modules are updated at the end of each ensemble's iteration), it is important that the initial selected feature is different in each module. Moreover, due to the fact that subset generation is approached greedily, the algorithm is prone to be trapped in local minima; hence, starting from an advantageous status is advisable. We propose to use the fast classical RelievedF [58] filter algorithm to rank the features that best separate our two classes and sequentially assign the best remaining feature to each module.

Then, every base learner is trained using all data in the training set exploiting the characteristics exhibited by the data in the space spanned by the selected features. A validation set is employed by each base classifier to estimate a continuous output representing the membership of every instance to each class, and finally, all the outputs provided by the classifiers are aggregated to obtain a single ensemble prediction.

The resulting predictions provided by the ensemble are compared with the true class label of the validation set and the ensemble performance P is assessed (e.g., by using the AUC), by which the ensemble iteration is completed.

Next, subsequent iterations consist in finding best candidate updates to every module in order to build the whole ensemble. Specifically, to form the next ensemble $\mathcal{L}(t+1)$ from $\mathcal{L}(t)$, we proceed as follows. For the i-th base classifier, three possibilities are considered: add the best feature to $\phi_i(t)$, remove the worst feature from $\phi_i(t)$, or leave $\phi_i(t)$ unchanged. The choice that leads to the highest *overall ensemble* performance will be selected. The best feature $B_i(t+1)$ for L_i is the feature that, when added to $\phi_i(t)$, leads to the best ensemble performance:

$$B_i(t+1) = \underset{\theta \in \Theta \setminus \phi_i(t)}{\arg\max}\, \mathrm{P}(\{L_1(\phi_1(t)), \ldots, L_i(\phi_i(t) \cup \{\theta\}), \ldots, L_N(\phi_N(t))\})$$

where P is the ensemble performance measure. Conversely, the worst feature $W_i(t+1)$ for L_i is the feature that, when removed from $\phi_i(t)$, leads to the best ensemble performance:

$$W_i(t+1) = \underset{\theta \in \phi_i(t)}{\arg\max}\, \mathrm{P}(\{L_1(\phi_1(t)), \ldots, L_i(\phi_i(t) \setminus \{\theta\}), \ldots, L_N(\phi_N(t))\})$$

Then, candidate $\phi_i(t+1)$ is set to either $\phi_i(t) \cup \{B_i(t+1)\}$, $\phi_i(t) \setminus \{W_i(t+1)\}$ or $\phi_i(t)$, depending on which choice leads to the best performance when $L_i(\phi_i(t+1))$ is used. This process to find the best candidate updating is repeated for all the base classifiers to form $\mathcal{L}(t+1)$. Changes are applied at the end of the iteration, when best candidate updates have been found for all the base classifiers. The reason for employing a batch mode is purely to improve computational speed.

The iterative process shown above continues until the stopping criterion is met. A Matlab implementation of the presented algorithm can be found at `http://www.cs.upc.edu/~avilamala/resources/BEL_Toolbox.zip`

4.4 Experimental evaluation of the proposed method

The proposed BEL algorithm was created with the problem of improving the predictive discriminatory capability between *gbm* and *met* in mind. Here, its suitability for such task is assessed and results are compared to single

classifiers and classical ensemble techniques. A discussion on the biological plausibility of automatically retrieved features as well as a theoretical interpretation of technical issues is provided.

4.4.1 Experimental setup

The data used to evaluate the proposed technique (Section 2.3) corresponds to a subset of the INTERPRET database, which is composed of 78 *gbm* and 31 *met* to be used as training set; and 30 *gbm* and 10 *met* from the eTumour project that conform the hold-out set.

Only 195 out of 512 available frequencies, validated by experts as corresponding to the most relevant frequency interval in the spectrum [32], are used in the current experiments. Data acquired at both LTE and STE are employed by concatenating both spectra (LTE + STE, 390 features); this setup has been shown in previous studies [94] to have a differential advantage for classification purposes. All data have been standardised prior to analysis.

The training phase in any of the experiments consisted in applying a leave-one-out cross-validation technique (LOO-CV, stated otherwise) over the training set (using the corresponding class labels) for parameter estimation and model selection, aiming at maximising the AUC measure of the whole ensemble; while the hold-out set was used to validate the ensemble performance in the prediction phase.

4.4.2 Single classifier vs. ensemble

This test consisted in assessing the appropriateness of each learner type in becoming the base learner of choice for the BEL algorithm, by picking the best performing one. Likewise, a comparison between single classifier versus its ensembled counterpart was also evaluated.

The first choice involved the selection of the hyperparameter controlling the number of base learners for BEL, which was set to 50.

The deterministic filter RelievedF algorithm was set to make use of only the nearest neighbour, which means setting the K parameter to 1. Prelim-

inary evaluations showed no significant difference in using more neighbours for the purpose of initialising BEL.

For probabilistic learners (i.e., NB, LDA and QDA), priors were set empirically as class proportions; μ was set to the empirical mean; σ^2 in NB was set to the empirical variance, Σ as the empirical covariance matrix for LDA, and Σ_c as the empirical class-conditional covariance matrices in the QDA learner.

Regarding LS-SVM, a linear kernel was chosen, setting the C parameter to default 1 value. In the case of CART, no pruning was set and the prior probabilities were set to be the class proportions. For the rest of parameters, defaults were also used: they include setting $k = 10$, which corresponds to the required number of instances per impure nodes to split, minimum number of observations per leaf equal to one, and using the Gini index [66] to guide the splitting.

The results obtained by each classification technique are summarised in Table 4.1, where different evaluation measures are presented. Notice that, given the deterministic nature of the used techniques (i.e., source of randomness in neither classifiers nor LOO-CV strategy), only point values are provided in the table, with the only exception being CART, which required 10-fold CV due to the excessive computation time involved in applying LOO.

The general pattern followed by all evaluated learning techniques is the improvement of predictive performance achieved when using an ensemble architecture as compared to a single classifier, fact that reinforces our hypothesis that BEL predicts better than single classifiers. If we now focus our attention on the best performing base classifier according to our results, we can conclude that LDA should be the learner of choice in our setting. This robust linear classifier yields better classification than the weak NB or CART, but, interestingly, also outperforms the more complex QDA. Linear LS-SVM also operates quite well and could therefore be considered an alternative choice.

Table 4.1: Breadth Ensemble Learning performance using different base classifiers

	n	AUC	AUH	ACC	F	BER
NB	1	0.59	0.68	0.80	0.80	0.40
	50	0.61	0.74	0.85	0.87	0.33
LDA	1	0.79	0.83	0.82	0.86	0.35
	50	**0.88**	**0.91**	**0.87**	**0.88**	**0.22**
QDA	1	0.58	0.68	0.77	0.79	0.37
	50	0.61	0.72	0.77	0.81	0.47
LS-SVM	1	0.68	0.76	0.80	0.86	0.35
	50	0.84	0.88	0.82	**0.88**	**0.22**
CART	1	0.58 ± 0.07	0.58 ± 0.07	0.75 ± 0.00	0.78 ± 0.05	0.46 ± 0.10
	50	0.65 ± 0.06	0.74 ± 0.04	0.81 ± 0.02	0.83 ± 0.02	0.37 ± 0.03

The learning techniques used as base classifiers were Naive Bayes (NB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Least-Squares Support Vector Machines (LS-SVM) and Classification and Regression Trees (CART). The ensemble performance was calculated using different measures: Area Under the ROC Curve (AUC), Area Under the ROC Convex Hull (AUH), accuracy (ACC), F-measure (F) and Balanced Error Rate (BER). The ensemble was composed of either 1 or 50 base classifiers.

4.4.3 Breadth Ensemble Learning *vs.* classical ensembles

The predictive performance of BEL was compared to state of the art general purpose ensemble techniques. Specifically, we evaluated Random Forests (RF), Bagging (Bag) and Boosting (Boost). The number of grown trees was set according to the authors' advice (i.e., 500 for RF, 100 for the other techniques). The rest of parameters were evaluated empirically:

For RF, we used 20 features per node, which approximately corresponds to the square root of the total number of features; in Bag, the maximum number of instances per node before splitting was set to 20 and increasing its fit by 0.5; finally, Boost was set as its predecessor but the parameter controlling the fit increment was set to 0.4.

Posterior probability values were calculated as the quotient of trees voting in favour of positive class over total of trees.

The first row in Table 4.2 shows the poor results obtained by the evaluated general purpose ensemble techniques. Given that one of our hypotheses in this chapter is that appropriate feature selection is required, we tried

Table 4.2: Performance of different ensemble methods on ^{1}H-MRS data

FSS	Ens.	AUC	AUH	ACC	F	BER
None	RF	0.67 ± 0.01	0.77 ± 0.01	0.77 ± 0.02	0.86 ± 0.01	0.44 ± 0.07
	Bag	0.69 ± 0.04	0.72 ± 0.04	0.75 ± 0.05	0.83 ± 0.04	0.35 ± 0.04
	Boost	0.71 ± 0.02	0.78 ± 0.02	0.74 ± 0.04	0.83 ± 0.03	0.36 ± 0.03
RelievedF ($m = 14$)	RF	0.59 ± 0.02	0.68 ± 0.02	0.75 ± 0.00	0.86 ± 0.00	0.50 ± 0.00
	Bag	0.62 ± 0.03	0.70 ± 0.03	0.73 ± 0.03	0.83 ± 0.02	0.39 ± 0.03
	Boost	0.62 ± 0.04	0.68 ± 0.03	0.76 ± 0.03	0.85 ± 0.02	0.40 ± 0.05
RF ($m = 23$)	RF	0.67 ± 0.01	0.74 ± 0.01	0.78 ± 0.02	0.86 ± 0.01	0.35 ± 0.02
	Bag	0.71 ± 0.04	0.73 ± 0.04	0.75 ± 0.06	0.83 ± 0.05	0.34 ± 0.04
	Boost	0.72 ± 0.02	0.77 ± 0.02	0.72 ± 0.04	0.81 ± 0.03	0.38 ± 0.03
Embed.	BEL	**0.88**	**0.91**	**0.87**	**0.88**	**0.22**

The proposed ensemble techniques are Random Forest (RF), Bagging (Bag) and Boosting (Boost) using CART, which were run with no feature selection prior to classification or with either RelievedF or RF (ending up keeping 14 and 23 features, respectively), and the proposed Breadth Ensemble Learning (BEL) using LDA as base learners. The ensemble performance was calculated using different measures: Area Under the ROC Curve (AUC), Area Under the ROC Convex Hull (AUH), accuracy (ACC), F-measure (F) and Balanced Error Rate (BER).

to help these algorithms by performing feature subset selection as a preprocessing step. More precisely, we used two different techniques to rank the features: using RelievedF filter with parameter $K = 1$ and by calculating the averaged Gini index of each feature after 100 RF runs. The final number of features m was chosen according to the *elbow criterion* [43]. None of these attempts helped to increase the predictive power of the models, as observed in rows 2 and 3 of the table.

4.4.4 Discussion

In light of these results, we conclude that regular ensemble learning algorithms based on random selection of features and composed of weak base learners (which are proved to be successful in many domains) do not perform well in our field. Trying to overcome such limitation by using a wiser feature selection strategy does not help in accomplishing the task.

The BEL algorithm achieves its goal by training the base learners using different subsets of features, which have been obtained by means of a parsimonious FSS strategy. In this context, sharing a feature between two

subsets is not prevented; a feature can be freely shared among as many subsets as required.

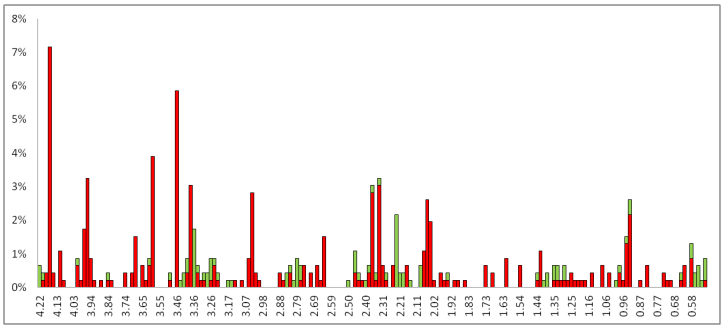


Figure 4.2: Single Voxel ^{1}H-MRS frequency appearances - Relative percentage of appearances for each feature (frequencies in ppm) from the SV-^{1}H-MRS spectrum using a BEL ensemble of 50 LDA. Deep red columns represent appearances in the Long Time of Echo spectrum whilst light green columns are the appearances in the Short Time of Echo.

With the purpose to show the relevance of each feature (frequencies in *ppm* from the SV-^{1}H-MRS spectrum) for the current discriminative task according to BEL standards, the relative percentage of feature appearances in our successful BEL classifier made up of 50 LDAs as base learners is shown in Figure 4.2. They must be compared and contrasted with results from existing literature, as well as domain knowledge, for enforcing the model's reliability and pointing towards new findings in form of relevant frequencies within the spectra.

Most of the highly selected features are consistent with those found relevant in previous studies. For instance, frequencies located between $3.38 - 3.45ppm$, which have been selected as relevant by our method in LTE, might correspond to Taurine as shown in [23]. Similarly, those in the interval $3.58 - 3.60ppm$, corresponding to Glycine, have also been picked up by both studies. A well-known important metabolite, namely Creatine, usually observed at $3.03ppm$ is properly captured by BEL. N-Acetyl Aspartate, at $2.05ppm$, has been selected by our model: a metabolite of interest for this specific discriminative task, as reported in [95]. This same study also observed the prevalence of important features at LTE with respect to

STE when both spectra are used in concatenation.

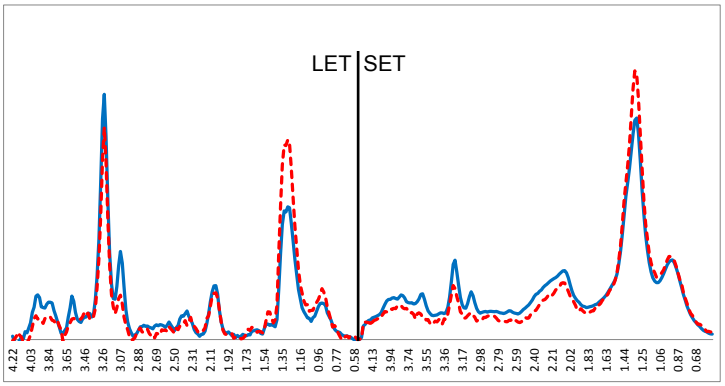


Figure 4.3: Average glioblastoma and metastasis spectra - Mean spectra as a function of frequency (in ppm) of *gbm* (solid blue) and *met* (dashed red) from the INTERPRET database, for both long and short echo times (LTE and STE, respectively).

Comparing the most selected frequencies in BEL with the mean spectra in Figure 4.3, we observe that frequencies showing high amplitude in the latter are not necessarily relevant, as appreciated in the former. In this respect, notice the Choline ($3.20ppm$) and Lipids/Macromolecules ($1.40ppm$) compounds, which present the highest peaks in the spectra, but are rarely selected by our method.

New and previously unreported findings arise in our study, according to the high occurrence of features located at $4.20ppm$ and $3.95ppm$, which might correspond to Choline and either Creatine or Alanine. They should be taken into account in future research to elucidate whether they may become consistent biomarkers.

We would not like to finish this discussion without commenting a few technical aspects. The first one is the computational cost of the BEL algorithm: we acknowledge the limitations of our approach in terms of time complexity, given the number of learners that have to be trained every time a new feature is considered. Nonetheless, taking the difficulty of the current goal into consideration (i.e., discriminating between *gbm* vs. *met*), together with the current situation in which no adequate models exist to solve such discrimination, and added to the fact that BEL is expected to work in an

offline environment, the benefits of accurate tumour classification using BEL are worth the price of slow model building.

Besides, if BEL was to be applied in other domains which require faster model learning, we should keep in mind that the algorithm has been designed to achieve easy parallelisation with the purpose of alleviating the time complexity bottleneck. Specifically, two levels of parallelisation can be applied: the first consists in running the trial of every new feature in each base learner distributedly; the second takes advantage of the batch mode updating of subsets (i.e., candidate features per base module are not updated until all modules have been treated), allowing module-basis parallelisation.

A final remark has to do with the bias-variance analysis of BEL's error improvement: according to [109], stable classifiers like LDA characteristically present low variance but can have high bias. Given the heterogeneity on the spectral signature that the tumours under consideration present (see Section 4.1), it seems plausible that single learners might show high bias. Following this line of reasoning, since every base learner specialises in a subspace, they better capture the singularities of their specific subproblem, hence reducing bias. However, further research including theoretical analysis of bias-variance decomposition of BEL should be carried out to validate this interpretation.

4.5 Conclusions

The proposed *Breadth Ensemble Learning* is a technique that builds a committee of experts with an embedded feature selection strategy specifically designed to overcome the limitations of current solutions attempting to discriminate glioblastomas from metastases using SV-^{1}H-MRS data. It has been conceived following the premises stated in Section 4.1, which we review next:

1. *Present an ensemble-like structure capable of subdividing the input space, with a base learner specialised in each subdivision*: stable LDA models have been used as base learners to become an expert in their sub-domain, which provide probabilistic outputs to better interpret

the reliability of the decision made, while allowing for a straightforward ensemble integration by means of averaging.

2. *Each subdivision is obtained by projecting the data into the space spanned by different subsets of features*: the data being projected in different subspaces not only allows the base learners to specialise, but also to provide the diversity that the ensemble requires. Notice that subspaces are created in breadth; that is, the new added dimension is the one that best improves the overall ensemble predictive capability.

3. *An embedded wise feature selection strategy must be considered, where non-relevant features are dropped and the dimensionality is kept low*: a sequential forward feature selection algorithm is chosen to take advantage of the low number of relevant features for the commended task. This strategy is tightly coupled with the subsequent base learner, working in a wrapper fashion.

The good results obtained in our benchmark to differentiate these two aggressive tumours rank with the best obtained to date for this kind of problem, analytically reinforcing the validity of our hypotheses.

Chapter 5

Stability of feature selection

It is widely accepted that ML models must provide high prediction accuracy. To fulfil this requirement, BEL followed, in the previous chapter, an ensemble approach with a wise FSS strategy to improve prediction ability in the specific problem of discriminating between *gbm* and *met* brain tumours. This is, however, a perfect example of a problem where providing interpretable outputs is as important (if not more important than) as achieving high classification accuracy. In such situations, FSS is not only useful from a technical viewpoint (i.e., assuaging the curse of dimensionality), but also by providing more interpretable models: a human radiologist will better understand the model's output when few features (i.e., SV-^{1}H-MRS frequencies in the aforementioned problem) have been employed to provide a decision.

Nonetheless, simply applying FSS techniques to our problem is not enough to obtain interpretable models that can be trusted by domain experts. An important hurdle in this respect relates to the instability of FSS algorithms: if little variations in the input data translate into a different selection of features considered relevant, the reliability on the model is hampered regardless its relative predictive accuracy. This phenomenon not only occurs in the domain under investigation, but in most of the situations in which we deal with few observations of high-dimensional data.

In this chapter, we stick to the problem of discriminating between *gbm* and *met* from SV-^{1}H-MRS data, but acting upon FSS algorithms to induce them to provide more stable subsets of important features at different

runs while maintaining their predictive power. We start by showing some properties of the domain data, which, together with the current technical limitations, set the premises over which to develop the new technique. Then, we review the literature, searching for the last reported improvements regarding feature subset stability. Later, the proposed technique for explicitly improving feature subset stability is presented, experiments on datasets from different domains are reported and finally, some final conclusions are summarised.

5.1 Motivation

When models capable of accurately discriminating between *gbm* and *met* using SV-^{1}H-MRS data were supplied to medical experts, they showed their scepticism about models' reliability, given the fact that FSS strategies often choose very different subsets of features as relevant for the classification every time the model is executed. The cause of such variability can be attributed to the own instability of FSS algorithms, an event often observed when using datasets containing:

- a small number of instances (small sample size),
- many features (high dimensionality).

In our domain, the number of spectra is in the order of tens per tumour type, while the number of SV-^{1}H-MRS frequencies is in the order of hundreds. In such circumstances, the hypotheses space is too large, while the number of constraints (i.e., instances) is limited, meaning that different configurations (i.e., subsets of features) might equally approximate the real hypothesis, leading to model overfitting.

As explored in the next section, the few studies addressing this problem mainly approach it through resampling strategies used to construct ensemble-based FSS models; the principal shortcomings of this approach are its high computational cost and the lost of learning capacity when sampling from an already small dataset. An alternative approach is grounded in the statistical concept of *importance sampling* [110]. This is an appealing

approach to be taken into consideration in our domain, due to its simplicity and efficiency. However, and despite sound formal analysis, empirical evaluation on the provided framework shows a number of limitations that need to be overcome.

We will therefore base our study on two main hypotheses:

1. There are some instances that are typical regarding their underlying distribution and others that show outlying behaviour. Due to the fact that the latter type induce FSS algorithms to be unstable, we could simply remove them for the sake of stability if the dataset is large enough. In our case, with a small sample size, we can obtain a similar effect by weighting their importance in the FSS process.

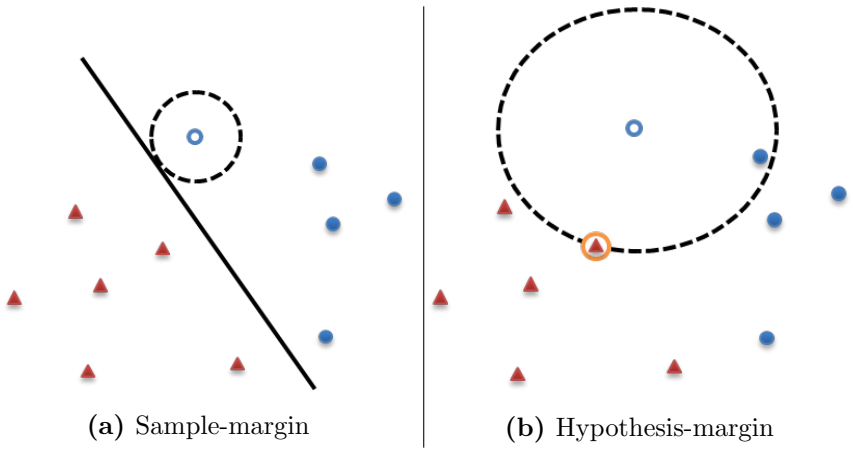
2. As discussed in the previous chapter, the high heterogeneity of instances lead them to cluster in local neighbourhoods. This means that FSS algorithms approaching the *hypothesis-margin* are likely to be more suitable than the ones aiming at reducing the *sample-margin*, as seen in the next section.

5.2 State of the art

The two types of margins mentioned previously are the first topic addressed in the current section. This is followed by the review of two widely used FSS algorithms, which turn out to be easy to adapt to the inclusion of information regarding instances' typicality. Next, we introduce several state of the art measures to evaluate the stability of FSS techniques. Last, a thorough survey of the few studies devoted to methods for explicitly increasing FSS stability is carried out.

5.2.1 Sample and hypothesis margins

In a research project published in [111], margins were defined as important elements to measure the confidence of a classifier with respect to its predictions. Specifically, the work exposes two different approaches to characterise the margin (or confidence) of a given instance:



(a) Sample-margin **(b)** Hypothesis-margin

Figure 5.1: Two types of margin - Radius of dashed circles represent the two types of margins for the hollow blue dot.

One is named *sample-margin* and is described as the distance between an instance and the decision boundary induced by the classification rule. Examples of learning algorithms using this type of margin include SVM, where a separating hyperplane that maximises the sample-margin is sought, or the K-NN version that defines a Voronoi tessellation.

Hypothesis-margin is the second type of margin, defined as the distance between the given instance and the closest hypothesis that assigns an alternative label to the given instance. This type of margin can be more easily computed in a different K-NN version than the previous one, by simple Euclidean distance calculation instead of requiring the generation of a tessellation.

Regarding FSS techniques, SVM-Recursive Feature Elimination (SVM-RFE) [112] is an example of algorithm of the sample-margin type, while the Relief family of algorithms falls into the hypothesis-margin type. See the next section for a review of these algorithms.

5.2.2 Feature selection techniques

In Chapter 3, we gave a broad introduction to FSS techniques and presented their three main types: filter, wrapper, or embedded methods. A different classification can be done if we focus on the output that FSS strategies

provide. In this respect, we might obtain a *set* of relevant features, all having the same importance; an ordered *list* of features, where features at the top of the list are more relevant than those ones at the bottom; and, finally, *weighting-score* features, where a quantification of the importance of each feature is provided. Notice that each of the subsequent presented output-types introduces one more level of information. Due to the requirements of our working domain (i.e., searching for equally relevant biomarkers), we stick to the use of the more general *set* of features.

The Relief family of filter FSS algorithms, employing the hypothesis-margin, aim at weighting each of the available features according to its relevance regarding the target concept. Its first version was presented in [61], consisting in randomly sampling P instances from the training set and updating the weight W of each feature j according to the distance of the selected instance $\mathbf{x}_i$ to the closest instance of different (nearest miss: $m(\mathbf{x}_i)$) and same (nearest hit: $h(\mathbf{x}_i)$) class. Feature weights are calculated as:

$$\mathrm{W}(j) = \sum_{i=1}^{P} \left(|\mathbf{x}_{i,j} - m(\mathbf{x}_i)_j| - |\mathbf{x}_{i,j} - h(\mathbf{x}_i)_j| \right).$$

This idea was further extended in [113], developing a more robust algorithm to deal with noisy data by averaging the distance to the K nearest hits and K nearest misses. The solution was named Relief-A:

$$\mathrm{W}(j) = \sum_{i=1}^{P} \frac{1}{K} \sum_{k=1}^{K} \left(|\mathbf{x}_{i,j} - m_k(\mathbf{x}_i)_j| - |\mathbf{x}_{i,j} - h_k(\mathbf{x}_i)_j| \right).$$

In that same study, Relief-F was proposed as a generalisation for multiple class prediction.

With the purpose of reducing variance due to the stochastic nature of Relief techniques, a deterministic version was proposed in Relieved [58], which proposed to use all N instances in the training set exactly once, instead of sampling from it and computing the distances to all hits and misses. An extension, defined to obtain a deterministic multi-class algorithm, was introduced under the name of Relieved-F [62].

Another interesting algorithm, this time using a classifier of the sample-margin approach with embedded FSS, is the SVM-RFE [112]. Its main idea

consists in using the weights of a maximum margin classifier to produce a feature ranking. In its initial version, a linear soft-margin SVM classifier is trained by minimising the following objective function:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i,$$

where ξ is a vector of slack variables or deviations from the hyperplane; C is the hyperparameter that controls the trade-off between separating with maximal margin and allowing misclassifications; and

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i$$

is the weight vector, $\alpha, y, \mathbf{x}$ being the Lagrangian parameters, class labels (s.t. $y_i \in \{-1, 1\}$) and instances, respectively. Once convergence is achieved, the weight vector is used to compute the ranking criterion for each feature as $c_j = |\mathbf{w}_j|$.

Notice that the proposed FSS techniques do not really provide a subset of selected features, but a weighting-score of features, instead. In order to obtain a subset of relevant features, an RFE strategy can be adopted by iteratively applying the proposed algorithms and removing the lowest ranking features.

5.2.3 Measures for assessing feature selection stability

The suitable figure of merit to evaluate the stability of a FSS technique will depend on the output-type the algorithm provides. There exist measures specifically designed to assess stability between feature rankings, feature scores and feature sets. In this part, we focus our attention on this latter type and review the most frequently used measures.

The matter of evaluation is the stability of a FSS algorithm in selecting a subset of k features out of the initial F features over a batch of M runs. Let $S_i(k)$ be the subset of selected features of length k in the i-th run; and $\mathcal{E} = \{S_1, S_2, ..., S_M\}$ the set containing all the retrieved feature subsets. The

first metric, termed Average Normalised Hamming Distance [114], makes use of the pairwise information theoretic Hamming distance:

$$\mathrm{HD}\left(S_i(k), S_j(k)\right) = \sum_{f=1}^{F} |S_{i,f}(k) - S_{j,f}(k)|,$$

which is averaged over the M runs to calculate the overall stability in $\mathcal{E}$, according to:

$$\mathrm{ANHD}\left(\mathcal{E}(k)\right) = \frac{2}{F \times M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathrm{HD}\left(S_i(k), S_j(k)\right).$$

This equation outputs unity-bounded values, ranging from low stability ($\approx$ 0) to high stability as we approach a value of 1. Its main drawback is that it does not account for the amount of intersection between two subsets.

Kalousis et al. [72] proposed to use the Tanimoto coefficient, which is a generalised Jaccard Index for dissimilarity between two subsets:

$$\begin{aligned} \mathrm{JI}\left(S_i(k), S_j(k)\right) &= \frac{|(S_i(k) \cap S_j(k)|}{|(S_i(k) \cup S_j(k)|} \\ &= 1 - \frac{|S_i(k)| + |S_j(k)| - 2|S_i(k) \cap S_j(k)|}{|S_i(k)| + |S_j(k)| - |S_i(k) \cap S_j(k)|}. \end{aligned}$$

This measure is also bounded between 0 and 1, the former meaning no intersection while the latter implies the two subsets to be the same.

Kuncheva [73] introduced the stability index (a.k.a. Kuncheva Index, KI) of $\mathcal{E}(k)$ by computing the average of pairwise consistency index:

$$\mathrm{KI}\left(\mathcal{E}(k)\right) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathrm{KCI}\left(S_i(k), S_j(k)\right),$$

where

$$\mathrm{KCI}\left(S_i(k), S_j(k)\right) = \frac{|S_i(k) \cap S_j(k)| - (k^2/F)}{k - (k^2/F)}.$$

KI values are bounded to values between -1 and 1, the latter meaning maximum stability. Values near 0 are interpreted as similarity drawn by chance, and negative values show high dissimilarity (more than random).

Similarly, using the same averaging equation, but substituting the consistency index (KCI) by the Jaccard Index (JI), Alelyani et al. [115] extended the Kalousis' similarity to multiple subsets.

Krizeck et al. [116] developed a stability measure based on the Shannon entropy, a concept borrowed from information theory:

$$\mathrm{KSE}\left(\mathcal{E}(k)\right) = -\sum_{i=1}^{C(F,k)} \overline{G}_i(k) \log_2 \overline{G}_i(k)$$

where $\overline{G}_i(k) = \frac{G_i(k)}{M}$; $G_i(k)$ being the number of occurrences of the set $S_i(k)$ in the sequence of M subsets of size k; and $C(F,k) = \binom{F}{k}$ the number of all possible subsets of size k from F. Its values range from a minimum stability of 0 to a maximum of $\log\left(\min\left\{M, C(F,k)\right\}\right)$.

Finally, Somol and Novovičová [74] made a thorough review on existing FSS stability measures, evaluated them and provided some modifications and improvements. They also proposed a new measure, Relative Weighted Consistency (RWC), aiming at achieving a unified measure while solving some of the limitations they found on the reviewed ones:

$$\mathrm{RWC}(\mathcal{E}) = \frac{F\left(A - V + Z\right) - A^2 + V^2}{F\left(W^2 + M(A - W) - V\right) - A^2 + V^2}$$

where A is the total number of occurrences of any feature in system $\mathcal{E}$; $V \equiv A \pmod{F}$; and $W \equiv A \pmod{M}$; and $Z = \sum_{f \in F} H_f(H_f - 1)$; H_f being the number of occurrences of feature f in system $\mathcal{E}$. The metric is bounded to values between 0 and 1 and is able to compare subsets of different size.

Among all the provided measures to evaluate FSS stability, KI will be the one to be used in this thesis, which, despite not showing the best properties (e.g., it is limited to subsets of same size), it is easy to interpret and matches the requirements of our problem. Moreover, the fact that it is the most widely used in the literature allows us to provide a direct comparison between our proposed approach and previous studies.

5.2.4 Previous studies on improving feature selection stability

According to the literature, few works address the problem of explicitly improving the stability of FSS techniques. Next, we review the most prominent ones.

One of the first research studies tackling the aforementioned problem was conducted by Saeys et al. [75]. They propose to stabilise the output of FSS strategies by means of ensemble feature learning. Similarly to the theory of ensemble learning exposed in Chapter 4, the goal is to build a committee of feature learning algorithms and aggregate their output. More precisely, four different ensembles made up of several feature learners from one of the following types were proposed: Relief and Symmetrical Uncertainty [117], from the filter family; and Random Forests and SVM-RFE, as embedded candidates. Required diversity was achieved by instance perturbation using bootstrap samples, and the aggregation strategy was either weighted average for feature rankings or voting for subsets of features. Experiments on Deoxyribonucleic Acid (DNA) microarray and mass spectrometry datasets provided evidence of the benefits in feature stability when the proposed ensemble feature learning was used, Relief being the least stable algorithm which most benefits from using the new strategy. Classification performance of all methods remained comparable to that of single FSS.

Another interesting study was presented in [118]. Its solution is based on two assumptions: first, the observation that in sample space, regions showing high density (as measured by probabilistic density estimation) are stable with respect to the features selected; second, that features near the core of high-density regions are highly correlated to one another and, therefore, should have similar relevance with respect to class labels; hence, they should be treated as a single group when ranking features. Having these premises in mind, the Dense Relevant Attribute Group Selector (DRAGS) framework is proposed. It consists of two main steps: finding dense instance regions (applying the Dense Group Finder –DGF– algorithm) and deciding their relevance. DGF uses the multivariate kernel density estimator [119] to evaluate the density of each feature; then, a number of unique density peaks in the data are identified using the mean shift procedure [120]; afterwards, dense features close to the same density peak are grouped together. The second step consists in finding the relevance of each feature group by averaging the relevance of features within the group according to the F-statistic. Finally, once relevance groups have been selected, one representative feature

per group (the one with the highest average similarity to all other features in the group) is picked up. The suitability of this method was assessed in experiments concerning DNA microarray data.

Authors acknowledged two main limitations of DRAGS: first, identifying dense feature groups using high dimensionality and low sample size makes density estimation difficult and unreliable; second, the algorithm might miss some of the most relevant individual features if they are located in the sparse region of the data distribution. With the purpose to overcome them, they published an improvement [76] called Consensus Group Stable (CGS) feature selection. An ensemble made up of DGF modules was constructed, using bootstrap samples from the data as a strategy to generate diversity, borrowing the *instance-based aggregation approach* from ensemble clustering. This algorithm models each feature as an entity and decides the similarity between each pair of instances based on how frequently they are grouped together. Moreover, when CGS computes the similarity of every feature pairs, agglomerative hierarchical clustering is applied to group features into a final set of consensus feature groups. As in DRAGS, the last step includes selecting a feature candidate from each group (i.e., the closest feature to the group centre) and determining the group relevance. In contrast to DRAGS, all consensus groups in this solution are taken into account during the relevance selection phase.

In a work published in [121], two main shortcomings of most current ensemble feature selection methods were identified: they do not account for interactions among features and they are not able to provide more than one equally suitable feature set. Algorithms capable to fulfil this second requirement might supply insight into the problem under investigation by showing different *viewpoints* (different, equally important sets of features). The research addresses these issues by studying current aggregation strategies and developing new ones in an ensemble environment similar to the ones exposed previously. In particular, they differentiate between Single Model Aggregation Strategies, where a unique feature set is provided, and Multiple Model Aggregation Strategies (MMAS), where several feature sets are outputted.

The former can be split, in turn, into three categories: Univariate Strategies, containing typical aggregation strategies such as voting and averaging, which are advisable for univariate FSS strategies; Model Component Combination Strategies (MCCS), which use Frequent Itemsets Mining [122] to detect subsets of features often appearing together in several base learners and bring them together to produce the final solution; and Exact Structure Preservation Strategies (ESPS), that select the best candidate from the feature subsets (often a median measure after applying a clustering algorithm to candidate subsets) generated by the base learners. The MMAS proposed in the study performs exactly as MCCS and ESPS, with the difference that, at the end, the top solutions are kept and not only the best one. Experiments performed on proteomics, genomics and text mining datasets show superior performance of MCCS as compared to ESPS, which show especially poor performance.

The final study to be reviewed is Han's doctoral thesis [123], analysing the instability of FSS algorithms from a theoretical viewpoint using a bias-variance decomposition approach. Specifically, instability is associated to the variance term in the decomposition, which is tightly coupled to sample size. Therefore, effort must be put on finding ways to decrease variance, a goal that can be obtained by variance reduction techniques such as importance sampling [110]. According to this technique, the variance of a Monte Carlo estimator can be reduced by increasing the number of instances taken from the regions which contribute more to the quantity of interest and decreasing the number of instances taken from other regions, instead of by i.i.d. sampling. Nevertheless, in practise (e.g., when using a limited biological dataset), it is not possible to perform this tailored sampling, although we can simulate its effect by weighting the instances accordingly. Based on these observations, the author proposes an empirical framework called Margin Based Instance Weighting (MBIW), which consists of three steps:

1. Transforming the original feature space into a Margin Vector Feature Space (MVFS) for an easy estimation of the importance of instances. For a dataset containing N instances, the MVFS is calculated following

the equation:

$$\mathbf{x}'_{i,j} = \sum_{l=1}^{M} |\mathbf{x}_{i,j} - m_l(\mathbf{x}_i)_j| - \sum_{l=1}^{H} |\mathbf{x}_{i,j} - h_l(\mathbf{x}_i)_j|. \tag{5.1}$$

where M and H are the total number of misses and hits (such that $M + H + 1 = N$); and $m_l(\mathbf{x})$ and $h_l(\mathbf{x})$ are the l-th nearest miss and hit with respect to instance $\mathbf{x}$.

2. Weighting each training instance according to its importance in the MVFS:

$$\omega(\mathbf{x}) = \frac{1/\bar{d}(\mathbf{x}')}{\sum_{i=1}^{N} 1/\bar{d}(\mathbf{x}'_i)}, \tag{5.2}$$

where

$$\bar{d}(\mathbf{x}') = \frac{1}{N-1} \sum_{p=1, \mathbf{x}'_p \neq \mathbf{x}'}^{N-1} \|\mathbf{x}' - \mathbf{x}'_p\|. \tag{5.3}$$

3. Finally, performing the FSS as usual. The only requirement is that the algorithm must be able to take instance weights into account. In this study, specifically-modified versions of SVM-RFE and RelievedF were employed.

The proposed framework was evaluated on synthetic data and real DNA microarray datasets, showing its suitability in reducing variance, which translates into an improvement of stability of FSS algorithms, while maintaining prediction performance and keeping the computational cost low, when compared to ensemble-like strategies.

5.3 Recursive Logistic Instance Weighting

The last study presented in the previous section supplies an empirical framework that appears to be the perfect candidate to overcome the problem of selecting stable feature subsets from SV-^{1}H-MRS data that are relevant for the task of differentiating between *gbm* and *met* tumours. A close look to its functioning, though, warns us of existing shortcomings that must be previously amended.

A first concern appears when analysing the mapping of instances to the new MVFS space. According to Eq. 5.1, a new coordinate is calculated for each dimension of an instance, and then the evaluation of the instance is typically carried out in the new space (Eq. 5.2). However, this explicit mapping seems avoidable, given that all dimensions are considered at a time by the Euclidean distance in Eq. 5.3. Hence, the evaluation of typicality for each instance can be performed directly in the original space. We reckon that this observation, despite being troublesome in terms of computational cost, does not influence the performance of the framework.

Imposing a normalisation factor in Eq. 5.2, such that the sum of all weights adds to 1, has a more serious effect. Given this constraint, the weight associated to each instance does not depend on its individual contribution, but on the total number of instances in the set (i.e., N), meaning that each weight is downgraded by a factor of N. As described in the experimental section, since the FSS algorithms employed in the study rely on distances between instances, an undesirable effect due to improper weights is shown to influence the algorithms' performance.

The work that we present in this section attempts to solve these inconveniences by providing a new framework for stable FSS using instance weighting.

5.3.1 A new instance weighting method

The first phase in our framework consists in weighting every instance of the training set according to whether they lay far from opposite-class instances. The reasoning is as follows: in a binary discrimination problem using small sample size datasets, instances close to opposite-class instances and far from same-class ones generate high instability, since the FSS outcome will highly vary depending on whether they have been picked up for the training set, or not. Contrarily, instances surrounded by same-class instances and far from opposite-class ones contribute positively to the stability of FSS algorithms. Therefore, we would like to reward the latter and punish the former.

Given the heterogeneity of the data used in our domain specific problem, we make use of the *hypothesis-margin* (see Section 5.2.1) to evaluate the

position of each instance with respect to same and opposite-class instances. Formally, let $D = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ be a training data set of length N, each instance $\mathbf{x}_i \in \mathbb{R}^d$ with its corresponding class label t_i , the margin of a hypothesis $\mathbf{x} \in \mathbb{R}^d$ can be calculated as:

$$\theta(\mathbf{x}) = \frac{1}{2}\left(\|\mathbf{x} - m(\mathbf{x})\| - \|\mathbf{x} - h(\mathbf{x})\|\right), \tag{5.4}$$

$m(\mathbf{x})$ and $h(\mathbf{x})$ being the nearest *miss* (instance of different class) and nearest hit (instance of same class) in D, respectively.

Notice that only accounting for the single closest neighbour of each type might be misleading if any of them present an atypical behaviour. Hence, a more robust evaluation can be calculated by averaging over all neighbours in D:

$$\theta(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{x} - m_i(\mathbf{x})\| - \frac{1}{H}\sum_{i=1}^{H}\|\mathbf{x} - h_i(\mathbf{x})\|, \tag{5.5}$$

where M, H are the total number of misses and hits. The sign of $\theta(\mathbf{x})$ is positive for those instances that are, on average, closer to same-class instances, while a negative sign is obtained whenever they are mostly surrounded by opposite-class instances; its value representing the strength in which the corresponding condition occurs.

The following step consists in bounding $\theta(\mathbf{x})$ in order to decouple its value, relative to the magnitude of the handled distances. For this purpose, we decided to limit the weight to be a positive value in the range $(0, 1)$ by using a logistic function:

$$\omega(\mathbf{x}) = \frac{1}{1 + \exp\left\{-\alpha\, z\left(\theta\left(\mathbf{x}\right)\right)\right\}}, \tag{5.6}$$

α being a hyperparameter controlling the slope, and $z(\cdot)$ the *standard score* $z(x) = (x - \hat{\mu}_D)/\hat{\sigma}_D$, where $\hat{\mu}_D$ and $\hat{\sigma}_D$ are the sample mean and standard deviation of $\theta(\mathbf{x})$, for all $\mathbf{x} \in D$, respectively. Suitable values for α are problem-dependent and must be set according to the user's needs. As a default value, we propose to set $\alpha = 3.03$, which corresponds to assigning a weight of 0.95 to an instance whose average margin is two standard deviations from the mean, that is $\theta(\mathbf{x}) = 2\hat{\sigma}_D$.

Finally, we divide each value by the mean. The reason for such operation is that the contribution of each instance (measured as distances within the environment) in the weighted FSS algorithms is to be multiplied by its weight, and we want to assign innocuous weights to typical instances (i.e., $\omega(\mathbf{x}) \approx 1$); values < 1 for atypically bad instances (regarding their location respect to all other instances); and > 1 for atypically good ones. Figure 5.2 shows an example of the ratings assigned by the proposed algorithm.

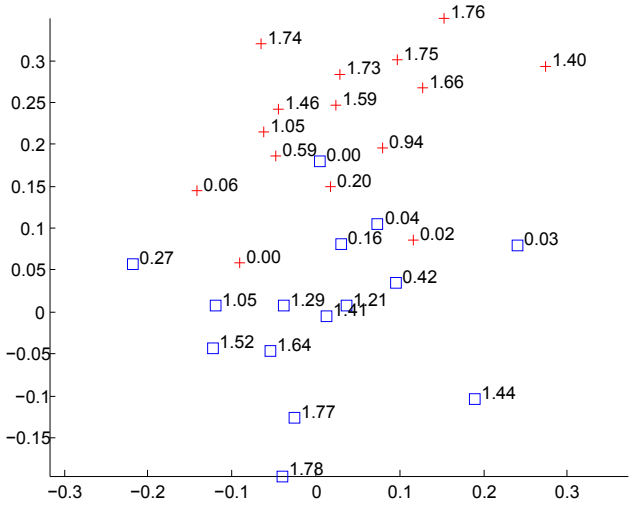


Figure 5.2: A weighting example - Ratings assigned by the proposed instance weighting approach to a synthetic dataset ($N = 30$). Data are generated by equally sampling from $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$, where $\mu_1 = [0, 0]$, $\mu_2 = [0, 0.25]$ and $\Sigma = \left[\begin{smallmatrix} 0.01 & 0.00 \\ 0.00 & 0.01 \end{smallmatrix}\right]$. Labels are set according to the distribution they come from. Notice the low values assigned to instances close to the boundary between classes and inside opposite-class region, while higher values are assigned to instances in the same-class region.

5.3.2 Weighted feature selection algorithms

The proposed weighted feature selection methods used in this study are a specifically modified version of the algorithms introduced in Section 5.2.2 to account for instance weights, as presented in [42]. The first one is a variant of the SVM-RFE:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \omega_i \xi_i,$$

where $\omega_i = \omega(\mathbf{x}_i)$ is the weight assigned to the i-th instance, according to Eq. 5.6.

The second weighted FSS alternative consists in introducing the ratings for the instance currently being treated, as well as for each miss and hit in the RelievedF formulation:

$$\mathrm{W}(j) = \sum_{i=1}^{N} \omega_i \sum_{l=1}^{k} \left(\omega_{i,l}^{M} |\mathbf{x}_{i,j} - m_l(\mathbf{x}_i)_j| - \omega_{i,l}^{H} |\mathbf{x}_{i,j} - h_l(\mathbf{x}_i)_j| \right), \tag{5.7}$$

where $\omega_i = \omega(\mathbf{x}_i)$, $\omega_{i,l}^{M} = \omega(m_l(\mathbf{x}_i))$ and $\omega_{i,l}^{H} = \omega(h_l(\mathbf{x}_i))$, obtained in Eq. 5.6.

Having presented all the required components, the *Recursive Logistic Instance Weighting* (RLIW) method is completed. It performs feature selection by repeatedly applying Eqs. (5.5) and (5.6) to compute the ω weights, uses them in a weighted FSS algorithm (e.g., either weighted SVM-RFE or weighted RelievedF), removing the worst feature (or features), recomputes the ω weights, and so on, until a stopping criterion is met. A Matlab toolbox containing the presented algorithm is available at `http://www.cs.upc.edu/~avilamala/resources/RLIW_Toolbox.zip`

5.4 Empirical evaluation

The ultimate goal of the RLIW method introduced in this thesis is to improve the stability of FSS algorithms in the discriminative task of diagnosing a tumour as *gbm* or *met* without losing predictive performance. This section shows two different groups of experiments from a technical viewpoint: firstly, limitations of MBIW are empirically verified using the same data as in its introductory study (i.e., synthetic and DNA microarray datasets); secondly, the suitability of our novel method is assessed using microarray DNA data and the SV-^{1}H-MRS dataset that is the main matter of the study of this thesis. A discussion on the benefits and risks of the new method is included, prior to revisit the initial hypotheses in the conclusions.

5.4.1 Experimental setup

Three different data sources were used to perform the experiments. One is a multivariate synthetic dataset [42] consisting of $M = 500$ training sets, each

of them of the form $\mathbf{X}^m \in \mathbb{R}^{N \times D}$, with $N = 100$ instances and $D = 1,000$ features, for $m = 1, \ldots, M$. Every instance is equiprobably drawn from one of two distributions: $\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma)$ or $\mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma)$, where

$$\mu_1 = (\underbrace{0.5, ..., 0.5}_{50}, \underbrace{0, ..., 0}_{950}), \quad \mu_2 = -\mu_1,$$

and

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{100} \end{bmatrix},$$

being $\Sigma_i \in \mathbb{R}^{10 \times 10}$, with 1 in its diagonal elements and 0.8 elsewhere. Class labels are assigned according to the expression:

$$\mathbf{y}_i = \text{sgn}\left(\sum_{j=1}^{D} \mathbf{X}_{i,j}\mathbf{r}_j\right), \quad \mathbf{r} = (\underbrace{0.02, ..., 0.02}_{50}, \underbrace{0, ..., 0}_{950}).$$

Notice that no test or hold-out sets are required to evaluate the stability of FSS.

The second type of data consists of seven different DNA microarray datasets, whose content, characteristics and pre-processing were discussed in Section 2.3.

Finally, the third source is a subset of SV-^{1}H-MRS data from the repository presented in Section 2.3. Specifically, 78 *gbm* and 31 *met* from the INTERPRET database were used as training set, whereas 30 *gbm* and 10 *met* from the eTumour database were used as hold-out set (a separate set is required because classification performance will also be assessed when using these data). Two different data modalities, one containing data acquired at LTE and the other containing data acquired at STE were employed. In these evaluations, the a priori and, according to medical expertise, most relevant 195 out of 512 frequencies were considered [32].

The experimental procedure is the same in all settings: given a normalised multivariate training set, importance of each instance is calculated (using either Eqs. 5.1, 5.2 and 5.3 in MBIW, or Eqs. 5.5, 5.6 and normalisation to the mean in RLIW) and instance ratings are provided to a weighted

FSS algorithm (either SVM-RFE or RelievedF-RFE), while removing the worst 10% of features per iteration until all of them have been eliminated. The procedure is repeated for each training set, calculating the KI at every feature subset size.

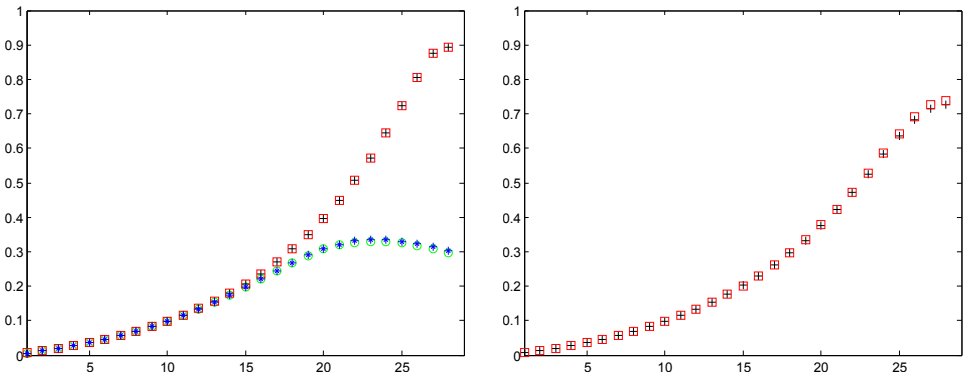
5.4.2 Limitations of Margin Based Instance Weighting

The undesirable effect of imposing a normalisation factor in Eq. 5.2 has been argued about in Section 5.3. We speculate that the improvement in FSS stability when using MBIW is not due to this preprocessing step, but to the influence that the normalisation factor has on the weighted FSS algorithms. Different configurations of parameters (available at Table 5.1) were designed to show such phenomenon.

Table 5.1: Configuration of different parameters in the Margin Based Instance Weighting experiments

FSS algorithm	Configuration	C	ω	Marker
SVM-RFE	default Std-FS	1	–	○
	rectified MBIW-FS	1	$N \times MBIW - FS$	$*$
	rectified Std-FS	N^{-1}	–	$+$
	default MBIW-FS	1	$MBIW - FS$	□
RelievedF-RFE	default Std-FS	–	–	$+$
	default MBIW-FS	–	$MBIW - FS$	□

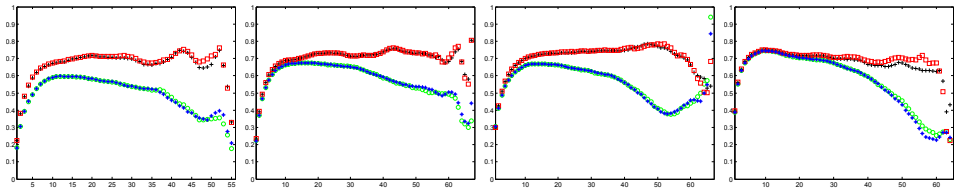
When the base FSS algorithm to use is SVM-RFE, an improvement in terms of feature subset stability on the synthetic dataset due to MBIW was reported in [42]. It corresponds to the configuration named *default MBIW-FS* ($C = 1$ using MBIW-FS to weight instances) and is compared to the poor performing *default Std-FS* ($C = 1$ using no instance weighting). As evidenced by Figure 5.3a, we obtain the same improvement using no instance weighting in the *rectified Std-FS* configuration, where C value has been divided by N (same effect as the normalisation factor induces). Bad results shown in a previous study in which no instance weighted was used has also been mimicked by the *rectified MBIW-FS*, where the rating of instances has been multiplied by N.



(a) SVM-RFE **(b)** RelievedF-RFE

Figure 5.3: Feature subset stability of MBIW on synthetic data - The plots show the KI (vertical axis) over a set of RFE iterations (horizontal axis). Parameters are set according to Table 5.1.

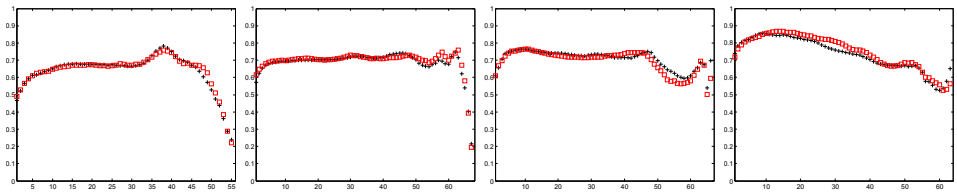
For the RelievedF-RFE (setting $K = 10$ as in previous study) as FSS base algorithm, neither the *default Std-FS* (no instance weighting) nor the *default MBIW-FS* (using MBIW-FS) configurations show any gain with respect to their counterpart. Notice that no scaling factor was applied in this setting because it does not affect the performance of RelievedF-RFE.



(a) Colon **(b)** Leukaemia **(c)** Prostate **(d)** Lung

Figure 5.4: Feature subset stability of MBIW using SVM-RFE on real microarray data - Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Parameters are set according to Table 5.1.

This effect has been verified in a larger cohort of data by performing a set of experiments over several DNA microarray datasets (the same ones as in [42]). Different training sets were obtained through a 10-times 10-fold cross-validation resampling strategy. KI was computed per feature subset length at every inner 10CV and then the average over the 10 times was calculated. Figure 5.4 and Figure 5.5 display the results obtained by SVM-RFE and



(a) Colon **(b)** Leukaemia **(c)** Prostate **(d)** Lung

Figure 5.5: Feature subset stability of MBIW using RelievedF-RFE on real microarray data - Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Parameters are set according to Table 5.1.

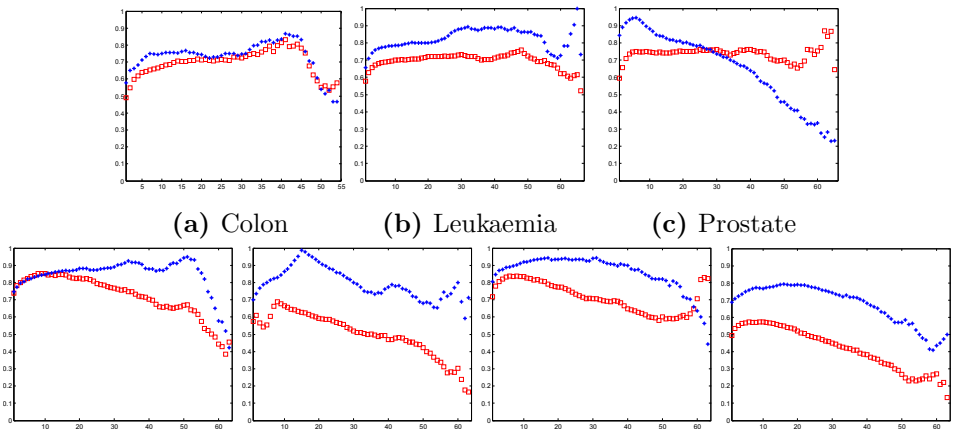
RelievedF-RFE, respectively, showing the same trend as in the experiments with synthetic data.

In light of these results, we reassert the initial hypothesis that MBIW by itself has little effect to the improvement on the stability of FSS algorithms.

5.4.3 Suitability of Recursive Logistic Instance Weighting

In this block, we shift our focus to the evaluation of the performance of the proposed RLIW method as compared to the use of standard FSS (Std-FS) algorithms without any instance weighting as preprocessing. For each experiment, the stability of the resulting feature subset as evaluated according to the KI, as well as the predictive performance of the subsequent classifier, measured by BAC, are provided. The FSS method of choice is RelievedF-RFE. The reason for not employing SVM-RFE is the high computational cost of adjusting the C parameter at each RFE iteration. Moreover, our preliminary results agree with the statement made in [75], stating that SVM-RFE is a highly stable algorithm, in contrast to Relief; therefore, unstable Relief is the family of filters that would most benefit from stability improvement strategies.

The first battery of experiments use all the DNA microarray datasets introduced in Section 2.3 with the purpose of selecting the subset of features that best discriminates among pathological and control subjects. Specifically, we employed a double 10-fold cross-validation resampling strategy to obtain the required number of independent sets allowing us to perform FSS, parameter adjustment and evaluate generalisation performance. Class



(a) Colon **(b)** Leukaemia **(c)** Prostate

(d) Lung **(e)** Breast **(f)** Melanoma **(g)** Parkinson

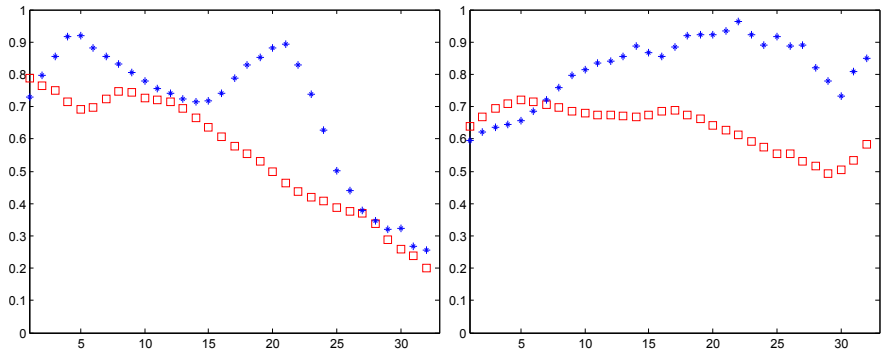
Figure 5.6: Feature subset stability of RLIW using RelievedF-RFE on the microarray data - Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Red squares: standard (unweighted) FSS; blue asterisks: RLIW.

predictions were obtained using linear-SVM, where the C parameter was adjusted according to the BAC measure in the inner 10CV in a logarithmic scale. The final subset of features corresponds to the one reaching maximum stability among those containing less than 20% of total features. The reported results are achieved in the outer 10cv for both feature subset stability (KI) and predictive performance (BAC). As seen in Figure 5.6, they show a clear gain in stability when RLIW is applied for most of RFE iterations in Colon, Leukaemia, Lung, Brest, Melanoma and Parkinson pathologies, an exception being the Prostate dataset, for which we have no clear explanation beyond the specificity of the dataset. Looking at the predictive capability of the selected subsets of features (Table 5.2), similar accuracies are shown for most of the datasets but Breast and Parkinson, for which a price of almost 10% less predictive capability is paid for the gains in stability.

The final experiment consists in assessing whether the proposed methodology is suitable for improving the stability of feature subset selection in the discrimination of *gbm* from *met* using SV-^{1}H-MRS data, which has been our ultimate goal from the beginning. The existence of a real test set permits to design the experiment using a 10 times 10-fold cross validation (10x10CV)

Table 5.2: Average balanced accuracies and their standard errors on the microarray datasets; feature subset size is shown in parentheses

Dataset	Std-FS	RLIW-FS
Colon	$0.82 \pm 0.05\,(22)$	$0.79 \pm 0.05\,(22)$
Leukaemia	$0.97 \pm 0.02\,(40)$	$0.98 \pm 0.02\,(3)$
Prostate	$0.94 \pm 0.02\,(5)$	$0.92 \pm 0.03\,(1239)$
Lung	$0.98 \pm 0.01\,(1026)$	$0.97 \pm 0.01\,(19)$
Breast	$0.76 \pm 0.05\,(1026)$	$0.66 \pm 0.05\,(1026)$
Melanoma	$0.98 \pm 0.02\,(3)$	$0.97 \pm 0.02\,(187)$
Parkinson	$0.78 \pm 0.04\,(1026)$	$0.68 \pm 0.05\,(923)$



(a) STE **(b)** LTE

Figure 5.7: Feature subset stability of RLIW using RelievedF-RFE on the real ^{1}H-MRS data - Each plot shows the KI over the successive RFE iterations. Red squares: standard (unweighted) FSS; blue asterisks: RLIW.

resampling technique for setting the C parameter in the linear-SVM learner and generating enough variability. The generalisation performance of the learner was assessed by calculating the BAC in the test set, while feature subset stability was evaluated as the average KI over the 10 times. According to the plots in Figure 5.7, RLIW achieves higher stability values, this being especially evident for the LTE dataset. Moreover, as shown in Table 5.3, the same or even better predictive performance is obtained in both validation and test sets when RLIW is applied with respect to Std-FS, using almost half of the features, fact that represents another advantage in this setting.

Table 5.3: Balanced accuracies and standard errors achieved by a linear SVM (number of selected features in parentheses) in discriminating between *gbm* and *met* using SV-^{1}H-MRS data.

			10x10CV	Test
STE	Std-FS	(36)	0.62 ± 0.01	0.67 ± 0.07
	RLIW-FS	(17)	0.65 ± 0.01	0.68 ± 0.07
LTE	Std-FS	(28)	0.62 ± 0.01	0.60 ± 0.08
	RLIW-FS	(15)	0.65 ± 0.01	0.60 ± 0.08

5.4.4 Discussion

Based on the obtained results in multiple datasets, despite some cases where the stability of feature subset selection algorithms can be increased without losing predictive performance, a sensible analysis would be to accept a trade-off between stability and accuracy, as a general trend. From this perspective, the decision on what measure to prioritise will be domain- and problem-specific. For instance, in an email spam filter, accuracy is the only measure that matters; nonetheless, in a knowledge discovery context, as for instance, identifying candidate genes as biomarkers to encode the existence of a specific pathology, focus on feature stability would be advisable. A third important actor in this decision might well be the final number of features to be used, that might relegate stability and accuracy to a subsidiary role. An example of this last case involves building a 3D viewer within a Decision Support System, where the number of final dimensions (features) must be exactly three.

Another observation we want to make is that, given the large dimensionality and small sample sizes of certain datasets, it might well be that previous results, obtained with little concern for stability are subject to large variability and over-optimistic in their evaluation of performance.

5.5 Conclusions

Recursive Logistic Instance Weighting is a novel technique that works as a data preprocessing step and whose task is to rate instances in a small dataset

(such as in the discriminative problem of glioblastomas *vs.* metastases from SV-^{1}H-MRS) according to their typicality in order to create an importance sampling effect in those situations where no sampling is available, to be supplied to FSS algorithms capable to deal with instance weights to obtain more stable subsets of features over different executions. The obtained solution is based on the assumptions introduced in Section 5.1:

1. *There are some instances that are typical regarding their underlying distribution, while others present outlying behaviour. Due to the fact that the latter type induces FSS algorithms to be unstable, we could simply remove them for the sake of stability if the dataset is large enough. In our case, the sample size being small, we can obtain a similar effect by weighting their importance in the FSS process*: a multivariate weighting technique based on distances to same- and opposite-class instances has been designed to evaluate typicality.

2. *The high heterogeneity of instances leads them to cluster in local neighbourhoods. This means that FSS algorithms approaching the hypothesis-margin are more likely to be suitable than the ones aiming at reducing the sample-margin*: we have adapted the *hypothesis-margin* based ReliefF FSS algorithm to deal with instance weights.

Results of the experiments on INTERPRET and eTumour datasets corroborate the suitability of the proposed technique as a candidate for efficiently improving stability in feature subset selection algorithms in the current domain, where high-dimensional sparse datasets are commonplace.

Chapter 6

Non-negative Matrix Factorisation

SV-^{1}H-MRS data have so far been used in this thesis to build models that accurately discriminate *gbm* from *met*, or to provide relevant biomarkers for better understanding which metabolites are directly involved in such differentiation. All these improvements have been achieved under the assumption that the measured biochemical components that are present in a voxel are of a single type (e.g., a specific tumour pathology or normal tissue). For a variety of reasons, including interferences from neighbouring voxels and coexistence of different tissues in the relatively large space conforming a voxel, measurements read by NMR scanners in real practise consist of a mixture of signals from different sources.

From this realisation, we now turn our attention towards strategies able to identify the signal generating sources and their relative contribution to the signal measured in a specific voxel. Previous studies analysing this phenomenon on similar data [124, 125, 126, 25] have shown success by employing BSS techniques, such as PCA or ICA; recently, a comparatively novel technique of this family, Non-negative Matrix Factorisation (NMF), which is the subject of this chapter, has shown encouraging results.

The ensuing sections in this chapter are structured as follows: a thorough explanation motivating the need for a new supervised algorithm for source extraction with a variety of restrictions imposed by the application domain

under investigation is exposed in the next section. We then review several studies that partially fulfil some of the requirements before introducing our proposed solution, deriving iterative algorithms for both training and prediction. Next, the performance of the new algorithm is assessed when addressing typical oncological questions using various data sets and, finally, our initial hypotheses are validated in the conclusions section.

6.1 Motivation

The assumption that the ^{1}H-MRS signal captured in each voxel can be exclusively attributed to a unique phenomenon occurring in the tissue of that specific voxel is a very strong one that does not often match real radiological practise. A frequently encountered pattern results in the measure being a mixture of various signals emitted from different components. The causes for this to happen include the existence of several different biological tissues within the voxel volume and the influence of interfering neighbouring voxels.

Another important issue, directly coupled with the previous statement, that needs to be questioned is the actual meaning of individual MRS frequencies in the spectra: under the conjecture that the measured signal is a composite, single point frequencies do not have entity by themselves (e.g., a biomarker corresponding to a specific tumour type), but they are instead a composite measure made up of the contributions from different sources. According to this new paradigm, it seems plausible to aim at assessing the contribution of each source to every frequency, as an alternative to singling out isolated frequencies to be labelled as biomarkers.

Importantly, when manipulating data to perform an analysis, a couple of restrictions must be kept in mind: it is common to use the ratio between certain metabolites (e.g., N-Acetyl Aspartate/Creatine, N-Acetyl Aspartate/Choline or Choline/Creatine) to analyse spectra in order to come out with a diagnosis; also, some metabolites at specific spectral frequencies may contain negative values. Therefore, any attempt to shape the data to fit the restrictions of our algorithms that overlooked these two issues might lead to biased results.

Having described the properties of SV-^{1}H-MRS data in this new context, we aim at designing a new solution that fulfils the following requirements:

1. It must be able to identify the underlying sources present in the retrieved signal.
2. It needs to assess the contribution of each source to the signal.
3. Both the sources and their contributions must be easily interpretable.
4. The solution must naturally deal with both negative and positive values.
5. Ratios between values of metabolites at certain frequencies must be preserved.
6. Distances between values of metabolites at specific frequencies must be kept.
7. Supervised information must be easily included in the solution when labelled data are available.

6.2 State of the Art

NMF is a low rank approximation technique that aims at factorising a given matrix of non-negative instances $\mathbf{X} \in \mathbb{R}_+^{D\times N}$ into a matrix of sources $\mathbf{S} \in \mathbb{R}_+^{D\times K}$, and a mixing matrix $\mathbf{H} \in \mathbb{R}_+^{K\times N}$; N being the number of instances, D the dimensionality of data, and K the number of sources. That is,

$$\mathbf{X}_+ = \mathbf{S}_+\mathbf{H}_+ + \mathbf{E} \approx \mathbf{S}_+\mathbf{H}_+$$

where $\mathbf{E}$ is some reconstruction error. A characteristic feature of this decomposition compared to other well-known BSS strategies is the constraint that all values in the matrices involved must be non-negative, a restriction that, for practical purposes, translates into facilitating the interpretability of the decomposition, given that any instance in $\mathbf{X}$ is approximated by a positive combination of the sources in $\mathbf{S}$, the contribution of which is encoded in $\mathbf{H}$.

Most of the algorithms obtain the decomposition by minimising a cost function that calculates the difference between the original data and the reconstructed signal by an iterative procedure. The cost function is denoted as

$$\Omega(\mathbf{X}||\mathbf{SH}).$$

6.2.1 Non-negative Matrix Factorisation variants

The first study to propose a solution for NMF was reported in [69], although they termed it Positive Matrix Factorisation. The cost function was denoted as follows:

$$\Omega_F(\mathbf{X}||\mathbf{SH}) = \min\|\mathbf{X} - \mathbf{SH}\|_F^2, \tag{6.1}$$

where $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$. The proposed Alternating Least Squares procedure consists in randomly initialising $\mathbf{S}$ and $\mathbf{H}$ and iteratively updating each matrix in turn according to:

$$\mathbf{H} \leftarrow \left(\mathbf{S}^\top\mathbf{S}\right)^{-1}\mathbf{S}^\top\mathbf{X}, \qquad \mathbf{S} \leftarrow \mathbf{X}\mathbf{H}^\top\left(\mathbf{H}\mathbf{H}^\top\right)^{-1},$$

setting all negative values to 0, until convergence is reached.

It was not until a publication in *Nature* by Lee and Seung [127], under the name of Non-negative Matrix Factorisation and mostly oriented towards image analysis, that the decomposition began to attract mainstream attention. In subsequent work [128], these authors proposed an information theoretic formulation based on the Kullback-Leibler divergence:

$$\Omega_{KL}\left(\mathbf{X}||\mathbf{SH}\right) = \min\sum_{d=1}^{D}\sum_{n=1}^{N}\left(\mathbf{X}_{d,n}\ln\left(\frac{\mathbf{X}_{d,n}}{[\mathbf{SH}]_{d,n}}\right) + [\mathbf{SH}]_{d,n} - \mathbf{X}_{d,n}\right). \tag{6.2}$$

A *multiplicative update rule* within a Gradient Descent strategy, preserving the non-negativity constraint, was provided to optimise $\Omega_{KL}\left(\mathbf{X}||\mathbf{SH}\right)$:

$$\mathbf{H}_{k,n} \leftarrow \mathbf{H}_{k,n}\frac{\left(\mathbf{S}^\top\mathbf{X}\right)_{k,n}/\left(\mathbf{J}_{KD}\mathbf{SH}\right)_{k,n}}{\left(\mathbf{S}^\top\mathbf{J}_{KN}\right)_{k,n}}, \tag{6.3}$$

$$\mathbf{S}_{d,k} \leftarrow \mathbf{S}_{d,k}\frac{\left(\mathbf{X}\mathbf{H}^\top\right)_{d,k}/\left(\mathbf{SHJ}_{NK}\right)_{d,k}}{\left(\mathbf{J}_{DN}\mathbf{H}^\top\right)_{d,k}}, \tag{6.4}$$

where $\mathbf{J}_{IL}$ is a $I \times L$ unit matrix. In the same study, similar update rules for $\Omega_F(\mathbf{X}||\mathbf{SH})$ cost function were also derived:

$$\mathbf{H}_{k,n} \leftarrow \mathbf{H}_{k,n}\frac{(\mathbf{S}^\top\mathbf{X})_{k,n}}{(\mathbf{S}^\top\mathbf{SH})_{k,n}}, \qquad \mathbf{S}_{d,k} \leftarrow \mathbf{S}_{d,k}\frac{(\mathbf{XH}^\top)_{d,k}}{(\mathbf{SHH}^\top)_{d,k}}.$$

Although Frobenius and Kullback-Leibler cost functions using multiplicative update rules are the most widely used strategies to solve the NMF decomposition, attempts to formalise the problem using other cost functions also exist. In this respect, we want to mention the Csiszár [129] and the Amari alpha divergences [130], which have recently been used for NMF purposes.

More efficient update rules have also been proposed in the literature, as the Alternating Least Squares using Projected Gradient bound-constrained optimisation method [131] for $\Omega_F(\mathbf{X}||\mathbf{SH})$:

$$\mathbf{H} \leftarrow \mathcal{P}\left[\mathbf{H} - \alpha\mathbf{S}^\top(\mathbf{SH} - \mathbf{X})\right], \quad \mathbf{S} \leftarrow \mathcal{P}\left[\mathbf{S} - \alpha(\mathbf{SH} - \mathbf{X})\mathbf{H}^\top\right],$$

where $\mathcal{P}[\cdot] = \max[\cdot, 0]$ is a bounding function ensuring the solution remains feasible; or a Second-Order Quasi-Newton optimisation for Amari alpha divergence [130].

A particularly interesting family of NMF variants is that in which the non-negativity constraint is relaxed, allowing values of any sign in both the original matrix $\mathbf{X}$ and the obtained sources $\mathbf{S}$, extending the applicability of NMF techniques to a broader range of applications. Semi Non-negative Matrix Factorisation (SNMF) [132] is a technique that specifically deals with this setting. In symbols,

$$\mathbf{X}_\pm \approx \mathbf{S}_\pm\mathbf{H}_+.$$

This study also derived a restricted version of SNMF, namely Convex NMF (CNMF), a formalism that forces the matrix of sources to be a convex combination of original instances (i.e., $\mathbf{S} = \mathbf{XW}$), gaining in interpretability, since the obtained sources can be read as class centroids:

$$\mathbf{X}_\pm \approx \mathbf{X}_\pm\mathbf{W}_+\mathbf{H}_+.$$

Now the Frobenius cost function is expressed as

$$\Omega_F\left(\mathbf{X}||\mathbf{XWH}\right) = \min\|\mathbf{X}-\mathbf{XWH}\|_F^2. \tag{6.5}$$

The corresponding update rules maintaining non-negativity constraints become

$$\begin{aligned}
\mathbf{H}_{k,n} &\leftarrow \mathbf{H}_{k,n}\sqrt{\frac{[\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{WH}+\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+]_{k,n}}{[\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{WH}+\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-]_{k,n}}},\\
\mathbf{W}_{n,k} &\leftarrow \mathbf{W}_{n,k}\sqrt{\frac{\left[(\mathbf{X}^\top\mathbf{X})^-\mathbf{WHH}^\top+(\mathbf{X}^\top\mathbf{X})^+\mathbf{H}^\top\right]_{n,k}}{\left[(\mathbf{X}^\top\mathbf{X})^+\mathbf{WHH}^\top+(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right]_{n,k}}},
\end{aligned}$$

where $(\mathbf{A})^+ = (|\mathbf{A}|+\mathbf{A})/2$ and $(\mathbf{A})^- = (|\mathbf{A}|-\mathbf{A})/2$.

6.2.2 Supervised Non-negative Matrix Factorisation

There are domains where the problems to be solved are clearly classification-oriented, meaning that desirable NMF role is not only to provide consistent interpretable bases, but also to supply class-separable subspaces. In those circumstances in which labelled instances are available, research has focused on enhancing NMF solutions by incorporating discriminant factors to the cost function. The first studies dealing with supervised NMF [133, 134] proposed to include Fisher's Linear Discriminants (LDA) to the $\Omega_{KL}\left(\mathbf{X}||\mathbf{SH}\right)$ cost function. To understand their functioning, let us first introduce the notion of *scatter matrices*. On the one hand, the *within-class scatter matrix* ($\mathbf{S}_w$) is a figure that evaluates the class-specific dispersion of instances; on the other hand, the *between-class scatter matrix* ($\mathbf{S}_b$) computes the intra-class variability. They can be calculated as follows:

$$\begin{aligned}
\mathbf{S}_w &= \sum_{r=1}^{R}\sum_{i\epsilon C_r}(\mathbf{u}_i-\boldsymbol{\mu}_r)(\mathbf{u}_i-\boldsymbol{\mu}_r)^\top,\\
\mathbf{S}_b &= \sum_{r=1}^{R}N_r(\boldsymbol{\mu}_r-\boldsymbol{\mu})(\boldsymbol{\mu}_r-\boldsymbol{\mu})^\top,
\end{aligned}$$

where,

$$\boldsymbol{\mu}_r = \frac{1}{N_r}\sum_{i\epsilon C_r}\mathbf{u}_i, \quad \boldsymbol{\mu} = \frac{1}{N}\sum_{r=1}^{R}N_r\boldsymbol{\mu}_r,$$

$\mathbf{u}_i$ being each D-dimensional instance, R the total number of classes, C_r the set of instances belonging to r-th class and N_r the cardinality of this set. Equivalently, if we arrange all instances as columns in matrix $\mathbf{U}$, the same computations can be expressed in matrix notation:

$$\begin{aligned} \mathbf{S}_w &= \mathbf{U}\mathbf{U}^\top - \mathbf{U}\tilde{\mathbf{M}}\mathbf{U}^\top, \\ \mathbf{S}_b &= \mathbf{U}\tilde{\mathbf{M}}\mathbf{U}^\top - \frac{1}{N}\mathbf{U}\mathbf{J}_N\mathbf{U}^\top, \end{aligned}$$

where $\mathbf{J}_N$ is a $N \times N$ square unit matrix and $\tilde{\mathbf{M}} = \mathbf{M}\mathbf{M}^\dagger(\mathbf{M}^\top)^\dagger\mathbf{M}^\top$; $\mathbf{A}^\dagger = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ being the left pseudo-inverse of $\mathbf{A}$ and $\mathbf{M} \in \{0,1\}^{N\times R}$ a matrix containing 1 in position $\mathbf{M}_{n,r}$ if instance n belongs to class r, and 0 otherwise.

The purpose of LDA is to find a mapping to a subspace such that $\mathbf{S}_w$ is minimised and $\mathbf{S}_b$ is maximised. The same rationale is applied when incorporating the scatter matrices into the cost function:

$$\Omega_{fisher}\left(\mathbf{X}||\mathbf{SH}\right) = \min\left[\Omega_{KL}\left(\mathbf{X}||\mathbf{SH}\right) + \gamma Tr[\mathbf{S}_w] - \lambda Tr[\mathbf{S}_b]\right], \tag{6.6}$$

$Tr[\mathbf{A}]$ being the trace of matrix $\mathbf{A}$, γ and λ two user-defined parameters that regulate the trade-off between prioritising a solution encompassing low reconstruction error or high separability. Scatter matrices are calculated on the low-rank projection of instances represented by the mixing matrix: $\mathbf{U} \leftarrow \mathbf{H}$. The two previously mentioned studies differ from each other in the way $\mathbf{S}_b$ is calculated (i.e., distance between each pairwise of class-centroids, or between each class-centroid and global mean).

The same idea of including Fisher discriminants to NMF cost function was used in [135], where the $\Omega_F(\mathbf{X}||\mathbf{SH})$ cost function was enhanced and only the between-class variance was included as a discriminant, resolving some of the issues in previous discriminant NMF algorithms.

Similarly, in [136], the $\Omega_F(\mathbf{X}||\mathbf{SH})$ cost function was employed; an important difference with previous studies is that scatter matrices are no longer calculated on the mixing matrix, but in the actual projection; that is $\mathbf{U} \leftarrow \mathbf{S}^\dagger\mathbf{X}$; or $\mathbf{U} \leftarrow \mathbf{S}^\top\mathbf{X}$ for real non-negativity solutions. Update rules

within an efficient Projected Gradient method were also provided:

$$\begin{aligned}\mathbf{H} &\leftarrow \mathcal{P}\left[\mathbf{H} - \alpha\mathbf{S}^\top(\mathbf{SH} - \mathbf{X})\right], \\ \mathbf{S} &\leftarrow \mathcal{P}\left[\mathbf{S} - \alpha(\mathbf{SH} - \mathbf{X})\mathbf{H}^\top + \gamma Tr[\mathbf{S}_w] - \lambda Tr[\mathbf{S}_b]\right],\end{aligned}$$

where $\mathcal{P}[\cdot]$ is $\max[\cdot, 0]$.

A different approach towards supervised NMF strategies involves the incorporation of SVM-like maximum sample-margin classification constraints [137] into the Kullback-Leibler-based cost function:

$$\Omega_{svm}(\mathbf{X}||\mathbf{SH}) = \lambda\Omega_{KL}(\mathbf{X}||\mathbf{SH}) + \frac{1}{2}Tr\left[\mathbf{A}\left(\mathbf{y}^\top\mathbf{y}\right)\mathbf{A}\left(\mathbf{H}^\top\mathbf{H}\right) - \mathbf{A}\mathbf{J}_{N1}\right],$$

where $\mathbf{A}$ is a diagonal matrix of Lagrange multipliers, and $\mathbf{y} \in \{-1, 1\}^N$ a row vector of class labels. Multiplicative update rules are supplied in the report.

We finish this section by reviewing an attempt to semi-supervised NMF for those situations where neither instances nor labels are ensured to be completed and missing values exist [138]. Let $\mathbf{Q} \in \{0, 1\}^{D\times N}$ encoding whether value d in instance n is observed or not; $\mathbf{R} \in \{0, 1\}^{R\times N}$ where the whole n-column is set to 1 whenever label for instance $\mathbf{x}_n$ is known, 0 otherwise; and $\mathbf{V} \in \mathbb{R}^{R\times K}$ the basis matrix for $\mathbf{M}^\top$; the cost function to optimise is:

$$\Omega_{semi}(\mathbf{X}||\mathbf{SH}) = \min\left[\|\mathbf{Q}\odot(\mathbf{X} - \mathbf{SH})\|_F^2 + \lambda\|\mathbf{R}\odot\left(\mathbf{M}^\top - \mathbf{VH}\right)\|_F^2\right],$$

where $\odot$ is the Hadamard product (i.e., element-wise multiplication).

An iterative procedure with multiplicative update rules for $\mathbf{S}$, $\mathbf{V}$ and $\mathbf{H}$ are provided in [138].

6.2.3 Non-negative Matrix Factorisation for Magnetic Resonance Spectroscopy in neuro-oncology

Since early studies such as [139] reported, where the Bayesian Spectral Decomposition (BSD) method was derived to decompose multivoxel Chemical Shift Images (CSI-MRS) of the human brain into a non-negative matrix of basic sources (representing muscle and brain tissue) and their corresponding non-negative matrix of tissue contribution to each voxel, several research

groups have contributed to enhance the literature of NMF variants applied to the study of brain structure and, ultimately, the diagnosis of brain tumours.

A much faster algorithm was presented in [140], namely Constrained Non-negative Matrix Factorisation (cNMF), where traditional NMF [127] was improved by incorporating a regulariser enforcing sparsity to the cost function. Its suitability was evaluated using the same dataset as in the previous study.

An evolution of the latter was materialised in [141] by stacking individual cNMF modules to obtain a hierarchical architecture, which was proved to achieve more meaningful physical sources for the same dataset. The paper also stresses the potential of these techniques to aid in the diagnosis of brain tumours.

Monitorisation of the response to chemotherapy in a patient suffering from oligodendroglioma was conducted in [142], by employing cNMF to process multivoxel CSI-MRS data of the brain.

In 2008, a study [143] was carried out in a group of 20 patients affected by gliomas of different degree (half of the patients presented low-grade and half high-grade gliomas). The goal was to extract relevant tissue types and contribution of each tissue to every voxel in the MV-^{1}H-MRS image by means of traditional NMF.

A similar decomposition was attempted in [144] on a dataset of High-Resolution Magic Angle Spinning MRS signals from several glioblastoma tumour patients. The algorithm of interest in this study was Sparse Non-negative Matrix Factorisation via Alternating Non-negativity-constrained Least Squares [145].

More recently, the performance of a range of NMF variants was characterised in the task of extracting meaningful sources from SV-^{1}H-MRS data of brain tumour and control patients [25]. Moreover, classification accuracy of the methods was evaluated by direct comparison of the mixing matrices, as well as using the aforementioned methods as a dimensionality reduction step, previous to regular supervised classification. Results reported superior performance in both tasks for CNMF.

Following this same idea, a semi-supervised technique was designed in [146], where labelled information representing tumour type was used to improve the quality of the retrieved sources. More precisely, a metric able to scale each dimension of the feature space according to the degree of relevance regarding class membership (Fisher Information metric [147]) was developed to subsequently project the unseen data to this new space, where regular CNMF was applied. Superior classification accuracy was achieved and higher quality interpretable sources were obtained for the same dataset as in the previous study.

That same year a hierarchical NMF implementation was defined [148]. The strategy parsimoniously retrieved two sources at each level in order to obtain compounding tissues in a dataset made of STE MRSI data of glioblastoma multiforme. Proper discrimination of the three most relevant tissue types (i.e., normal, tumour and necrosis) was obtained, a goal that one-level NMF variants failed to solve.

One mandatory hyperparameter shared by any of the aforementioned methods is that to select the appropriate number of sources to be extracted. A recent study [149] aims at automatically determining such value by using a Variational Bayes NMF method that uses priors enforcing sparsity. The iterative process achieves its goal by discarding sources whose contribution is negligible (i.e., either values in the mixing matrix are near zero or they are highly correlated to other existing sources). Suitability of this new approach has been gage on SV-^{1}H-MRS data of patients affected by different brain tumour pathologies.

For a thorough review of the most significant applications of NMF to MRS data in the field of tissue typing methods for tumour diagnosis, please refer to the recently published [150]; or to [151] for applicability on the more general field of computational biology.

6.3 Discriminant Convex Non-negative Matrix Factorisation

Previous studies investigating SV-^{1}H-MRS data for brain tumour diagnosis demonstrated the appropriateness of the LDA learning algorithms for discriminating among tumour types [82, 92, 88]; another recent publication [25] showing the suitability of CNMF as a technique to identify different types of biological tissues in a voxel and their contribution to the retrieved signal has been discussed in the previous section. Therefore, we considered that developing a supervised version for CNMF by incorporating Fisher linear discriminants to the cost function was a natural step forward. This development is described next.

6.3.1 Objective function

Out of the two most used cost functions for NMF (i.e., Frobenius norm and Kullback-Leibler divergence), the Frobenius norm was chosen to be the base cost function to assess the error between real instances and their reconstructed versions. The reason for this choice is the fact that negative values exist in the matrices being employed and the Kullback-Leibler function, designed to measure divergence among probabilities, does not handle negativity. Hence, Eq. 6.5 is reformulated as:

$$\Omega_F\left(\mathbf{X}||\mathbf{XWH}\right) = Tr\left(\mathbf{XX}^\top + \mathbf{XWHH}^\top\mathbf{W}^\top\mathbf{X}^\top - 2\mathbf{XWHX}^\top\right).$$

In contrast to [134] and [136], scatter matrices are calculated in the reconstruction space, since we want to obtain simplified versions of the original instances containing higher discrimination capability despite losing similarity with their counterparts. Moreover, we want to compute this discrimination using the same unities (same order of magnitude) in the original space. For this, we set:

$$\tilde{\mathbf{S}}_w = \mathbf{XWHH}^\top\mathbf{W}^\top\mathbf{X}^\top - \mathbf{XWH}\tilde{\mathbf{M}}\mathbf{H}^\top\mathbf{W}^\top\mathbf{X}^\top$$

and

$$\tilde{\mathbf{S}}_b = \mathbf{XWH}\tilde{\mathbf{M}}\mathbf{H}^\top\mathbf{W}^\top\mathbf{X}^\top - \frac{1}{N}\mathbf{XWHJ}_N\mathbf{H}^\top\mathbf{W}^\top\mathbf{X}^\top.$$

Therefore, the complete objective function that the method optimises is:

$$\Omega_{DC}\left(\mathbf{X}||\mathbf{XWH}\right) = (1-\alpha)\,\Omega_F\left(\mathbf{X}||\mathbf{XWH}\right) + \alpha\left(Tr(\tilde{\mathbf{S}}_w) - Tr(\tilde{\mathbf{S}}_b)\right), \quad (6.7)$$

α being a user-defined parameter that controls the balance between approximating the reconstructed instances to the real ones or giving more power to the discriminative factors; its values ranging from 0 to 1. Let $\tilde{\mathbf{X}} = \mathbf{XWH}$; we can alternatively express the cost function as:

$$\begin{aligned}\Omega_{DC}\left(\mathbf{X}||\mathbf{XWH}\right) &= Tr\big[(1-\alpha)\mathbf{X}\mathbf{X}^\top + \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + (2\alpha-2)\tilde{\mathbf{X}}\mathbf{X}^\top \quad (6.8)\\ &\quad -2\alpha\tilde{\mathbf{X}}\tilde{\mathbf{M}}\tilde{\mathbf{X}}^\top + \frac{\alpha}{N}\tilde{\mathbf{X}}\mathbf{J}_N\tilde{\mathbf{X}}^\top\big].\end{aligned}$$

6.3.2 Optimisation procedure

We provide an iterative algorithm to optimise Eq. 6.8 based on multiplicative update rules that alternatively update matrices $\mathbf{W}$ and $\mathbf{H}$ until convergence. The procedure is summarised in Algorithm 1.

Algorithm 1 DCNMF algorithm

1) Normalise data $\mathbf{X}$ (L2-norm)
2) Initialise matrices $\mathbf{W}$ and $\mathbf{H}$ using K-means
3) Repeat until convergence:
   a) Update $\mathbf{H}$ according to Eq. 6.9
   b) Update $\mathbf{W}$ according to Eq. 6.10
4) Calculate $\mathbf{S} = \mathbf{XW}$
5) Normalise $\mathbf{S}$ (L2-norm) and $\mathbf{H}$ (L1-norm)

The proposed multiplicative expressions to update values of mixing and unmixing matrices have the property that no negative value will occur. For $\mathbf{H}$ matrix, the update is:

$$\mathbf{H}_{k,n} = \mathbf{H}_{k,n}\sqrt{\frac{\left(\breve{\mathbf{B}}_H\right)_{k,n}}{\left(\breve{\mathbf{V}}_H\right)_{k,n}}}, \text{ where} \quad (6.9)$$

$$\breve{\mathbf{B}}_H = \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{WH}\left(1+\frac{\alpha\mathbf{J}_N}{N}\right) + \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\left((1-\alpha)+2\alpha\mathbf{WH}\tilde{\mathbf{M}}\right),$$

$$\breve{\mathbf{V}}_H = \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{WH}\left(1+\frac{\alpha\mathbf{J}_N}{N}\right) + \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\left((1-\alpha)+2\alpha\mathbf{WH}\tilde{\mathbf{M}}\right).$$

And for $\mathbf{W}$ matrix, the rule is:

$$\begin{aligned}
\mathbf{W}_{n,k} &= \mathbf{W}_{n,k}\sqrt{\frac{\left(\breve{\mathbf{B}}_W\right)_{n,k}}{\left(\breve{\mathbf{V}}_W\right)_{n,k}}}, \text{ where} \qquad (6.10)\\
\breve{\mathbf{B}}_W &= (\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\left(1+\frac{\alpha\mathbf{J}_N}{N}\right)\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^+\left((1-\alpha)+2\alpha\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\right)\mathbf{H}^\top,\\
\breve{\mathbf{V}}_W &= (\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\left(1+\frac{\alpha\mathbf{J}_N}{N}\right)\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^-\left((1-\alpha)+2\alpha\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\right)\mathbf{H}^\top.
\end{aligned}$$

For a detailed derivation of Eq. 6.9 and Eq. 6.10 please refer to Appendix A. For an analysis of convergence, see Appendix B.

6.3.3 Prediction of unseen instances

The previous section introduced a procedure to obtain a set of sources $\mathbf{S}$ which correspond to the underlying processes generating the data, as well as a matrix of mixing coefficients $\mathbf{H}$ containing the contribution of each source to the retrieved signal in every voxel of measured data $\mathbf{X}$.

Now we are interested in mapping unseen instances into the new reconstructed space showing better discriminative capability. However, there is a key issue that complicates such mapping: the evidence we found in preliminary studies that the obtained sources do not capture much of the discriminative power imposed in the objective function, but the mixing matrix is the one absorbing most of the discrimination information. In other words, the effect of including class labels in the objective function has a strong influence on $\mathbf{H}$, such that the reconstructed instances $\tilde{\mathbf{X}}$ get more easily separable in the original space, but not much influence is applied over $\mathbf{S}$; hence no direct use of $\mathbf{S}$ can be applied to predict unseen instances. With the purpose of circumventing this limitation, we present two different approaches to predict new instances: the former repeatedly uses the *expectation-maximisation* framework to obtain predictive values for the mixture of sources of every unseen instance; the latter uses $\tilde{\mathbf{S}}$, a convex combination of reconstructed instances, instead of $\mathbf{S}$ as the matrix of generating sources.

6.3.3.1 Prediction using Expectation-Maximisation

Let $\mathbf{Y} \in \mathbb{R}^{D \times N_y}$ be the matrix of unlabelled instances whose low rank projection $\mathbf{Q}$ has to be predicted. That is, given a fixed source matrix $\mathbf{S}$, our goal is to find the best $\mathbf{Q}$ for $\mathbf{Y} \approx \mathbf{SQ}$, such that $\tilde{\mathbf{Y}} = \mathbf{SQ}$ presents high discriminatory capabilities. Given that $\mathbf{S}$ barely captures discrimination ability, we propose to minimise an objective function to obtain $\mathbf{Q}$ that jointly uses training instances $\mathbf{X}$ and labels $\mathbf{M}_x$ to guide the optimisation procedure. The cost function of choice corresponds to Eq. 6.6, changing Ω_{KL} by Ω_F, and using the appropriate matrices: for the Ω_F part of the equation, $\mathbf{S}$ is fixed from the training phase, and the observed and mixing matrices are:

$$\begin{aligned} \bar{\mathbf{X}} &= \mathbf{X} \cup \mathbf{y}, \\ \bar{\mathbf{H}} &= \mathbf{H} \cup \mathbf{q}, \end{aligned}$$

where $\mathbf{y} = \mathbf{Y}_{i,:}$ and $\mathbf{q} = \mathbf{Q}_{i,:}$ correspond to a single instance to be predicted. Here, $\mathbf{A} \cup \mathbf{B}$ operation means to append columns in $\mathbf{B}$ to the end of $\mathbf{A}$ matrix.

Despite using $\mathbf{H}$ to drive the optimisation process, we do not want its values to be influenced, but only those values in $\mathbf{q}$. That is why we split the matrix $\bar{\mathbf{H}}$ into

$$\bar{\mathbf{H}} = \mathbf{H} \cup \mathbf{q} = \bar{\mathbf{H}}\mathbf{U}^\top \cup \bar{\mathbf{H}}\mathbf{v}^\top = \mathbf{HU} + \mathbf{qv},$$

$\mathbf{U} \in \{0,1\}^{N \times (N+1)}$ and $\mathbf{v} \in \{0,1\}^{1 \times (N+1)}$ being an auxiliary matrix and a vector, respectively. $\mathbf{U}$ is a *mask matrix* used to extract the training part of the matrices. It contains 1's in its diagonal elements from position $(1,1)$ until position (n,n) and 0's elsewhere. Vector $\mathbf{v}$ is used to separate the unseen instances part of the matrices: it contains a 1 in its position $(1, n+1)$ and 0's elsewhere. Similarly, we split observed instances into training and predictive factors:

$$\bar{\mathbf{X}} = \mathbf{X} \cup \mathbf{y} = \bar{\mathbf{X}}\mathbf{U}^\top \cup \bar{\mathbf{X}}\mathbf{v}^\top = \mathbf{XU} + \mathbf{yv}.$$

Using this technique, we express the task of matrix factorisation as:

$$\bar{\mathbf{X}} \approx \mathbf{S}\bar{\mathbf{H}} = \mathbf{S}\left(\mathbf{HU} + \mathbf{qv}\right) = \left(\mathbf{SHU} + \mathbf{Sqv}\right),$$

where only an update rule for $\mathbf{q}$ is required.

The second part of the cost function deals with maximising separability among reconstructed instances according to the calculated scatter matrices. Let us suppose we have a matrix $\mathbf{m}_y \in [0,1]^{1\times R}$ containing the probability of the current instance $\mathbf{y}$ to belong to each class; then, the matrix of known class labels is

$$\mathbf{M} = \mathbf{M}_x \cup \mathbf{m}_y.$$

Following a similar approach as in the previous paragraph, we split the data as:

$$\mathbf{M} = \mathbf{M}_x \cup \mathbf{m}_y = \mathbf{M}\mathbf{U}^\top \cup \mathbf{M}\mathbf{v}^\top = \mathbf{M}_x\mathbf{U} + \mathbf{m}_y\mathbf{v}.$$

The scatter matrices using all training instances plus one instance from the prediction set become

$$\bar{\mathbf{S}}_b = \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{M}^\top\bar{\mathbf{X}}^\top - \frac{1}{N}\bar{\mathbf{X}}\mathbf{M}\mathbf{J}_K\mathbf{M}^\top\bar{\mathbf{X}}^\top,$$

where $\mathbf{E} = \left(\mathbf{P}_{K,1}\mathbf{P}_{K,1}^\top\mathbf{M}^\top\mathbf{J}_{N,1}\mathbf{P}_{K,1}^\top + \mathbf{R}_{K,1}\mathbf{R}_{K,1}^\top\mathbf{M}^\top\mathbf{J}_{N,1}\mathbf{R}_{K,1}^\top\right)$, $\mathbf{J}_K$ is a $K\times K$ unit matrix, and

$$\begin{aligned}\bar{\mathbf{S}}_w &= (\bar{\mathbf{X}}\mathbf{U}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{M}^\top\mathbf{U}^\top)(\bar{\mathbf{X}}\mathbf{U}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{M}^\top\mathbf{U}^\top)^\top \\ &+ (\bar{\mathbf{X}}\mathbf{V}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{P}_{K,1})\mathbf{V}\mathbf{M}\mathbf{P}_{K,1}(\bar{\mathbf{X}}\mathbf{V}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{P}_{K,1})^\top \\ &+ (\bar{\mathbf{X}}\mathbf{V}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{R}_{K,1})\mathbf{V}\mathbf{M}\mathbf{R}_{K,1}(\bar{\mathbf{X}}\mathbf{V}^\top - \bar{\mathbf{X}}\mathbf{M}\mathbf{E}^{-1}\mathbf{R}_{K,1})^\top.\end{aligned}$$

Now, recalling

$$\bar{\Omega}_F = \|\bar{\mathbf{X}} - \mathbf{S}\mathbf{H}\mathbf{U} - \mathbf{S}\mathbf{q}\mathbf{V}\|_F^2,$$

we already have all the components to define the cost function for prediction:

$$\bar{\Omega}_{DC} = (1-\alpha)\bar{\Omega}_F + \alpha\left(Tr(\bar{\mathbf{S}}_w) - Tr(\bar{\mathbf{S}}_b)\right),$$

otherwise expressed as:

$$\begin{aligned}\bar{\Omega}_{DC} &= (1-\alpha)\bar{\mathbf{X}}\bar{\mathbf{X}}^\top - (2-2\alpha)\bar{\mathbf{X}}\mathbf{U}^\top\mathbf{H}^\top\mathbf{S}^\top - (2-2\alpha)\bar{\mathbf{X}}\mathbf{V}^\top\mathbf{q}^\top\mathbf{S}^\top \\ &+ \mathbf{S}\mathbf{H}\bar{\mathbf{A}}\mathbf{H}^\top\mathbf{S}^\top + \mathbf{S}\mathbf{H}\bar{\mathbf{B}}\mathbf{q}^\top\mathbf{S}^\top + \mathbf{S}\mathbf{q}\bar{\mathbf{C}}\mathbf{q}^\top\mathbf{S}^\top \\ &- \mathbf{S}\mathbf{H}\bar{\mathbf{D}}\mathbf{H}^\top\mathbf{S}^\top - \mathbf{S}\mathbf{H}\bar{\mathbf{E}}\mathbf{q}^\top\mathbf{S}^\top - \mathbf{S}\mathbf{q}\bar{\mathbf{F}}\mathbf{q}^\top\mathbf{S}^\top,\end{aligned} \tag{6.11}$$

which uses the following constants:

$$
\begin{aligned}
\mathbf{P}_{K,K} &= \mathbf{P}_{K,1}\mathbf{P}_{K,1}^{\top}, \\
\mathbf{R}_{K,K} &= \mathbf{R}_{K,1}\mathbf{R}_{K,1}^{\top}, \\
\tilde{\mathbf{P}} &= \mathbf{P}_{K,1}\mathbf{V}\mathbf{M}\mathbf{P}_{K,1}, \\
\tilde{\mathbf{R}} &= \mathbf{R}_{K,1}\mathbf{V}\mathbf{M}\mathbf{R}_{K,1}, \\
\tilde{\mathbf{E}} &= \mathbf{E}^{-1}\mathbf{M}^{\top}\mathbf{U}^{\top}\mathbf{U}\mathbf{M}\mathbf{E}^{-1}, \\
\tilde{\mathbf{F}} &= \mathbf{E}^{-1}(\tilde{\mathbf{P}}\mathbf{P}_{K,1}^{\top} + \tilde{\mathbf{R}}\mathbf{R}_{K,1}^{\top})\mathbf{E}^{-1}, \\
\bar{\mathbf{A}} &= 1 + \alpha\mathbf{U}\mathbf{M}\left(\tilde{\mathbf{E}} + \tilde{\mathbf{F}} + \frac{\mathbf{J}_k}{N}\right)\mathbf{M}^{\top}\mathbf{U}^{\top}, \\
\bar{\mathbf{B}} &= \alpha\mathbf{U}\mathbf{M}\left(2\tilde{\mathbf{E}} + \tilde{\mathbf{F}} + \tilde{\mathbf{F}}^{\top} + \frac{2\mathbf{J}_k}{N}\right)\mathbf{M}^{\top}\mathbf{V}^{\top}, \\
\bar{\mathbf{C}} &= (1-\alpha) + \alpha\mathbf{V}\mathbf{M}\left[(\mathbf{P}_{K,1} + \mathbf{R}_{K,1}) + \left(\tilde{\mathbf{E}} + \tilde{\mathbf{F}} + \frac{\mathbf{J}_k}{N}\right)\mathbf{M}^{\top}\mathbf{V}^{\top}\right], \\
\bar{\mathbf{D}} &= 3\alpha\mathbf{U}\bar{\mathbf{M}}\mathbf{U}^{\top}, \\
\bar{\mathbf{E}} &= \alpha\mathbf{U}\mathbf{M}\mathbf{E}^{-1}\left(4 + \tilde{\mathbf{P}} + \tilde{\mathbf{R}} + \mathbf{R}_{K,K} + \mathbf{P}_{K,K}\right)\mathbf{M}^{\top}\mathbf{V}^{\top}, \\
\bar{\mathbf{F}} &= \alpha\mathbf{V}\mathbf{M}\left[(\mathbf{P}_{K,K} + \mathbf{R}_{K,K})\mathbf{E}^{-1}\mathbf{M}^{\top}\mathbf{V}^{\top} + \mathbf{E}^{-1}(\tilde{\mathbf{P}} + \tilde{\mathbf{R}} + \mathbf{M}^{\top}\mathbf{V}^{\top})\right].
\end{aligned}
$$

The update rule for vector $\mathbf{q}$ is obtained by applying the same procedure as for deriving the updating expression for $\mathbf{H}$ (for detailed information refer to Appendix A):

$$
\begin{aligned}
\mathbf{q}_k &= \mathbf{q}_k\sqrt{\frac{\left(\breve{\mathbf{B}}_q\right)_k}{\left(\breve{\mathbf{V}}_q\right)_k}}, \text{ where} \qquad (6.12) \\
\breve{\mathbf{B}}_q &= (2-2\alpha)(\mathbf{S}^{\top}\bar{\mathbf{X}})^{-}\mathbf{V}^{\top} + (\mathbf{S}^{\top}\mathbf{S})^{-}\left(\mathbf{H}\bar{\mathbf{B}} + 2\mathbf{q}\bar{\mathbf{C}}\right) + (\mathbf{S}^{\top}\mathbf{S})^{+}\left(\mathbf{H}\bar{\mathbf{E}} + 2\mathbf{q}\bar{\mathbf{F}}\right), \\
\breve{\mathbf{V}}_q &= (2-2\alpha)(\mathbf{S}^{\top}\bar{\mathbf{X}})^{+}\mathbf{V}^{\top} + (\mathbf{S}^{\top}\mathbf{S})^{+}\left(\mathbf{H}\bar{\mathbf{B}} + 2\mathbf{q}\bar{\mathbf{C}}\right) + (\mathbf{S}^{\top}\mathbf{S})^{-}\left(\mathbf{H}\bar{\mathbf{E}} + 2\mathbf{q}\bar{\mathbf{F}}\right).
\end{aligned}
$$

However, there is still one missing detail. Until now, we have assumed that class membership probabilities for the instance to be predicted $\mathbf{m}_y$ exist, but we do not have such information. This limitation is overcome by estimating class membership $\hat{\mathbf{m}}_y$ using the *expectation-maximisation* algorithm [152]:

First, the algorithm initialises $\mathbf{q}$ to be the mean vector of training mixing matrix $\mathbf{H}$, and calculates the class specific mean $\boldsymbol{\mu}_r$ and the covariance

$\mathbf{\Sigma}_r$ from $\mathbf{H}$. The *expectation* step consists in calculating the posterior probability of the current instance belonging to each class by approximating a multivariate Gaussian distribution:

$$p(C_r|\hat{\mathbf{q}}) = \frac{p(\hat{\mathbf{q}}|C_r)p(C_r)}{\sum_j p(\hat{\mathbf{q}}|C_j)p(C_j)},$$

such that $p(C_r) = \frac{N_r}{N}$ are the empirical priors (or known priors, were they available), and

$$p(\hat{\mathbf{q}}|C_r) = \frac{1}{(2\pi)^{R/2}|\mathbf{\Sigma}_r|^{1/2}} \exp\left\{-\frac{1}{2}(\hat{\mathbf{q}} - \boldsymbol{\mu}_r)^\top \mathbf{\Sigma}_r^{-1}(\hat{\mathbf{q}} - \boldsymbol{\mu}_r)\right\}.$$

Once $\hat{\mathbf{m}}_y = (p(C_1|\hat{\mathbf{q}}), ..., p(C_R|\hat{\mathbf{q}}))$ has been retrieved, the *maximisation* phase is conducted, consisting in iteratively updating $\mathbf{q}$ according to Eq. 6.12 until convergence; then, in the next *expectation* step, the $\hat{\mathbf{m}}_y$ vector is re-estimated using the new value for $\mathbf{q}$. This *expectation-maximisation* procedure is repeated until convergence, obtaining the final $\mathbf{q}$ vector. Given that only one instance is predicted at a time, the whole process needs to be done for each instance in $\mathbf{Y}$ to be predicted.

6.3.3.2 Prediction using Reconstructed Sources

Given the complexity of the prediction procedure using the expectation-maximisation framework, described in the previous section, we have derived an alternative approach. The idea is simple: instead of using $\mathbf{S}$ to predict new samples, since this matrix of sources does not capture all the discriminatory power included in $\mathbf{H}$, we can use the reconstructed instances to calculate a reconstructed version of the sources. That is:

$$\mathbf{Y} \approx \tilde{\mathbf{S}}\mathbf{Q} = \tilde{\mathbf{X}}\mathbf{W}\mathbf{Q} = \mathbf{X}\mathbf{W}\mathbf{H}\mathbf{W}\mathbf{Q}.$$

The prediction of the mixing matrix $\mathbf{Q}$ for the test instances $\mathbf{Y}$ can be achieved by applying the update rule for the mixing matrix, as in the CNMF algorithm:

$$\mathbf{Q}_{ik} = \mathbf{Q}_{ik}\sqrt{\frac{\left[(\tilde{\mathbf{S}}^\top\mathbf{Y})^+ + (\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}})^-\mathbf{Q}\right]_{ik}}{\left[(\tilde{\mathbf{S}}^\top\mathbf{Y})^- + (\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}})^+\mathbf{Q}\right]_{ik}}}.$$

Matlab code of the proposed algorithms is available at `http://www.cs.upc.edu/~avilamala/resources/DCNMF_Toolbox.zip`

6.4 Empirical evaluation

DCNMF has been designed to extract meaningful class-specific sources in complicated classification problems by including discriminative information from the training instances. This section presents the benchmark used to assess the appropriateness of the proposed technique employing two different sources of data: synthetically generated SV-^{1}H-MRS-like and real SV-^{1}H-MRS instances corresponding to the most prevalent questions in brain tumour diagnosis. Final remarks to better understand the obtained results and design decisions are also provided.

6.4.1 Experimental setup

The proposed method is evaluated on realistic regular practise problems, using data from two different sources: one consists of real SV-^{1}H-MRS data from INTERPRET repository (Section 2.3). Specifically, 22 astrocytomas grade II ($ac2$), 86 glioblastomas (gbm), 38 metastases (met) and 22 normal controls (nom), at STE using 195 out of 512 frequencies are included; and also 20 $ac2$, 78 gbm, 31 met and 15 nom for data at LTE. The second source of data contains synthetically generated SV-^{1}H-MRS-like samples, which have been built from fixed template sources (within-class tissue averages), mixed using an example mixing matrix. Then, for every diagnostic problem, the samples of each tumour type were averaged, becoming the artificial sources, resulting in as many sources as classes. Afterwards, Gaussian noise of different and increasing levels (5%, 15%, 25% and 35%) was added to the standardised synthetic data, ensuring that noise added to each dimension was proportional to its true standard deviation. The final data set includes the same number of instances as the real dataset just presented, plus 50 instances per class to be used as a test set.

The initialisation of the algorithm entails normalising the current training set to vector unit length and setting appropriate values to $\mathbf{H}$ and $\mathbf{W}$, this

last choice being of great relevance, since the proposed method converges to a local minimum. In a research study published in [153], an advantageous initialisation based on K-means clustering algorithm was proposed; it works by creating as many clusters as sources were desired to be extracted, defining a matrix $\mathbf{C} \in \{0,1\}^{K \times N}$ of cluster membership, and setting $\mathbf{H}^0 = \mathbf{C} + 0.2\mathbf{E}$, $\mathbf{W}^0 = (\mathbf{C} + 0.2\mathbf{E})^\top \mathbf{D}^{-1}$, $\mathbf{E}$ being a $K \times N$ unit matrix and $\mathbf{D}$ a $K \times K$ diagonal matrix with the number of instances belonging to each cluster as diagonal entries. Such initialisation was proven to be an adequate procedure in our domain [25]. The algorithm finishes its execution when convergence is achieved. This convergence is expressed as the lack of sufficient variation (i.e., common threshold set to 10^{-4}) in the objective function (Eq. 6.7) between two consecutive iterations.

The composition of signal and the class membership for unseen instances were predicted using the methods previously mentioned, namely *expectation-maximisation* (EM in the acronym used in the tabulated results), *reconstructed sources* (RS), and a mixture of both strategies (EMRS), where the reconstructed sources were used in the EM algorithm aiming at making the most of each method. Standard unsupervised CNMF was also employed for comparison purposes.

The initialisation of $\mathbf{Q}$ was done differently depending on the prediction algorithm: for CNMF and RS strategies, the distance between each instance to be predicted and centroids of K-means from the training phase was used to assign initial class memberships and, then, the initial values of $\mathbf{Q}$ were set following the same approach as for $\mathbf{H}$, while the average value of $\mathbf{H}$ was used to initialise $\mathbf{Q}$ in EM and EMRS.

For the sake of interpretability, $\mathbf{S}$ was normalised to vector unit length and each element of $\mathbf{H}$ and $\mathbf{Q}$ was divided by L1 norm of its column at the end of every execution. The class membership for unknown labels was assigned to the source contributing the most to the signal composition, as evaluated by the highest value in $\mathbf{H}$ and $\mathbf{Q}$.

The most adequate value for parameter $\alpha \in (0,1)$ was estimated using grid search at intervals of 0.05, such that the average BAC (class prediction metric, Section 3.1.3) and the Pearson linear correlation (COR) between

sources in **S** and class centroids over 10-fold cross-validation in the training set was maximised. COR between two random variables X and Y is mathematically defined as:

$$\mathrm{COR} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

Cov being the covariance and σ_X the standard deviation of X.

The reported results are evaluated over the test set for *synthetic* data; and repeated double 10-fold cross-validation, where the inner loop was used in the training phase and the outer one for testing, in the *real* SV-^{1}H-MRS data scenario.

6.4.2 Results

Table 6.1: BAC/COR results for the test set using the *synthetic* data. The values for each method are displayed columnwise. Each row corresponds to one of the analysed diagnostic discrimination problems

Problem	TE	CNMF	EM	RS	EMRS
met vs *ac*2	Short	0.96/0.99	0.97/0.99	0.97/0.99	0.97/0.99
gbm vs *ac*2		0.94/0.93	0.96/**0.97**	0.94/**0.97**	0.94/**0.97**
gbm vs *met*		0.52/0.69	**0.56**/**0.99**	0.54/**0.99**	**0.56**/**0.99**
met vs *ac*2	Long	0.86/0.94	**0.96**/**0.96**	0.86/**0.96**	0.94/**0.96**
gbm vs *ac*2		0.78/0.87	0.84/0.88	0.81/**0.89**	**0.90**/**0.89**
gbm vs *met*		0.59/0.82	0.60/**0.98**	0.61/**0.98**	0.61/**0.98**

The results obtained for the *synthetic* test data can be found in Table 6.1. They consistently show equal or greater performance in classification when using DCNMF variants as compared to CNMF alone, according to the BAC measure. This gain is more evident in LTE data than STE, reaching up to 12% in the discrimination between *gbm* from *ac*2 using the EMRS technique. For LTE, the highest accuracy is obtained for the *met* vs. *ac*2 problem, by increasing it up to 10% for EM and 8% using EMRS. Notice the high difficulty of differentiating *gbm* from *met* at either LTE or STE, where barely 60% accuracy is obtained at best. This result enforces our hypothesis presented in the previous chapters that *ad-hoc* techniques

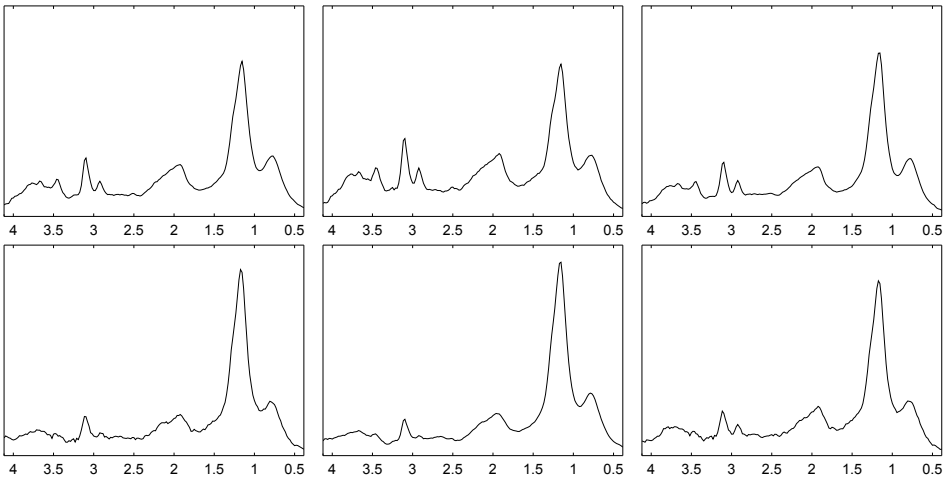


Figure 6.1: Correlation between *glioblastomas, metastases* and sources at short TE for the analysed synthetic data - The top row, from left to right, shows the *gbm* average spectrum in the training set, the estimated source for *gbm* using the EM algorithm, and the *gbm* average spectrum in the test set. The bottom row contains the same information for *met*. The X-axes units are ppm.

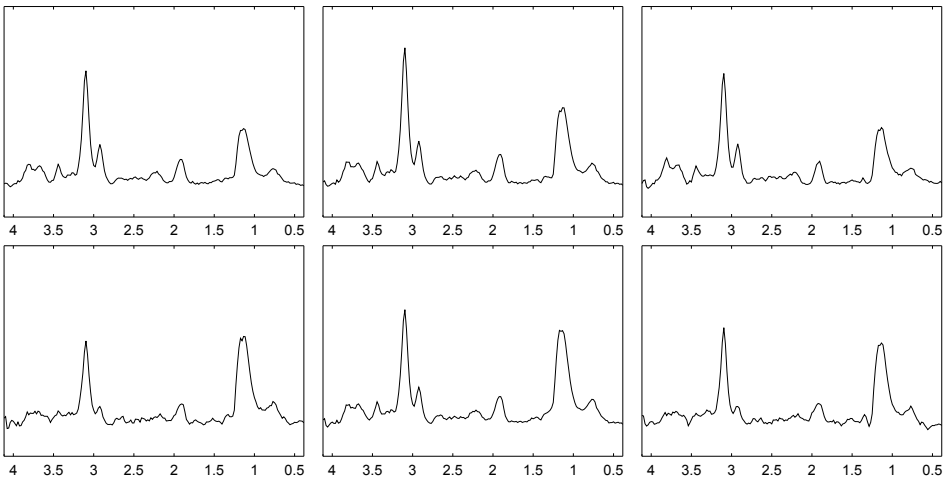


Figure 6.2: Correlation between *glioblastomas, metastases* and sources at long TE for the analysed synthetic data - The top row, from left to right, shows the *gbm* average spectrum in the training set, the estimated source for *gbm* using the RS algorithm, and the *gbm* average spectrum in the test set. Bottom row contains the same information for *met*. The X-axes units are ppm.

including robust feature selection and ensemble techniques are required to address it.

Nonetheless, the real value of the proposed solution in its multiple variants is its ability to retrieve class-specific sources; that is, the extracted individual sources that highly correlate with mean spectrum of a tumour type. Evidence of such quality can be appreciated noting that most of DCNMF obtain a Pearson correlation coefficient over 0.89 (89%), equalling or improving the CNMF's results. Specially striking is the improvement acquired by DCNMF in the *gbm* vs. *met* setting at STE, being more than 30% to its CNMF counterpart; this exemplifies the added value of DCNMF which succeeds in generating very class-specific sources, even for a problem with so much overlapping and ambiguity.

Table 6.2: Repeated double cross-validation BAC (means ± standard deviations) results for the real SV-^{1}H-MRS data

Problem	TE	CNMF	EM	RS	EMRS
met vs *ac*2	Short	0.97 ± 0.05	0.95 ± 0.09	0.97 ± 0.05	0.95 ± 0.11
gbm vs *ac*2		0.92 ± 0.04	0.92 ± 0.05	0.90 ± 0.06	0.92 ± 0.05
gbm vs *met*		0.58 ± 0.10	0.59 ± 0.10	0.58 ± 0.11	0.58 ± 0.06
*as*2 vs *nom*		0.87 ± 0.13	0.85 ± 0.17	0.87 ± 0.21	0.80 ± 0.20
met vs *nom*		0.97 ± 0.05	0.97 ± 0.05	0.97 ± 0.05	0.97 ± 0.05
gbm vs *nom*		0.91 ± 0.07	0.92 ± 0.04	0.91 ± 0.08	0.92 ± 0.06
met vs *ac*2	Long	0.84 ± 0.11	0.84 ± 0.11	0.86 ± 0.14	0.85 ± 0.12
gbm vs *ac*2		0.71 ± 0.09	0.71 ± 0.09	0.74 ± 0.13	0.72 ± 0.10
gbm vs *met*		0.59 ± 0.15	0.59 ± 0.15	0.59 ± 0.17	0.57 ± 0.17
*ac*2 vs *nom*		1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
met vs *nom*		0.90 ± 0.08	0.90 ± 0.08	0.92 ± 0.12	0.92 ± 0.12
gbm vs *nom*		0.72 ± 0.01	0.63 ± 0.19	0.67 ± 0.16	0.72 ± 0.01

To illustrate the DCNMF potential regarding the generation of class-specific sources, Figs. 6.1 and 6.2 show the spectra that have been acquired by our method when comparing *gbm* and *met* at STE using EM and LTE with RS, respectively. Notice the high similarity between training averages, retrieved sources, and unseen test means.

Turning our attention to *real* SV-^{1}H-MRS dataset, the limitations of dealing with such small sample size set, as well as the lack of a proper test

set, become apparent. Nevertheless, and in consonance to previous experiments, DCNMF solutions yield equal or better BAC values when compared to CNMF ones (Table 6.2), although differences are smaller in this case. Another observation is that there is little difference among DCNMF algorithms, the discrimination of low-grade astrocytomas and high-grade tumours being quite complete, while the differentiation between *gbm* and *met* remains very difficult. The extra discriminations involving *nom* controls are fairly easy.

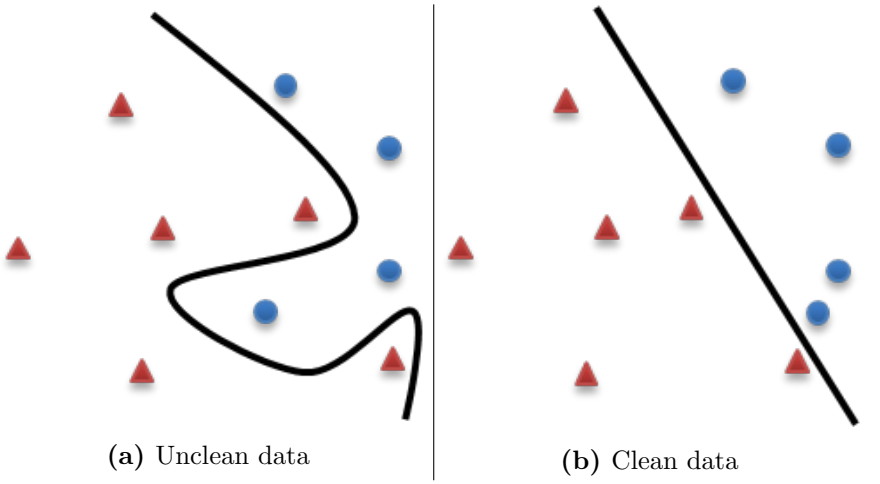
Focusing on the correlation between extracted sources and class-specific averages (Table 6.3), which is the ultimate goal of the study, we observe a coherently similar or better performance when using DCNMF methods, in comparison to CNMF. Once more, the improvement of 33% in the *gbm* vs. *met* problem at STE using EMRS and the nearly 23% in *gbm* vs. *nom* at LTE obtained by EM are especially significant. It seems obvious, though, that finding the best DCNMF variant is a problem-dependent matter.

Table 6.3: Mean correlations (± standard deviations) between tumour type averages and estimated sources in a repeated double 10-fold cross-validation for the real SV-^{1}H-MRS data, displayed as in the previous table

Problem	TE	CNMF	EM	RS	EMRS
met vs *ac*2	Short	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.03	0.99 ± 0.00
gbm vs *ac*2		0.97 ± 0.00	0.98 ± 0.00	0.96 ± 0.04	0.98 ± 0.00
gbm vs *met*		0.65 ± 0.02	0.71 ± 0.02	0.96 ± 0.04	$\mathbf{0.98 \pm 0.00}$
*ac*2 vs *nom*		0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.01
met vs *nom*		1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00
gbm vs *nom*		0.98 ± 0.00	0.98 ± 0.00	0.97 ± 0.01	0.98 ± 0.00
met vs *ac*2	Long	0.93 ± 0.00	0.94 ± 0.00	0.90 ± 0.06	0.94 ± 0.00
gbm vs *ac*2		0.78 ± 0.01	0.80 ± 0.01	$\mathbf{0.82 \pm 0.03}$	0.80 ± 0.01
gbm vs *met*		0.76 ± 0.02	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.86 \pm 0.15}$	$\mathbf{0.80 \pm 0.02}$
*ac*2 vs *nom*		1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
met vs *nom*		0.96 ± 0.00	0.96 ± 0.00	0.92 ± 0.05	0.96 ± 0.00
gbm vs *nom*		0.70 ± 0.02	$\mathbf{0.93 \pm 0.01}$	0.72 ± 0.09	0.71 ± 0.03

6.4.3 Discussion

To begin with this discussion, we would like to remind of the primary goal that the current solution was designed for: accurately identifying the interpretable latent sources out of which the measured signal is made of, with the



(a) Unclean data **(b)** Clean data

Figure 6.3: DCNMF data cleaning - An example of the type of cleaning that DCNMF is performing in original feature space.

aid of labelled data that improves class-specific source determination. The proposed algorithm is especially appealing for complicated problems where the added value of supervised strategy is made clear.

The assessment of predictive ability by means of BAC has been performed with the purpose of guiding the process and ensuring that no harmful effect is caused by our method when compared to previous approaches. However, no explicit effort has been made to improve classification performance and hence, results are rather suboptimal. A more prediction-oriented attempt might build a classifier using the low-dimensional representation of instances stored in $\mathbf{H}$.

A related decision in the construction of the evaluation benchmark has been to constrain the number of generating sources to be equal to the number of tumour types in the problem (i.e., $r = k$). The reason for such decision is that we wanted to obtain class-specific sources resembling class-average spectra, therefore a one-to-one relationship was forced. Nonetheless, nothing prevents our algorithm to be applied using different values for r in those cases where predictive performance is prioritised (allowing for a higher low-dimensional representation of instances where to build the classifiers) or several heterogeneous sources are allowed to represent a tumour type.

The last remark we would like to make is about the ability of the method to *clean* the data. By cleaning, we mean here obtaining a reconstructed version of the initial signal data points in the original data space presenting better discriminative properties. For instance, when using the synthetic dataset in Figure 6.3, noisy or outlying instances falling in the opposite-class region are reconstructed to a version falling in the same-class region.

6.5 Conclusions

The proposed *Discriminant Convex Non-negative Matrix Factorisation* is a supervised signal processing method specifically designed to decompose the measured multivariate data into two interpretable components: the underlying class-specific generating sources and the contribution of each source to the measured signal in form of positive coefficients. The domain of application for which it has been designed is that of SV-^{1}H-MRS data for brain tumour diagnosis, where multiple phenomena contributing to the signal measured by MR scanners in a voxel coexist. The following requirements introduced in Section 6.1 are revisited to evaluate the completeness of our approach:

1. *It must be able to identify the underlying sources present in the retrieved signal*: each column in matrix $\mathbf{S}$ contains an estimated source.

2. *It needs to assess the contribution of each source to the signal*: matrix $\mathbf{H}$ contains the positive coefficients representing the source contribution to every instance.

3. *Both the sources and their contributions must be easily interpretable*: source matrix $\mathbf{S}$ is provided in the same space as the input vectors, while the instance coefficients in $\mathbf{H}$ are normalised to sum up to 1 in the range $(0, 1)$, becoming easily interpretable outcomes.

4. *The solution must naturally deal with both negative and positive values*: this is addressed by imposing sources to be a convex combination of instances.

5. *Ratios between values of metabolites at certain frequencies must be preserved*: this is the rationale behind specifically dealing with negative values instead of shifting the whole spectrum.

6. *Distances between values of metabolites at specific frequencies must be kept*: the same reasoning as in previous statement applies here.

7. *Supervised information must be easily included in the solution when labelled data are available*: the discriminant part of the objective function takes care of this requirement.

A benchmark including synthetic and real neuro-oncological instances from INTERPRET database selected data sets has been designed to demonstrate the ability of the proposed technique to extract tumour type-specific sources in difficult discriminative problems such as the differentiation of glioblastomas from metastases, obtaining notable improvements in certain settings with respect to available state of the art algorithms.

Chapter 7

Probabilistic Matrix Factorisation

In the previous Chapter, we have developed a supervised method to extract the heterogeneous sources responsible to generate the observed SV-^{1}H-MRS signal produced by the tissues in a specific voxel, as captured in a regular scanning. This method was also able to quantify the contribution of each source to the final signal.

During the development of the DCNMF technique, we have identified some important issues that would require specific attention: one of these issues is the selection of the most appropriate number of sources that conform the retrieved signal. As the discussion in Chapter 6 argues, the heuristic used to select the appropriate number of sources has consisted in matching the number of tumour types in the current classification problem. This decision, even if practical for interpretation purposes, might be far from optimal. A second unaddressed issue is the assessment of the confidence that can be placed on the possibility that the provided sources (or pieces of them) are good candidates for the description of the tissues they represent. Such concern is even more prominent in those situations where few instances are used to extract the sources.

To confront the aforementioned issues and some other lower-level ones, we derive a *probabilistic* interpretation of *convex* and *semi* NMF in the current chapter, with the purpose of retaining all the known strengths of

these techniques when applied to our current domain, while incorporating the bonus features that the Bayesian framework may provide.

The structure of the present chapter is described next: a few paragraphs motivating the change of paradigm we adopt is first presented; a guided journey from classical Matrix Factorisation (MF) to full Bayesian treatment for MF, while introducing important concepts to understand Bayesian techniques is then presented. Thereafter, a more specific revision of published research dealing with probabilistic versions of NMF is carried out as an introduction to our detailed derivation of *Probabilistic CNMF* and fully *Bayesian SNMF*. Experiments testing the proposed methods in the brain oncology domain are performed and their results discussed. The last section concludes the chapter by validating the initial hypotheses.

7.1 Motivation

Recall the plausible assumption from previous chapter, stating that the measured SV-^{1}H-MRS signal is composed of a mixture of various signals emitted by different compounds, which contribute with varying intensity. We have seen how the purpose of CNMF is to retrieve, from measurements, both the underlying sources and their contributions, taking into account all those restrictions enumerated in Section 6.1, which, for the sake of brevity, are not repeated here. Notice, though, that all the requirements in that list but number 7 still hold in the current research.

In another order of things, we have seen in Chapter 5 the problem that overfitting poses to our models in the context of feature selection, which is equally relevant in source extraction. The constructed models will be especially prone to this phenomenon when small size data samples are used in their learning phase.

Bearing all these inputs in mind, the solution we propose must incorporate a list of new preconditions to the first six requirements in Section 6.1, intending to ease the construction of models, avoid overfitting and provide elements for a better interpretation and reliability of results:

1. The possibility to incorporate prior knowledge on sources and their contributions.
2. Automatic control of regularisation hyperparameters.
3. Appropriately handle uncertainty and provide an interpretable measure of confidence for the retrieved sources.
4. Suitable selection of the most appropriate number of underlying sources.

7.2 State of the Art

The NMF variants introduced in the previous chapter represent a (non-negatively) constrained subset of a wider problem whose goal is to approximate a real-valued matrix of observations by a lower-rank one. Singular Value Decomposition is a technique often used to obtain such low-rank matrix as a product of various low-dimensional matrices. However, there exist certain domains (e.g., MF in recommender systems - RecSys [154]) where this technique cannot be employed (due to the usually extreme sparsity of the original matrix), and those matrices are found using optimisation-based strategies. In this section, we review some work in the RecSys domain as an example to introduce different scenarios that provide a probabilistic interpretation of MF, while linking them to their classical counterparts.

7.2.1 Classical Matrix Factorisation

Let $\mathbf{X} \in \mathbb{R}^{D\times N}$ be the matrix of observed instances, we aim at finding two lower rank matrices $\mathbf{S} \in \mathbb{R}^{D\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times N}$ such that $\mathbf{X} \approx \mathbf{SH}$. A typical objective function to evaluate the proposed solution is one that minimises the sum-of-squares error:

$$\Omega_{LS}(\mathbf{X}||\mathbf{SH}) = \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}(\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n})^2. \tag{7.1}$$

Notice the importance of choosing the rank K parameter in order to appropriately capture the latent factors underlying the distribution: a too small value might incur a large error in reconstructing the instances (underfitting),

while a too big one might identify each latent factor as the source generating a single instance (overfitting). The choice of the most adequate value for K is a domain-dependent problem, which is difficult to solve using only prior knowledge. A traditional strategy to avoid overfitting, widely employed in Machine Learning literature, is *regularisation*. It consists in adding a term to the objective function that forces the learned function to be as smooth as possible, while keeping the faithfulness of data modelling acceptable. In our context, the regularisation term is not applied directly to the K parameter, but it is dealt with indirectly by keeping the values of the factorised matrices low. This is accomplished in the Ridge Regression technique, which adds the L_2-norm as a regularisation term to the objective, in the form:

$$\Omega_{ridge}(\mathbf{X}||\mathbf{SH}) = \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}(\mathbf{X}_{d,n}-(\mathbf{SH})_{d,n})^2+\frac{\lambda}{2}\sum_{d=1}^{D}\|\mathbf{S}_{d,:}\|_F^2+\frac{\gamma}{2}\sum_{n=1}^{N}\|\mathbf{H}_{:,n}\|_F^2, \tag{7.2}$$

λ and γ being two hyperparameters that regulate the trade-off between learning with maximum fit and keeping the function smooth. Alternatively, a common technique called Lasso [155] is used to obtain a more sparse decomposition by employing the L_1-norm:

$$\Omega_{lasso}(\mathbf{X}||\mathbf{SH}) = \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}(\mathbf{X}_{d,n}-(\mathbf{SH})_{d,n})^2+\frac{\lambda}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}|\mathbf{S}_{d,k}|+\frac{\gamma}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}|\mathbf{H}_{k,n}|.$$

7.2.2 Probabilistic Matrix Factorisation

Unconstrained MF is often applied in the RecSys domain: each cell in the matrix of observations $\mathbf{X} \in \mathbb{R}^{D\times N}$ contains the rating that a user d gives to a certain item n. The task of the system is to provide predictions of ratings for unknown user-item pairs, in order to recommend those items that best suit each user preferences.

Let us now reformulate the MF from a probabilistic perspective within the RecSys domain [156]. Given that the fitting measure to minimise in Eq. 7.1 corresponds to the least-squares error, an equivalent probabilistic formulation would entail using a linear model with Gaussian observation

noise. Hence:

$$p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) = \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{d,n} \mid (\mathbf{SH})_{d,n}, \sigma^2\right), \tag{7.3}$$

where

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

is the probability density function of the Gaussian distribution with mean μ and variance σ^2.

Estimating the values for $\mathbf{S}_{d,k}$ and $\mathbf{H}_{k,n}$ such that Eq. 7.3 is maximised is known as *maximum likelihood* estimation, which is equivalent to minimising Eq. 7.1 and is usually solved by minimising the negative logarithm of the likelihood. That is:

$$-\log p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) = \frac{1}{2\sigma^2} \sum_{d=1}^{D} \sum_{n=1}^{N} (\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n})^2 + \frac{DN}{2} \log \sigma^2 + C,$$

where C is a constant that does not depend on the parameters.

As when using least-squares as the loss function, maximum likelihood is also prone to overfitting. The probabilistic approach to control model complexity consists in specifying a Bayesian prior for each $\mathbf{S}_{d,k}$ and $\mathbf{H}_{k,n}$, which are often set to be random variables from a zero-mean Gaussian distribution:

$$p\left(\mathbf{S} \mid \sigma_S^2\right) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{S}_{d,k} \mid \mu_S, \sigma_S^2\right), \tag{7.4}$$

$$p\left(\mathbf{H} \mid \sigma_H^2\right) = \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{H}_{k,n} \mid \mu_H, \sigma_H^2\right), \tag{7.5}$$

where $\mu_S = \mu_H = 0$; leading to the following objective function:

$$p\left(\mathbf{S}, \mathbf{H} \mid \mathbf{X}, \sigma^2, \sigma_S^2, \sigma_H^2\right) \propto p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) p\left(\mathbf{S} \mid \sigma_S^2\right) p\left(\mathbf{H} \mid \sigma_H^2\right). \tag{7.6}$$

By minimising the negative logarithm of Eq. 7.6, we obtain the so-called *maximum a posteriori* (MAP) estimate:

$$\begin{aligned} -\log p\left(\mathbf{S}, \mathbf{H} \mid \mathbf{X}, \sigma^2, \sigma_S^2, \sigma_H^2\right) \propto\ & \frac{1}{2\sigma^2} \sum_{d=1}^{D} \sum_{n=1}^{N} (\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n})^2 \\ +\ & \frac{1}{2\sigma_S^2} \sum_{d=1}^{D} \mathbf{S}_{d,:}^{\top} \mathbf{S}_{d,:} + \frac{1}{2\sigma_H^2} \sum_{n=1}^{N} \mathbf{H}_{:,n}^{\top} \mathbf{H}_{:,n} \\ +\ & \frac{DN}{2} \log \sigma^2 + \frac{DK}{2} \log \sigma_S^2 + \frac{KN}{2} \log \sigma_H^2 + C, \end{aligned}$$

which is equivalent to the Ridge Regression strategy in Eq. 7.2 with $\lambda = \sigma^2/\sigma_S^2$ and $\gamma = \sigma^2/\sigma_H^2$. The optimisation procedure can be carried out using an Iterative Gradient Descent algorithm, where matrices $\mathbf{S}$ and $\mathbf{H}$ are alternatively updated at each iteration while keeping the other fixed.

So far, we have seen that a probabilistic formulation that uses Gaussian observation noise can be compared to a classical Least Squares loss function and also that using Gaussian priors for the latent factors has the same effect as $L2$-norm regulariser in Ridge Regression. Another typical setting consists in using the Laplacian distribution to model the prior knowledge for latent factors, which is equivalent to the $L1$-norm regularisation strategy in Lasso. We might think that probabilistic NMF is just a reinterpretation of classical NMF, but we start devising the first benefits of this new approach: prior probabilities are not only useful for regularisation, but they also allow to model our prior beliefs on the latent factors, pulling the estimated values towards such priors, an interesting property when few data are available.

7.2.2.1 Hierarchical Bayes

Another useful characteristic of the probabilistic approach is the automatic control of the hyperparameters (i.e., λ and γ) by using the so-called hierarchical Bayes techniques, instead of adjusting their values by explicitly examining a set of candidates in a cross-validation set-up, as it is often the case in classical approaches. Basically, it consists in treating the unknown hyperparameters the same way as the other unknown parameters in the formulation: they are random variables drawn from a distribution. Now, the objective function we aim at optimising becomes:

$$\begin{aligned} p\left(\mathbf{S}, \mathbf{H}, \sigma^2, \sigma_S^2, \sigma_H^2 \mid \mathbf{X}\right) \quad \propto \quad & p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) p\left(\mathbf{S} \mid \sigma_S^2\right) p\left(\mathbf{H} \mid \sigma_H^2\right) \\ & p\left(\sigma^2\right) p\left(\sigma_S^2\right) p\left(\sigma_H^2\right), \end{aligned} \tag{7.7}$$

$p\left(\sigma^2\right)$, $p\left(\sigma_S^2\right)$ and $p\left(\sigma_H^2\right)$ being the appropriately chosen prior distributions for the hyperparameters. Following the same procedure as in the previous block, the MAP point estimate for Eq. 7.7 can be obtained by minimising $-\log p\left(\mathbf{S}, \mathbf{H}, \sigma^2, \sigma_S^2, \sigma_H^2 \mid \mathbf{X}\right)$, using an Iterative Gradient Descent optimisation algorithm, including the hyperparameters in the updating loop.

The last strategy we want to comment before introducing the full Bayesian approach is known as *empirical Bayes* and lies somewhat between the MAP technique with manual hyperparameters setting and the MAP with automatic control of hyperparameters. Now, we also attempt to optimise Eq. 7.7, but we use a different approach to estimate the hyperparameters: at each iteration of the gradient descent algorithm, the prior hyperparameter distributions are approximated by a δ-function at their mode, according to the available data. That is:

$$\begin{aligned} p\left(\sigma^2\right) &\approx \delta\left(p\left(\sigma^2 \mid \mathbf{X}, \mathbf{S}, \mathbf{H}\right)\right) \propto \arg\max\left\{p\left(\mathbf{X}, \mathbf{S}, \mathbf{H} \mid \sigma^2\right) p\left(\sigma^2\right)\right\}, \\ p\left(\sigma_S^2\right) &\approx \delta\left(p\left(\sigma_S^2 \mid \mathbf{S}\right)\right) \propto \arg\max\left\{p\left(\mathbf{S} \mid \sigma_S^2\right) p\left(\sigma_S^2\right)\right\}, \\ p\left(\sigma_H^2\right) &\approx \delta\left(p\left(\sigma_H^2 \mid \mathbf{H}\right)\right) \propto \arg\max\left\{p\left(\mathbf{H} \mid \sigma_H^2\right) p\left(\sigma_H^2\right)\right\}. \end{aligned}$$

7.2.3 Bayesian Probabilistic Matrix Factorisation

Nonetheless, the real value of using a probabilistic interpretation of a learning problem consists in carrying out a full Bayesian treatment and make use of *marginalisation* techniques to get rid of nuisance parameters to obtain the final solution. Recall the Bayes' rule from Chapter 3, where:

$$posterior = \frac{likelihood \times prior}{evidence}; \tag{7.8}$$

if we only focus on finding the function that best fits our data, we are dealing exclusively with the *likelihood* part of the equation, and finding the best parameterisation can be achieved by estimating the *maximum likelihood* as shown above. Yet, we can incorporate prior information to find the best fit, providing the benefits we just explained; in this scenario, we would be using the $likelihood \times prior$, whose best estimates are often obtained using MAP. But what would really supply all the power of Bayesian inference (e.g., uncertainty handling) is the calculation of the whole posterior distribution, and not just a point estimate. In this case, we would need to calculate complex integrals in either the numerator or denominator of the equation, which often require resorting to approximate inference.

An approach to full Bayesian MF in RecSys can be found in [157]. We use this study as an example to introduce a few relevant concepts in the Bayesian

inference framework: as a starting point, let us assume the likelihood of the observed data to be given by Eq. 7.3 and the priors for the latent factors to be normal distributed as in Eqs. 7.4 and 7.5, but now expressed as uncorrelated multivariate Gaussian using precision instead of variance:

$$\begin{aligned} p\left(\mathbf{S} \mid \boldsymbol{\mu}_S, \boldsymbol{\Lambda}_S\right) &= \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{S}_{d,:} \mid \boldsymbol{\mu}_S, \boldsymbol{\Lambda}_S^{-1}\right), \\ p\left(\mathbf{H} \mid \boldsymbol{\mu}_H, \boldsymbol{\Lambda}_H\right) &= \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{H}_{:,n} \mid \boldsymbol{\mu}_H, \boldsymbol{\Lambda}_H^{-1}\right), \end{aligned}$$

where $\boldsymbol{\Lambda}_S = 1/\sigma_S^2\mathbf{I}$ and $\boldsymbol{\Lambda}_H = 1/\sigma_H^2\mathbf{I}$, $\mathbf{I}$ being the identity matrix.

7.2.3.1 Conjugate priors

Recall the use of hierarchical Bayes introduced above, consisting in placing a prior distribution to the unknown hyperparameters. Here, we proceed in a similar way defining prior distributions for hyperparameters related to matrices $\mathbf{S}$: $\boldsymbol{\theta}_S = \{\boldsymbol{\mu}_S, \boldsymbol{\Lambda}_S\}$; and $\mathbf{H}$: $\boldsymbol{\theta}_H = \{\boldsymbol{\mu}_H, \boldsymbol{\Lambda}_H\}$; but we go one step forward to introduce the concept of conjugate priors. Now, the chosen prior distribution does not only need to capture our prior beliefs on the random variable, but it also has to be mathematically suitable, so that when multiplied by the likelihood, the resulting posterior is of the same family as the prior. In this case, the prior and the posterior are called conjugate distributions, the former being a conjugate prior for the current likelihood. The practical applicability of conjugate priors is that they facilitate the computation of the posterior, which might be otherwise intractable.

In our MF example [157], hyperparameter priors were modelled using the Gaussian-Wishart distribution, which is a conjugate prior of a multivariate normal, and is defined as:

$$\begin{aligned} p\left(\boldsymbol{\theta}_S \mid \boldsymbol{\theta}_0\right) &= \mathcal{N}\left(\boldsymbol{\mu}_S \mid \boldsymbol{\mu}_0, (\beta_0\boldsymbol{\Lambda}_S)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_S \mid \mathbf{W}_0, \nu_0\right), \\ p\left(\boldsymbol{\theta}_H \mid \boldsymbol{\theta}_0\right) &= \mathcal{N}\left(\boldsymbol{\mu}_H \mid \boldsymbol{\mu}_0, (\beta_0\boldsymbol{\Lambda}_H)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_H \mid \mathbf{W}_0, \nu_0\right); \end{aligned}$$

$$\mathcal{W}\left(\boldsymbol{\Lambda} \mid \mathbf{W}_0, \nu_0\right) = \frac{1}{C}|\boldsymbol{\Lambda}|^{(\nu_0 - K - 1)/2} \exp\left(-\frac{1}{2} Tr\left(\mathbf{W}_0^{-1}\boldsymbol{\Lambda}\right)\right)$$

being the Wishart distribution, with constant C and initial hyperparameters $\boldsymbol{\theta}_0 = \{\boldsymbol{\mu}_0, \nu_0, \beta_0, \mathbf{W}_0\}$, set, for convenience, to be $\boldsymbol{\mu}_0 = 0$, $\nu_0 = K$ and $\mathbf{W}_0 \in \mathbb{R}^{K \times K} \equiv \mathbf{I}_K$ (identity matrix).

7.2.3.2 Sampling approximations

Unlike in BSS domains, in a RecSys context we are not much interested in retrieving the posterior distribution for the factors, but in predicting new ratings for a user to a specific item, which corresponds to filling out a cell in our $\mathbf{X}$ matrix. Therefore, the *predictive distribution* we aim at finding is:

$$p\left(\mathbf{X}^*_{d,n} \mid \mathbf{X}, \boldsymbol{\theta}_0\right) = \iint p\left(\mathbf{X}^*_{d,n} \mid \mathbf{S}_{d,:}, \mathbf{H}_{:,n}\right) p\left(\mathbf{S}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\theta}_S, \boldsymbol{\theta}_H\right) \\ p\left(\boldsymbol{\theta}_S, \boldsymbol{\theta}_H \mid \boldsymbol{\theta}_0\right) \mathrm{d}\left\{\mathbf{S}, \mathbf{H}\right\} \mathrm{d}\left\{\boldsymbol{\theta}_S, \boldsymbol{\theta}_H\right\}. \tag{7.9}$$

Analysing Eq. 7.9, we realise that the first term of the integral corresponds to the prediction of a new rating, given the model parameters $\{\mathbf{S}_{d,:}, \mathbf{H}_{:,n}\}$, while the second and third terms are a factorisation of the *full posterior* $p(\mathbf{S}, \mathbf{H}, \boldsymbol{\theta}_S, \boldsymbol{\theta}_H \mid \mathbf{X})$, whose purpose is to infer the best values for the parameters and hyperparameters of the model given the data. However, we are not really interested in explicitly calculating any parameterisation, but in predicting new ratings, instead. Therefore, all these nuisance parameters are integrated out.

In the very rare cases where such posterior predictive distribution can be calculated analytically, a point estimate best summarising the distribution (i.e., usually the posterior mean) is given as the most probable value, and the width of the distribution is used as an indicator of confidence on the prediction (i.e., a peaked distribution implies high confidence in the predicted value, while a more uncertain outcome is accompanied by a wider distribution).

Unfortunately, most of the time, as in the case of our explanatory RecSys example, the predictive distribution is intractable and approximate inference techniques are required, out of which Monte Carlo simulation variants [110] are most frequently employed.

The basic idea behind the Monte Carlo method for approximating intractable integrals is very simple [158]: let $\mathbf{x}$ be a vector of N random

variables sampled from distribution $p(\cdot)$; then, the task is to evaluate the expectation:

$$E\left[f(\mathbf{x})\right] = \int_{\mathbf{x}} f(x)p(x)\mathrm{d}x.$$

The Monte Carlo integration evaluates the expectation by drawing N samples from $p(\cdot)$ and approximating:

$$E\left[f(\mathbf{x})\right] \approx \frac{1}{N}\sum_{n=1}^{N} f\left(x^{(n)}\right). \tag{7.10}$$

According to the law of large numbers, the approximation can be arbitrarily accurate by increasing the sample size N, provided $x^{(n)}$ are independent.

The problem with the aforementioned technique arises when sampling from $p(\cdot)$ is not feasible, given the complexity of the distribution. To solve this limitation, the Markov Chain Monte Carlo (MCMC) approach proposes to obtain a non-independent set of random variables in the same proportions as if they were sampled independently from $p(\cdot)$; hence, evaluating the expectation as in Eq. 7.10. One way to construct an MCMC estimator with the desired properties, when $p(x)$ can be evaluated, is by using the Metropolis-Hastings algorithm (MH, [159]).

The MH algorithm consists in approximating the target density $p(\cdot)$ by generating a sequence of instances, each one of them obtained from its predecessor, by following three basic steps [160]:

1. Given the last generated instance $x^{(n)}$, sample a candidate instance from the proposal distribution:

$$x^* \sim q\left(x \mid x^{(n)}\right).$$

2. Calculate the acceptance probability, according to:

$$\rho\left(x^{(n)}, x^*\right) = \min\left\{1, \frac{p(x^*)}{p(x^{(n)})}\frac{q\left(x^{(n)} \mid x^*\right)}{q\left(x^* \mid x^{(n)}\right)}\right\}.$$

3. Set $x^{(n+1)} = x^*$ with probability $\rho\left(x^{(n)}, x^*\right)$, otherwise set $x^{(n+1)} = x^{(n)}$.

A particular instance of the MH algorithm, which is especially suitable for high-dimensional distributions, is the Gibbs sampling [159]. It is recommended in those cases where the methods mentioned above are difficult to apply, but, instead, sampling from the full conditional distributions is easy. That is, given a D-dimensional random variable x, the full conditional distribution of x_d is defined as:

$$p(x_d \mid x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D) = \frac{p(x_1, \ldots, x_D)}{p(x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D)}.$$

Following this Gibbs sampling technique, the joint distribution over D dimensions $p(x_1, \ldots, x_D)$ (the target density $p(\cdot)$) is approximated by iteratively sampling from the full conditionals:

$$\begin{aligned}
x_1^{(n)} &\sim p(x_1 \mid x_2^{(n-1)}, \ldots, x_D^{(n-1)}),\\
x_2^{(n)} &\sim p(x_2 \mid x_1^{(n)}, x_3^{(n-1)}, \ldots, x_D^{(n-1)}),\\
&\vdots\\
x_d^{(n)} &\sim p(x_d \mid x_1^{(n)}, \ldots, x_{d-1}^{(n)}, x_{d+1}^{(n-1)}, \ldots, x_D^{(n-1)}),\\
&\vdots\\
x_D^{(n)} &\sim p(x_D \mid x_1^{(n)}, \ldots, x_{D-1}^{(n)}).
\end{aligned}$$

Finally, expectation is calculated as in Eq. 7.10.

Two important questions worth mentioning before finishing this block on MCMC are related to the convergence of the chain to the target distribution. One is called *burn-in* and is defined as the number of samples that need to be discarded at the beginning of the chain before it actually samples from the desired stationary distribution; the second consists in establishing the number of samples to be kept to ensure that the chain has converged to the target distribution. For the time being, there are no formal derivations successfully addressing these two issues and they are usually dealt with through heuristics.

Back to our MF example, the intractable *predictive distribution* in Eq. 7.9 will now be calculated by a Monte Carlo approximation given by:

$$p(\mathbf{X}^*_{d,n} \mid \mathbf{X}, \boldsymbol{\theta}_0) \approx \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{X}^*_{d,n} \mid \mathbf{S}^{(m)}_{d,:}, \mathbf{H}^{(m)}_{:,n}), \tag{7.11}$$

where $\mathbf{S}_{d,:}^{(m)}$ and $\mathbf{H}_{:,n}^{(m)}$ are sampled from a Markov chain with stationary distribution equivalent to the posterior over parameters and hyperparameters of the model: $p(\mathbf{S}, \mathbf{H}, \boldsymbol{\theta}_S, \boldsymbol{\theta}_H \mid \mathbf{X})$.

The desired Markov chain is going to be constructed using the Gibbs sampling technique introduced above, according to the following algorithm:

1. Initialise model parameters $\{\mathbf{S}^{(1)}, \mathbf{H}^{(1)}\}$.
2. For $m = 1, \ldots, M$:
   (a) Sample the hyperparameters:
   $$\begin{aligned} \boldsymbol{\theta}_S^{(m)} &\sim p(\boldsymbol{\theta}_S \mid \mathbf{S}^{(m)}, \boldsymbol{\theta}_0) \\ \boldsymbol{\theta}_H^{(m)} &\sim p(\boldsymbol{\theta}_H \mid \mathbf{H}^{(m)}, \boldsymbol{\theta}_0) \end{aligned}$$
   (b) For $d = 1, \ldots, D$:
   $$\mathbf{S}_{d,:}^{(m+1)} \sim p(\mathbf{S}_{d,:} \mid \mathbf{X}, \mathbf{H}^{(m)}, \boldsymbol{\theta}_S^{(m)})$$
   (c) For $n = 1, \ldots, N$:
   $$\mathbf{H}_{:,n}^{(m+1)} \sim p(\mathbf{H}_{:,n} \mid \mathbf{X}, \mathbf{S}^{(m)}, \boldsymbol{\theta}_H^{(m)})$$
3. Calculate Eq. 7.11.

Due to the use of conjugate priors, the conditional distribution to sample hyperparameter values for the matrix of users in the Gibbs algorithm is a Gaussian-Wishart. It turns out to be an easy-to-sample-from distribution, whose parameters are set using closed form equations, which take parameters from the prior and update them as data have been seen:

$$\begin{aligned} p(\boldsymbol{\theta}_S \mid \mathbf{S}, \boldsymbol{\theta}_0) &= p\left(\boldsymbol{\mu}_S, \boldsymbol{\Lambda}_S \mid \mathbf{S}, \boldsymbol{\mu}_0, \nu_0, \mathbf{W}_0\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_S \mid \boldsymbol{\mu}_0^*, (\beta_0^* \boldsymbol{\Lambda}_S)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_S \mid \mathbf{W}_0^*, \nu_0^*\right), \end{aligned}$$

where

$$\begin{aligned} &\boldsymbol{\mu}_0^* = \frac{\beta_0 \boldsymbol{\mu}_0 + D\bar{\mathbf{S}}}{\beta_0 + D}, \quad \beta_0^* = \beta_0 + D, \quad \nu_0^* = \nu_0 + D, \\ &\mathbf{W}_0^* = \left[\mathbf{W}_0^{-1} + D\bar{\mathbf{C}} + \frac{\beta_0 D}{\beta_0 + D}\left(\boldsymbol{\mu}_0 - \bar{\mathbf{S}}\right)^\top \left(\boldsymbol{\mu}_0 - \bar{\mathbf{S}}\right)\right]^{-1}, \\ &\bar{\mathbf{S}} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{S}_{d,:}, \quad \bar{\mathbf{C}} = \frac{1}{D}\sum_{d=1}^{D}\left(\mathbf{S}_{d,:} - \bar{\mathbf{S}}\right)^\top \left(\mathbf{S}_{d,:} - \bar{\mathbf{S}}\right). \end{aligned}$$

For sampling user feature vector values in the Gibbs algorithm, we use a conditional Gaussian distribution defined as:

$$\begin{aligned} p\left(\mathbf{S}_{d,:} \mid \mathbf{X}, \mathbf{H}, \boldsymbol{\theta}_S\right) &= p\left(\mathbf{S}_{d,:} \mid \mathbf{X}, \mathbf{H}, \boldsymbol{\mu}_S, \boldsymbol{\Lambda}_S, \alpha\right) \\ &= \mathcal{N}\left(\mathbf{S}_{d,:} \mid \boldsymbol{\mu}_d^*, [\boldsymbol{\Lambda}_d^*]^{-1}\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_d^* &= \boldsymbol{\Lambda}_S + \alpha \sum_{n=1}^{N} \mathbf{H}_{:,n} \mathbf{H}_{:,n}^\top, \\ \boldsymbol{\mu}_d^* &= [\boldsymbol{\Lambda}_d^*]^{-1} \left(\alpha \sum_{n=1}^{N} \mathbf{H}_{:,n} \mathbf{X}_{d,n} + \boldsymbol{\Lambda}_S \boldsymbol{\mu}_S \right). \end{aligned}$$

The conditional distribution over item feature vector $p(\mathbf{H}_n \mid \mathbf{X}, \mathbf{S}^{(m)}, \boldsymbol{\theta}_H^{(m)})$ and item hyperparameters $p(\boldsymbol{\theta}_H \mid \mathbf{H}^{(m)}, \boldsymbol{\theta}_0)$ have exactly the same form.

An alternative to sampling would be the use of Variational inference [79] to deterministically approximate the posteriors of our MF example [161]. The basic idea of Variational methods is to pick a tractable distribution that approximates the intractable true posterior; and then to try to make this approximation as close as possible to the true posterior: this reduces inference to an optimisation problem [162, 163].

7.2.3.3 Model selection

An important factor in Bayes' rule (Eq. 7.8), which has so far been overlooked is the *evidence* or *marginal likelihood* term, corresponding to the denominator of the equation. We have not included it in any of our earlier computations aiming at inferring the *posterior distribution* of model parameters or obtaining the *predictive distribution* for a new data point, since marginal likelihood is meant to be constant with respect to the model parameters and is therefore subsumed into the proportionality constant. Nevertheless, there are some situations where explicitly calculating this model evidence is unavoidable. This is the case when applying a Bayesian approach to model selection.

Model selection is the task of picking the right model from a set of models of different complexity; for instance, in an MF context, choosing the

right number of latent factors K is considered a model selection problem. Classical approaches use cross-validation (CV) to estimate the generalisation error of every model in order to select the one that minimises such error. The main pitfall of this strategy is that it requires to fit each model as many times as folds in the CV. A more efficient procedure consists in computing the posterior over models [79]:

$$p(m \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid m)p(m)}{\sum_{m \in \mathcal{M}} p(m, \mathbf{X})}. \tag{7.12}$$

Bayesian model selection achieves its purpose by computing the MAP estimate: $\hat{m} = \arg\max p(m \mid \mathbf{X})$, which, in the case of no prior preference for any model: $p(m) \propto 1$, translates into choosing the model that maximises the marginal likelihood we have been referring to:

$$p(\mathbf{X}) \equiv p(\mathbf{X} \mid m) = \int p(\mathbf{X} \mid \theta)p(\theta \mid m)\mathrm{d}\theta.$$

Unfortunately, computing the marginal likelihood is often intractable and approximate techniques are usually required [164].

7.2.4 Probabilistic Non-negative Matrix Factorisation

In the previous section, we have introduced Bayesian inference with the assistance of an unconstrained MF example. Here, we present further relevant work that draws somehow closer to our domain of application by constraining all matrices in the decomposition to contain only non-negative values. Reviewing this work will provide insights on different avenues to adapt Bayesian techniques to convex or semi NMF.

The first study, published in [165], is an attempt to formalise classical NMF with the Frobenius cost function (Eq. 6.1) using the Bayesian framework. In particular, the likelihood is defined as the multiplication of source and mixing matrices with zero mean Gaussian noise and the non-negative

constraints of the factors are encoded using Gamma distributions. That is:

$$
\begin{aligned}
p\left(\mathbf{S},\mathbf{H} \mid \mathbf{X},\boldsymbol{\theta}\right) &\propto p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\boldsymbol{\theta}\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}\right), \\
p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\boldsymbol{\theta}\right) &= \prod_{d=1}^{D}\prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{d,n}; (\mathbf{SH})_{d,n}, \sigma_n^2\right), \qquad (7.13) \\
p\left(\mathbf{S} \mid \boldsymbol{\theta}\right) &= \prod_{d=1}^{D}\prod_{k=1}^{K} \mathcal{G}\left(\mathbf{S}_{d,k}; \alpha_k, \beta_k\right), \qquad (7.14) \\
p\left(\mathbf{H} \mid \boldsymbol{\theta}\right) &= \prod_{k=1}^{K}\prod_{n=1}^{N} \mathcal{G}\left(\mathbf{H}_{k,n}; \gamma_k, \lambda_k\right), \qquad (7.15)
\end{aligned}
$$

where $\boldsymbol{\theta} = \{\sigma_n^2\}_{n=1}^{N} \cup \{\alpha_k, \beta_k, \gamma_k, \lambda_k\}_{k=1}^{K}$.

A MAP approach was chosen to obtain point estimates for each element in matrices $\mathbf{S}$ and $\mathbf{H}$, a procedure that can be seen as a generalisation of the widely-used Positive Matrix Factorisation algorithm [69], but now containing a different regularisation parameter per source. Alternating iterative gradient descent was employed to optimise the cost function; and empirical hierarchical Bayes was the strategy of choice to estimate hyperparameter values. The suitability of the proposed method was investigated using a synthetically-generated toy example.

These same authors extended their previous work with the purpose of providing full Bayesian inference capabilities [166] to the model. Starting from the same formalisation, they derived a hybrid Gibbs-Metropolis-Hastings MCMC procedure, where Gibbs method was used to sample from the posterior to compute the marginal posterior mean point estimate:

$$
\left(\hat{\mathbf{S}}, \hat{\mathbf{H}}\right) = \mathbb{E}_{p(\mathbf{S},\mathbf{H}|\mathbf{X},\boldsymbol{\theta})}\left\{\mathbf{S},\mathbf{H}\right\}.
$$

For those complicated steps within Gibbs (i.e., posterior densities of sources and mixing coefficients, as well as prior densities for shape parameters of Gamma distributions), Metropolis-Hastings was the method of choice to obtain the appropriate samples. Experiments on synthetic and real data, consisting on the analysis of the spectral mixture (as measured by a near infrared spectrometer) of a compound obtained by experimentally mixing three chemical species, were carried out to evaluate the performance of the proposed strategy.

As the authors pointed out in their article, by setting the shape parameters of the Gaussian distributions (i.e., α_k and γ_k) in Eqs. 7.14 and 7.15 to 1, the distribution becomes an *exponential*, simplifying the computation of the posterior and avoiding the need for MH steps:

$$p\left(\mathbf{S} \mid \boldsymbol{\theta}\right) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{E}\left(\mathbf{S}_{d,k}; \lambda_k\right), \quad p\left(\mathbf{H} \mid \boldsymbol{\theta}\right) = \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{E}\left(\mathbf{H}_{k,n}; \gamma_k\right),$$

where $\{\lambda_k, \gamma_k\}_{k=1}^{K}$. This is precisely the formulation in [167], where a fast and direct Gibbs sampling procedure was derived, by sampling from a rectified normal density (i.e., the product of a normal by an exponential distribution) and exploiting independence to allow simultaneous computation. A specially relevant contribution of this study is the model selection technique based on Chib's method to appropriately choose the best number of sources in the factorisation as a byproduct of Gibbs draws. A point estimate is obtained from the posterior using Iterated Conditional Modes [168]. The suitability of the proposed methodology was evaluated on synthetically-generated data, as well as on real data from chemical shift imaging of a human head and images for face recognition.

In the last study we review [169], authors modelled the divergence between observations and factorised matrices as a Poisson distribution, which corresponds to the Kullback-Leibler divergence variant of NMF (Eq. 6.2). Hence, Eq. 7.13 was replaced by

$$p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}\right) = \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{PO}\left(\mathbf{X}_{d,n}; (\mathbf{SH})_{d,n}\right),$$

keeping Eqs. 7.14 and 7.15 in their original form.

Focusing only on the likelihood, an *expectation-maximisation* algorithm for *maximum likelihood* was described, which proved to be an equivalently theoretically-grounded version of the update rules in Eqs. 6.3. Full Bayesian inference was proposed by means of Variational methods, providing a MAP point estimate using ICM and a strategy to perform model selection. MCMC-like counterparts based on Gibbs were also derived and marginal likelihood

estimation for model selection using Chib's method was provided. Evaluation of the different solutions were performed on both synthetic and real-world images for face detection.

7.3 Probabilistic Semi and Convex Non-negative Matrix Factorisation

Given a matrix of observations $\mathbf{X} \in \mathbb{R}_{\pm}^{D \times N}$, where N is the number of instances and D the dimensionality (number of features or variables), SNMF aims at decomposing this matrix as a linear combination of K D-dimensional sources of mixed sign $\mathbf{S} \in \mathbb{R}_{\pm}^{D \times K}$ and a matrix $\mathbf{H} \in \mathbb{R}_{+}^{K \times N}$ of positive mixing coefficients. As already stated, CNMF is a particular case where sources in $\mathbf{S}$ are obtained as a convex combination of data instances, thus linking them with the notion of centroids in clustering problems. In symbols,

$$\mathbf{X}_{\pm} = \mathbf{S}_{\pm}\mathbf{H}_{+} + \mathbf{E}_{\pm} = \mathbf{X}_{\pm}\mathbf{W}_{+}\mathbf{H}_{+} + \mathbf{E}_{\pm}$$

where $\mathbf{W} \in \mathbb{R}_{+}^{N \times K}$ is the so-called unmixing matrix and $\mathbf{E} \in \mathbb{R}_{\pm}^{D \times N}$ is the error matrix.

7.3.1 A probabilistic formulation for Convex Non-negative Matrix Factorisation

The probabilistic approach for CNMF that we propose in this section uses empirical Bayes strategies to formulate the matrix decomposition using three components: first, a likelihood function to account for the difference between the outcome of the model and the observations; second, a prior distribution for values in the unmixing matrix $\mathbf{W}$; and, finally, a prior distribution over the mixing matrix $\mathbf{H}$. Notice that any assumptions that we make for any of those components must be encoded by these distributions (e.g., non-negativity of elements in mixing and unmixing matrices).

Following the Bayes rule, the joint posterior distribution of the adaptive matrices $\mathbf{W}$ and $\mathbf{H}$ is:

$$p\left(\mathbf{W}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\theta}\right) \propto p\left(\mathbf{X} \mid \mathbf{W}, \mathbf{H}, \boldsymbol{\theta}\right) \cdot p\left(\mathbf{W} \mid \boldsymbol{\theta}\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}\right),$$

$\boldsymbol{\theta}$ being a vector containing all the required hyperparameters associated to the chosen distributions.

In particular, observed instances, as well as the mixing and unmixing coefficients associated to each source, are assumed to be independent and identically distributed (i.i.d.). Residuals conforming the likelihood are assumed to be drawn from a normal distribution centred at 0 and with variances $\{\sigma_n^2\}_{n=1}^N$. Prior densities for latent factors $\mathbf{W}$ and $\mathbf{H}$, as explained in Section 7.2.4, are conveniently chosen to be exponential:

$$\begin{aligned} p\left(\mathbf{X} \mid \mathbf{W}, \mathbf{H}, \boldsymbol{\theta}\right) &= \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{d,n}; (\mathbf{XWH})_{d,n}, \sigma_n^2\right), \\ p\left(\mathbf{W} \mid \boldsymbol{\theta}\right) &= \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{E}\left(\mathbf{W}_{n,k}; \lambda_k\right), \\ p\left(\mathbf{H} \mid \boldsymbol{\theta}\right) &= \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{E}\left(\mathbf{H}_{k,n}; \gamma_k\right), \end{aligned}$$

where $\boldsymbol{\theta} = \{\sigma_n^2\}_{n=1}^N \cup \{\lambda_k, \gamma_k\}_{k=1}^K$.

7.3.1.1 Maximum a Posteriori approach

From the formulation above, a direct way to obtain a point estimate for this distribution is by using the MAP approach, which consists in minimising the negative log-posterior:

$$F\left(\mathbf{W}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\theta}\right) = -\log p\left(\mathbf{W}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\theta}\right),$$

which can be expanded as

$$F\left(\mathbf{W}, \mathbf{H} \mid \mathbf{X}, \boldsymbol{\theta}\right) = F_L\left(\mathbf{X} \mid \mathbf{W}, \mathbf{H}, \boldsymbol{\theta}\right) + F_{P1}\left(\mathbf{W} \mid \boldsymbol{\theta}\right) + F_{P2}\left(\mathbf{H} \mid \boldsymbol{\theta}\right),$$

where

$$
\begin{aligned}
F_L\left(\mathbf{X} \mid \mathbf{W}, \mathbf{H}, \boldsymbol{\theta}\right) &= \sum_{n=1}^{N} \frac{1}{2\sigma_n^2} \sum_{d=1}^{D} \left(\mathbf{X}_{d,n} - (\mathbf{XWH})_{d,n}\right)^2 \\
&= Tr\left[\frac{1}{2}(\mathbf{X} - \mathbf{XWH})\mathbf{V}(\mathbf{X} - \mathbf{XWH})^\top\right], \\
F_{P1}\left(\mathbf{W} \mid \boldsymbol{\theta}\right) &= \sum_{k=1}^{K} \lambda_k \sum_{n=1}^{N} \mathbf{W}_{n,k} = Tr\left[\boldsymbol{\lambda}\mathbf{W}^\top\mathbf{e}^\top\right], \\
F_{P2}\left(\mathbf{H} \mid \boldsymbol{\theta}\right) &= \sum_{k=1}^{K} \gamma_k \sum_{n=1}^{N} \mathbf{H}_{k,n} = Tr\left[\mathbf{e}\mathbf{H}^\top\boldsymbol{\gamma}\right].
\end{aligned}
$$

Here, $\mathbf{V}$ is an $N \times N$ matrix of variance hyperparameters for the Gaussian distribution with σ_n^{-2} in its diagonal; $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_k]$, $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_k]^\top$ are scale hyperparameters of the exponential distributions; and $\mathbf{e}$ is a row unit vector of length K; $\mathbf{A}^\top$ represents the transpose of $\mathbf{A}$ and $Tr[\mathbf{A}]$ its trace. Hence, the cost function to optimise is expressed as:

$$
\begin{aligned}
F &= \frac{1}{2}Tr\left[\mathbf{XVX}^\top + \mathbf{XWHVH}^\top\mathbf{W}^\top\mathbf{X}^\top - 2\mathbf{XVH}^\top\mathbf{W}^\top\mathbf{X}^\top\right] \\
&+ Tr\left[\boldsymbol{\lambda}\mathbf{W}^\top\mathbf{e}^\top\right] + Tr\left[\mathbf{e}\mathbf{H}^\top\boldsymbol{\gamma}\right].
\end{aligned} \tag{7.16}
$$

A closed-form expression to obtain the minimum of the cost function cannot be derived; therefore, an optimisation procedure based on gradient descent, able to deal with mixtures of positively and negatively-valued matrices, is proposed.

First, an update rule for $\mathbf{W}$ will be derived: we start by adding a matrix of Lagrangian multipliers $\boldsymbol{\beta}_{N,K}$ to the cost function to ensure that each $\mathbf{W}_{n,k} \geq 0$:

$$
\begin{aligned}
F &= \frac{1}{2}Tr\left[\mathbf{XVX}^\top + \mathbf{XWHVH}^\top\mathbf{W}^\top\mathbf{X}^\top - 2\mathbf{XVH}^\top\mathbf{W}^\top\mathbf{X}^\top\right] \\
&+ Tr\left[\boldsymbol{\lambda}\mathbf{W}^\top\mathbf{e}^\top\right] + Tr\left[\mathbf{e}\mathbf{H}^\top\boldsymbol{\gamma}\right] - Tr\left[\boldsymbol{\beta}\mathbf{W}^\top\right].
\end{aligned}
$$

Then we calculate the gradient of the objective with respect to $\mathbf{W}$, which must equal 0 at convergence:

$$
\frac{\partial F}{\partial \mathbf{W}} = \mathbf{X}^\top\mathbf{XWHVH}^\top - \mathbf{X}^\top\mathbf{XVH}^\top + \mathbf{e}^\top\boldsymbol{\lambda} - \boldsymbol{\beta} = 0.
$$

According to the the Karush-Kuhn-Tucker (KKT) complementary slackness condition, a fixed point equation that the solution must satisfy at convergence is obtained:

$$\begin{aligned}\left(\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{V}\mathbf{H}^\top - \mathbf{X}^\top\mathbf{X}\mathbf{V}\mathbf{H}^\top + \mathbf{e}^\top\boldsymbol{\lambda}\right)_{n,k}\mathbf{W}_{n,k} \\ = \boldsymbol{\beta}_{n,k}\mathbf{W}_{n,k} = \boldsymbol{\beta}_{n,k}\mathbf{W}_{n,k}^2 \quad = \quad 0.\end{aligned}$$

Next, by decomposing $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^- = (|\mathbf{A}| + \mathbf{A})/2 - (|\mathbf{A}| - \mathbf{A})/2$, the previous equation is transformed into a non-negative one:

$$\begin{aligned}\left((\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{V}\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^-\mathbf{V}\mathbf{H}^\top + \mathbf{e}^\top\boldsymbol{\lambda}\right)_{n,k}\mathbf{W}_{n,k} \\ = \left((\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{V}\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^+\mathbf{V}\mathbf{H}^\top\right)_{n,k}\mathbf{W}_{n,k} \quad .\end{aligned}$$

Solving on $\mathbf{W}$, we obtain an update rule for this matrix:

$$\mathbf{W}_{n,k} \leftarrow \mathbf{W}_{n,k}\sqrt{\frac{((\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{V}\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^+\mathbf{V}\mathbf{H}^\top)_{n,k}}{((\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{V}\mathbf{H}^\top + (\mathbf{X}^\top\mathbf{X})^-\mathbf{V}\mathbf{H}^\top + \mathbf{e}^\top\boldsymbol{\lambda})_{n,k}}},$$

which satisfies the above fixed point equation at convergence.

Using the same approach, an update rule for $\mathbf{H}$ can be derived:

$$\mathbf{H}_{k,n} \leftarrow \mathbf{H}_{k,n}\sqrt{\frac{(\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{V} + \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{V})_{k,n}}{(\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{V} + \mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{V} + \boldsymbol{\gamma}\mathbf{e})_{k,n}}}.$$

7.3.1.2 Hyperparameter estimation

At this point, there is still a crucial decision that needs to be made, which corresponds to the estimation of hyperparameter values for $\boldsymbol{\theta}$. In this study we decided to use the hierarchical empirical Bayes technique to find the best candidates, which are selected as the mode of the distribution over hyperparameters.

First, we estimate σ_n^2 as:

$$\begin{aligned}&p\left(\sigma_n^2 \mid \mathbf{X}, \mathbf{W}, \mathbf{H}\right) \propto \\ &\left(\frac{1}{\sigma_n^2}\right)^{\frac{D}{2}} \exp\left\{-\frac{1}{2\sigma_n^2}\sum_{d=1}^{D}\left(\mathbf{X}_{d,n} - (\mathbf{X}\mathbf{W}\mathbf{H})_{d,n}\right)^2\right\} \times p\left(\sigma_n^2\right).\end{aligned}$$

The prior for noise variance σ_n^2 is distributed as an inverse gamma:

$$\sigma_n^2 \sim \mathcal{IG}\left(\alpha_\sigma^o, \beta_\sigma^o\right) = \frac{(\beta_\sigma^o)^{\alpha_\sigma^o}}{\Gamma(\alpha_\sigma^o)}(\sigma_n^2)^{-\alpha_\sigma^o-1}\exp\left(-\frac{\beta_\sigma^o}{\sigma_n^2}\right).$$

Using a conjugate prior, we obtain the posterior

$$p\left(\sigma_n^2 \mid \mathbf{X}, \mathbf{W}, \mathbf{H}\right) \sim \mathcal{IG}\left(\alpha_\sigma^p, \beta_\sigma^p\right),$$

where

$$\alpha_\sigma^p = \alpha_\sigma^o + \frac{D}{2}, \quad \beta_\sigma^p = \beta_\sigma^o + \frac{1}{2}\sum_{d=1}^{D}\left(\mathbf{X}_{d,n} - \mathbf{XWH}_{d,n}\right)^2.$$

The point estimate is obtained as the mode of the previous $\mathcal{IG}$:

$$\hat{\sigma}_n^2 = \frac{\beta_\sigma^o + \frac{1}{2}\sum_{d=1}^{D}\left(\mathbf{X}_{d,n} - \mathbf{XWH}_{d,n}\right)^2}{\alpha_\sigma^o + \frac{D}{2} + 1}.$$

For the scale factor λ_k:

$$p\left(\lambda_k \mid \mathbf{W}\right) = \lambda_k \exp\left\{-\lambda_k \mathbf{W}\right\} \times p\left(\lambda_k\right)$$

we assume the prior for the parameter λ_k to be distributed as a gamma density

$$\lambda_k \sim \mathcal{G}\left(\alpha_\lambda^o, \beta_\lambda^o\right) = \frac{(\beta_\lambda^o)^{\alpha_\lambda^o}}{\Gamma(\alpha_\lambda^o)}(\lambda_k)^{\alpha_\lambda^o-1}\exp\left(-\beta_\lambda^o\lambda_k\right).$$

By means of conjugacy, we obtain

$$p\left(\lambda_k \mid \mathbf{W}\right) \sim \mathcal{G}\left(\alpha_\lambda^p, \beta_\lambda^p\right),$$

where

$$\alpha_\lambda^p = \alpha_\lambda^o + N, \quad \beta_\lambda^p = \beta_\lambda^o + \sum_{n=1}^{N}\mathbf{W}_{n,k}.$$

The point estimate is chosen as the mode of the above $\mathcal{G}$ density:

$$\hat{\lambda}_k = \frac{\alpha_\lambda^o + N - 1}{\beta_\lambda^o + \sum_{n=1}^{N}\mathbf{W}_{n,k}}.$$

Analogously, we can estimate γ_k as

$$\hat{\gamma}_k = \frac{\alpha_\gamma^o + N - 1}{\beta_\gamma^o + \sum_{n=1}^{N}\mathbf{H}_{k,n}}.$$

A Matlab implementation of the presented algorithm can be found at `http://www.cs.upc.edu/~avilamala/resources/ProbCNMF_Toolbox.zip`

7.3.1.3 Empirical evaluation

A probabilistic formulation for CNMF has been designed to overcome some of the limitations of their classical counterpart. In this section, we report experiments carried out to assess the appropriateness of the MAP estimate in our application domain that concerns the analysis of real SV-^{1}H-MRS data. These results are then discussed in some detail.

Experimental setup Data from the online-accessible and curated INTERPRET repository (Section 2.3) are used to evaluate the current method. In particular, the most clinically relevant 195 spectral frequencies of the SV-^{1}H-MRS instances are selected for each of the 78 *gbm*, 31 *met*, 20 *ac*2 and 15 *nom* spectra acquired at LTE; and for the 86 *gbm*, 38 *met*, 22 *ac*2 and 20 *nom* spectra acquired at STE. Correctly distinguishing the aforementioned types is of great relevance in medical practise. An extra relevant discrimination problem was added to those involving specific tumour types, namely the discrimination of aggressive tumours ($agg = gbm + met$) from other types.

Experiments consisted in estimating the most appropriate tumour type label for each of the available instances (binary classification problem) according to the coefficients in $\mathbf{H}$, while simultaneously providing reliable sources representing each class (columns of $\mathbf{S}$). The quality of the retrieved sources was assessed through a measure of correlation (COR) between each source and the type-specific average spectra. A tumour type label was assigned to every instance according to the source contributing the most to the reconstruction of the observed signal, expressed in $\mathbf{H}$. The AUC was the metric of choice to gauge overall tumour-type imputation.

The hyperparameter controlling the number of sources was set to a value equal to the number of tumour types in each classification problem (i.e., $K = 2$) for purely practical reasons, despite not being the optimal value for source reconstruction. Normalisation to vector unit length was performed to every instance before any further treatment.

Given that the joint optimisation of $\mathbf{W}$ and $\mathbf{H}$ in Eq. 7.16 is not convex, the proposed method is bound to converge to a local minimum, which

means that an adequate and careful initialisation of parameters and hyper-parameters is required. Following [132], K-means initialisation was used, setting K as the number of sources we want to extract. Matrices $\mathbf{H}$ and $\mathbf{W}$ were initialised as $\mathbf{H}^0_{k,n} = l_k + 0.2$, where $l_k \in \{0, 1\}$, with the latter indicating membership to k-th cluster; and $\mathbf{W}^0_{n,k} = (l_k + 0.2)/c_k$; c_k being the number of instances belonging to cluster k. Convergence of the algorithm was assumed when a minimum variation in the cost function between two consecutive iterations was observed: $\epsilon < 10^{-4}$. The *hyperpriors* for noise variance were set to be uninformative: $\alpha^o_\sigma = \beta^o_\sigma = 0.001$; and the *priors* for $\mathbf{H}_{k,n}$ and $\mathbf{W}_{n,k}$ parameters were chosen to match the data amplitude; that is, $p(\mathbf{H}_{n,k} < 1.5) = p(\mathbf{W}_{k,n} < 1.5) = 0.95$, $\alpha^o_\lambda = \beta^o_\lambda = \alpha^o_\gamma = \beta^o_\gamma = 2$.

Results Tables 7.1 and 7.2 show the results obtained by *Probabilistic CNMF* as compared to standard CNMF and K-means algorithm. Our proposed method presents analogous and sometimes better source extraction properties (COR), when compared to CNMF and similar ones in the task of discriminating tumour types (AUC). Both algorithms consistently provide higher classification ability than K-means, as measured by AUC.

Table 7.1: AUC / COR results for LTE data

	K-means	CNMF	Probabilistic CNMF
gbm vs. *met*	0.58 / **0.91**	**0.63** / 0.79	**0.63** / 0.81
gbm vs. *ac*2	0.72 / 0.86	0.93 / 0.80	**0.93** / **0.91**
met vs. *ac*2	0.90 / **0.99**	**0.95** / 0.93	**0.95** / 0.91
*ac*2 vs. *nom*	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
met vs. *nom*	0.92 / **0.98**	**0.97** / 0.96	**0.97** / 0.94
gbm vs. *nom*	0.72 / 0.78	0.92 / 0.71	**0.93** / **0.82**
agg vs. *nom*	0.73 / 0.77	**0.93** / 0.72	**0.93** / **0.79**
agg vs. *ac*2	0.73 / 0.87	**0.95** / 0.83	0.94 / **0.90**

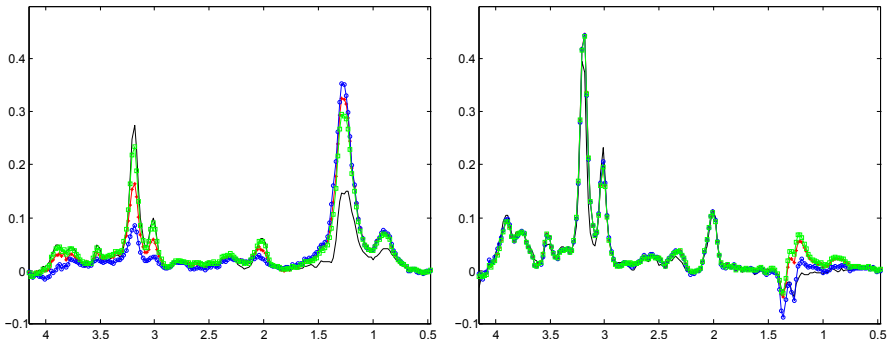
The separate analysis of the results according to data acquisition time modality (LTE or STE), reveals that in the experiments involving LTE spec-

tra, CNMF-like algorithms coherently exhibit better class alignment than the K-means algorithm, as can be seen in the gap of more than 20% AUC in the *gbm* vs. *ac*2, *gbm* vs. *nom*, *agg* vs. *nom* and *agg* vs. *ac*2 classification tasks. By using the current probabilistic approach, a gain of up to 11% in the correlation between extracted sources and class centroids (i.e., *gbm* vs. *ac*2 and *gbm* vs. *nom*) is obtained when compared to its non-probabilistic counterpart. In certain cases, *Probabilistic CNMF* is able to outperform the K-means in source extraction (up to 5% in *gbm* vs. *ac*2 and 4% in *gbm* vs. *nom*).

The source extraction capabilities of our method are exemplified in Figure 7.1, which displays the tumour type representatives obtained by each algorithm in contrast to the class average for the discrimination between *gbm* from *ac*2 in LTE. Although all candidate methods perform reasonably well in retrieving the *ac*2 source despite small irregularities around 1.3*ppm*, big differences exist in the *gbm* tumour type candidate: all of them overemphasize the lipids peak at 1.3*ppm*, the *Probabilistic CNMF* being the one with less deviation from the average; major differences between algorithms can be appreciated in the characteristic Choline (3.3*ppm*), Creatine (3.0*ppm*) and N-Acetyl Aspartate (2.0*ppm*) peaks, which are very well approximated by *Probabilistic CNMF*, according to the class-average source.

Table 7.2: AUC / COR results for STE data

	K-means	CNMF	Probabilistic CNMF
gbm vs. *met*	0.59 / **0.93**	0.64 / 0.70	**0.65** / 0.72
gbm vs. *ac*2	0.92 / 0.99	**0.98** / 0.98	**0.98** / 0.97
met vs. *ac*2	0.97 / 1.00	**1.00** / 0.99	**1.00** / 0.99
*ac*2 vs. *nom*	0.93 / 1.00	0.99 / 0.99	**1.00** / 0.99
met vs. *nom*	0.97 / 1.00	**1.00** / 1.00	**1.00** / 0.99
gbm vs. *nom*	0.92 / 0.98	**0.99** / 0.98	**0.99** / 0.98
agg vs. *nom*	0.93 / 0.98	**0.99** / 0.99	**0.99** / 0.98
agg vs. *ac*2	0.93 / 0.98	**0.98** / 0.99	**0.98** / 0.98



(a) *gbm* LTE **(b)** *ac*2 LTE

Figure 7.1: Sources retrieved by the different algorithms in the *gbm* vs. *ac2* problem using data acquired at LTE - The black solid line represents the average spectrum of *gbm* (a: left figure) and *ac*2 (b: right figure), the lines with asterisk symbols are the sources retrieved by K-means; with circle symbols by CNMF; and with square symbols by Probabilistic CNMF. Y-axes represent unit-free metabolite concentrations and X-axes represent frequency as measured in parts per million (ppm).

Shifting our attention now towards STE data, the considerably good results obtained by all algorithms in almost all discriminative tasks but the known-to-be difficult *gbm* vs. *met* discrimination, leave little room to appreciate the differences in performance amongst the different strategies, even though the same general trend of higher AUC for CNMF versions with respect to K-means can be observed. In the special case already mentioned, K-means performs best in tumour-type specific source retrieval (COR), but at the price of lower class discrimination (AUC) than CNMF variants

7.3.2 Full Bayesian Semi Non-negative Matrix Factorisation

In this last contribution of the thesis, we make the probabilistic formulation of SNMF to be pure Bayesian. This means that, unlike in empirical Bayes, we are not using the observations to estimate any a priori information. The side effect of this decision is that CNMF formulation is not valid any longer, due to the fact that sources $\mathbf{S}$ are a linear combination of the observations. However, the chosen formulation also aims at obtaining highly interpretable results. In this respect (recall Eq. 7.3), elements of source matrix $\mathbf{S} \in \mathbb{R}_{\pm}^{D \times K}$

are encoded as samples from a Gaussian distribution; while the values of the mixing matrix $\mathbf{H} \in \mathbb{R}_{+}^{K \times N}$ are conveniently obtained from an exponential density. Residuals in $\mathbf{E} \in \mathbb{R}_{\pm}^{D \times N}$ are assumed to be i.i.d. zero mean.

Now, according to the Bayes' rule, the joint posterior is defined as:

$$p\left(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X}\right) = \frac{p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right)}{p\left(\mathbf{X}\right)}. \quad (7.17)$$

Notice that calculating the marginal likelihood $p\left(\mathbf{X}\right)$ involves the computation of an intractable integral:

$$p\left(\mathbf{X}\right) = \int_{\mathbf{S}} \int_{\mathbf{H}} \int_{\sigma^2} p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right) \mathrm{d}\left\{\mathbf{S}, \mathbf{H}, \sigma^2\right\}.$$

However, given that the marginal likelihood is constant with respect to the model parameters, we subsume it into the proportionality constant. Hence,

$$p\left(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) \cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) \cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right),$$

where

$$p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) = \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{d,n}; (\mathbf{SH})_{d,n}, \sigma^2\right)$$

is the likelihood function, denoted as

$$\mathcal{N}\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; \quad (7.18)$$

$$p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{S}_{d,k}; \mu_o, \sigma_o^2\right),$$

where $\boldsymbol{\theta}_S = \{\mu_o, \sigma_o^2\}$ are the priors for the source signals, as expressed in Eq. 7.18; and

$$p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) = \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{E}\left(\mathbf{H}_{k,n}; \lambda_o\right),$$

with $\boldsymbol{\theta}_H = \{\lambda_o\}$, corresponds to the prior distribution for the values in the mixing matrix; where $\mathcal{E}\left(x; \lambda\right) = \lambda \exp\left\{-\lambda x\right\}$ is the exponential density. Finally, the prior for the noise variance is appropriately chosen to be an inverse gamma of the form:

$$p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right) = \mathcal{IG}\left(\sigma^2; \alpha_o, \beta_o\right) = \frac{\beta_o^{\alpha_o}}{\Gamma(\alpha_o)} (\sigma^2)^{-\alpha_o - 1} \exp\left(-\frac{\beta_o}{\sigma^2}\right);$$

$\boldsymbol{\theta}_\sigma = \{\alpha_o, \beta_o\}$ being its hyperparameters.

From this joint posterior, we would be interested in estimating the marginal density of each **S** and **H** factor, but this procedure involves the computation of an intractable integral. In the next section, this shortcoming is overcome by deriving an MCMC sampling method.

7.3.2.1 Gibbs sampling approach

In this section, we derive a Gibbs sampling method for our model; Gibbs being a particular instance of the MCMC sampling strategy (see Section 7.2.3.2). It is of special interest when the calculation of any of the following becomes intractable:

- the joint posterior distribution,
- the marginal distribution of any subset of factors,
- the expected value of any of the factors.

Assuming that sampling from the full conditional posterior distribution is feasible, drawing a set of instances from this density converges to a sample from the joint posterior. If samples from the marginal distribution of a subset of factors are required, only the samples for that subset are kept; finally, the expected value of any factor can be computed by averaging over all its samples.

For our problem, we are interested in the second output; hence, we formulate the conditional density of **S**, which is proportional to a normal distribution multiplied by a normal prior. That is: $\mathcal{N}\left(x; \mu_p, \sigma_p^2\right) \propto \mathcal{N}\left(x; \mu, \sigma^2\right) \mathcal{N}\left(x; \mu_o, \sigma_o^2\right)$.

Let $\mathbf{A}_{\setminus(i,j)}$ represent all elements of $\mathbf{A}$ except $\mathbf{A}_{i,j}$; the full conditional density of $\mathbf{S}_{d,k}$ is:

$$p(\mathbf{S}_{d,k} \mid \mathbf{X}, \mathbf{S}_{\setminus(d,k)}, \mathbf{H}, \sigma^2) = \mathcal{N}\left(\mathbf{S}_{d,k}; \mu_p, \sigma_p^2\right), \tag{7.19}$$

where

$$\begin{aligned}\mu_p &= \sigma_p^2\left(\frac{\mu_o}{\sigma_o^2}+\frac{\sum_{n=1}^{N}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)\mathbf{H}_{k,n}}{\sigma^2}\right),\\ \sigma_p^2 &= \frac{\sigma^2\cdot\sigma_o^2}{\sigma^2+\sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}.\end{aligned}$$

Focusing on the mixing matrix, the full conditional density of $\mathbf{H}$ is proportional to a normal multiplied by an exponential, which turns out to be a rectified normal density of the form $\mathcal{R}\left(x;\mu_p,\sigma_p^2,\lambda_p\right)\propto\mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{E}\left(x;\lambda_o\right)$. That is:

$$p(\mathbf{H}_{k,n}\mid\mathbf{X},\mathbf{S},\mathbf{H}_{\setminus(k,n)},\sigma^2)=\mathcal{R}\left(\mathbf{H}_{k,n};\mu_p,\sigma_p^2,\lambda_p\right), \tag{7.20}$$

where

$$\mu_p=\frac{\sum_{d=1}^{D}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)\mathbf{S}_{d,k}}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2},$$
$$\sigma_p^2=\frac{\sigma^2}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2},\quad\lambda_p=\lambda_o.$$

Finally, the full conditional density of σ^2 is proportional to a normal multiplied by an inverse-gamma, denoted as $\mathcal{IG}\left(x;\alpha_p,\beta_p\right)\propto\mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{IG}\left(x;\alpha_o,\beta_o\right)$. Specifically:

$$p\left(\sigma^2\mid\mathbf{X},\mathbf{S},\mathbf{H}\right)=\mathcal{IG}\left(\sigma^2;\alpha_p,\beta_p\right), \tag{7.21}$$

where

$$\alpha_p=\frac{DN}{2}+\alpha_o,\quad\beta_p=\frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n}-(\mathbf{SH})_{d,n}\right]^2}{2}+\beta_o;$$

A detailed explanation on the derivations conducted to obtain the full conditional densities and their parameterisation can be found in Appendix C.

The resulting Gibbs sampler procedure for the *Bayesian SNMF* formulation is depicted in Algorithm 2.

7.3.2.2 Marginal likelihood for model selection

We have talked in Section 7.2.3.3 about the benefits of Bayesian model selection, which in our domain would translate into a strategy to appropriately

Algorithm 2 Bayesian SNMF Gibbs sampler

1) Normalise data $\mathbf{X}$ (L2-norm)
2) Randomly initialise $\mathbf{S}$, $\mathbf{H}$ and σ^2
3) For each sample $m \in \{1, \ldots, M\}$
   a) For each $d \in \{1, \ldots, D\}$ and $k \in \{1, \ldots, K\}$:
      i) Sample $\mathbf{S}_{d,k}$ according to Eq. 7.19
   b) For each $k \in \{1, \ldots, K\}$ and $n \in \{1, \ldots, N\}$:
      i) Sample $\mathbf{H}_{k,n}$ according to Eq. 7.20
   c) Sample σ^2 according to Eq. 7.21
   d) Store $\mathbf{S}^{(m)} = \mathbf{S}; \mathbf{H}^{(m)} = \mathbf{H}; \sigma^{2(m)} = \sigma^2$
4) Return $\{\mathbf{S}^{(m)}, \mathbf{H}^{(m)}, \sigma^{2(m)}\}_{m=1}^{M}$

assess the number of tissues sources over which the matrix factorisation should be performed. However, we have also mentioned the difficulty to calculate the marginal likelihood due to an intractable integral. In this section, we use the Chib's method [170] to estimate the marginal likelihood by using only posterior draws provided by the Gibbs sampler.

Recall the SNMF joint posterior expressed in Eq. 7.17, from which the marginal likelihood can be isolated:

$$p(\mathbf{X}) = \frac{p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right) \cdot p(\mathbf{S} \mid \boldsymbol{\theta}_S) \cdot p(\mathbf{H} \mid \boldsymbol{\theta}_H) \cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right)}{p(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X})}. \tag{7.22}$$

Computing the above equation for any value $\boldsymbol{\Phi}$ will result to a specific evaluation of the marginal likelihood at the point $\boldsymbol{\Phi}$ (selected to be a high density point for the most accurate estimation). Comparison among models (e.g., each using different number of sources) will be performed by comparing their marginal likelihood estimates at $\boldsymbol{\Phi}$: $p(\mathbf{X} \mid \boldsymbol{\Phi})$.

Obtaining the density at $\boldsymbol{\Phi}$ for any of the factors in the numerator is straight forward. The problem arises when calculating it in the denominator. The Chib's method solves it by segmenting the parameters in the denominator into B blocks, and applying the chain rule to write the denominator as the product of B terms. That is:

$$p(\boldsymbol{\Phi} \mid \mathbf{X}) = p(\boldsymbol{\Phi}_1 \mid \mathbf{X}) \times p(\boldsymbol{\Phi}_2 \mid \boldsymbol{\Phi}_1, \mathbf{X}) \times \ldots \times p(\boldsymbol{\Phi}_B \mid \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{B-1}, \mathbf{X}). \tag{7.23}$$

The blocks of parameters are appropriately chosen to be amenable to Gibbs sampling, such that each term is approximated by averaging over the con-

ditional density:

$$p\left(\mathbf{\Phi}_b \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{X}\right) \approx \frac{1}{M} \sum_{m=1}^{M} p\left(\mathbf{\Phi}_b \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{\Phi}_{b+1}^{(m)}, \ldots, \mathbf{\Phi}_B^{(m)}, \mathbf{X}\right),$$

where $\left\{\mathbf{\Phi}_{b+1}^{(m)}, \ldots, \mathbf{\Phi}_B^{(m)}\right\}$ are Gibbs samples from

$$p\left(\mathbf{\Phi}_{b+1}, \ldots, \mathbf{\Phi}_B \mid \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_{b-1}, \mathbf{X}\right),$$

and M the number of samples.

In our setting, each column of $\mathbf{S}$, each row of $\mathbf{H}$ and σ^2 are selected to be the blocks in Eq. 7.23. Therefore, given that $\mathbf{A}^*$ represents a matrix of high density points, $\mathbf{A}_{:,i}$ corresponds to all the values in the i-th column and $\mathbf{A}_{j,:}$ all the values in the j-th row:

$$\begin{aligned} p\left(\mathbf{S}^*, \mathbf{H}^*, {\sigma^2}^* \mid \mathbf{X}\right) &= p\left(\mathbf{S}_{:,1}^* \mid \mathbf{X}\right) \times p\left(\mathbf{S}_{:,2}^* \mid \mathbf{S}_{:,1}^*, \mathbf{X}\right) \times \ldots \times \\ &\times p\left(\mathbf{S}_{:,K}^* \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K-1}^*, \mathbf{X}\right) \times \\ &\times p\left(\mathbf{H}_{1,:}^* \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K}^*, \mathbf{X}\right) \times \ldots \times \\ &\times p\left(\sigma^{2^*} \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K}^*, \mathbf{H}_{1,:}^*, \ldots, \mathbf{H}_{K,:}^*, \mathbf{X}\right). \end{aligned} \tag{7.24}$$

Notice that all the above rationale still holds and computations are simplified if we apply the calculations in the logarithmic scale. Hence, Eq. 7.22 becomes

$$\begin{aligned} \log\left\{p\left(\mathbf{X}\right)\right\} &= \log\left\{p\left(\mathbf{X} \mid \mathbf{S}, \mathbf{H}, \sigma^2\right)\right\} + \log\left\{p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right)\right\} + \log\left\{p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right)\right\} \\ &+ \log\left\{p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right)\right\} - \log\left\{p\left(\mathbf{S}, \mathbf{H}, \sigma^2 \mid \mathbf{X}\right)\right\}. \end{aligned}$$

Similarly, Eq. 7.24 is now

$$\begin{aligned} \log\left\{p\left(\mathbf{S}^*, \mathbf{H}^*, {\sigma^2}^* \mid \mathbf{X}\right)\right\} &= \log\left\{p\left(\mathbf{S}_{:,1}^* \mid \mathbf{X}\right)\right\} + \log\left\{p\left(\mathbf{S}_{:,2}^* \mid \mathbf{S}_{:,1}^*, \mathbf{X}\right)\right\} + \ldots + \\ &+ \log\left\{p\left(\mathbf{S}_{:,K}^* \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K-1}^*, \mathbf{X}\right)\right\} + \\ &+ \log\left\{p\left(\mathbf{H}_{1,:}^* \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K}^*, \mathbf{X}\right)\right\} + \ldots + \\ &+ \log\left\{p\left(\sigma^2 \mid \mathbf{S}_{:,1}^*, \ldots, \mathbf{S}_{:,K}^*, \mathbf{H}_{1,:}^*, \ldots, \mathbf{H}_{K,:}^*, \mathbf{X}\right)\right\}. \end{aligned}$$

In order to compute the Bayes Factor between two models, namely $\mathcal{M}_i$ and $\mathcal{M}_j$, each one of them set to obtain a different number of sources K, we proceed to evaluate the marginal likelihood at $\left\{\mathbf{S}^*, \mathbf{H}^*, {\sigma^2}^*\right\}$ for both models, and compare them as follows:

$$\hat{B}_{ij} = \exp\{\log \hat{p}\left(\mathbf{X} \mid \mathcal{M}_i\right) - \log \hat{p}\left(\mathbf{X} \mid \mathcal{M}_j\right)\}.$$

This Bayes Factor allows us to select the most adequate model out of a pool of models, the difference among them being the number of sources employed to build it.

Matlab code of the proposed algorithms can be downloaded from `http://www.cs.upc.edu/~avilamala/resources/BayesianSNMF_Toolbox.zip`

7.3.2.3 Empirical evaluation

The suitability of the proposed method will be validated in the current section by a qualitative study on real SV-^{1}H-MRS data. In particular, we will use Chib's method to estimate the most appropriate number of underlying sources, the composition of which generates each of the observed instances within a tumour type, in a principled way. Secondly, each of these sources will be individually retrieved and analysed. A confidence measure on the proposed signals will also be supplied by providing a 90% interval around the signal.

Experimental setup For this study we again use data from the online-accessible and curated INTERPRET repository (Section 2.3). In particular, the 195 most clinically relevant spectral frequencies of the SV-^{1}H-MRS instances are selected for each of the 15 *nom*, 78 *gbm*, 31 *met* and 20 *ac*2 spectra acquired at LTE. Data acquired at STE have not been reported for this evaluation, given that their results did no provide much qualitative difference with respect to LTE data.

Given that all data points were normalised (L2-norm) prior to any treatment, the parameters for the prior distributions were chosen to match the amplitude of the data. These include $\mu_o = 0.01$ and $\sigma_o^2 = 0.2$ to limit the values of the sources $\mathbf{S}_{d,k}$ between -1 and 1 with $p > 0.95$; setting the $\lambda_o = 3$ to bound the values of the mixing matrix $\mathbf{H}_{k,n}$ to the $[0, 1]$ interval ($p > 0.95$); and $\alpha_o = 1; \beta_o = 0.001$ as flat priors for the noise variance σ^2. Moreover, the number of samples M generated at each Gibbs sampler run was set to 100,000; the first 50,000 were discarded to allow *burn-in*.

Table 7.3: Logarithm of the marginal likelihood ($\times 10^3$) according to the number of sources for each tumour type at LTE

	1	2	3	4	5
nom	**4.55**	3.82	2.94	2.70	1.95
gbm	24.31	**26.56**	25.89	26.32	26.40
met	**9.71**	8.68	8.67	8.62	8.58
*ac*2	6.49	**6.60**	6.15	5.73	5.44

Results As can be seen in Table 7.3, the values of the marginal likelihood for different number of sources obtained by the Chib's method clearly favour the models presenting low complexity; that is, those ones employing either one or two sources. This is a clear example of the Ockham's razor at work (Section 3.5). Notice that this estimate of the *best* number of sources to represent the observed instance from a source extraction point of view, might not necessarily be the most adequate for interpretability purposes. This will become clear in the following lines. Note also that the choice of best number of sources does not preclude other choices, given that the marginal likelihood provides a real-valued measure, not a binary one; in other words, it is a relative measure of relevance.

Let's focus our attention to Figure 7.2: the first column shows the average spectrum of each tumour type in our dataset; clearly showing the existent high intra-class variability, which is represented in the figure as a shadow zone. The number of sources chosen to decompose each tumour type follows the advise provided by the marginal likelihood. In this respect, the first row, corresponding to the normal tissue, can be represented by a single pure source (Figure 7.2b), where the characteristic peaks of N-Acetyl Aspartate (2.0ppm), Choline (3.2ppm) and Creatine (3.0ppm) are appropriately captured; the Glutamine and Glutamate are also retrieved at 2.05 - 2.46ppm.

The second row shows the decomposition of *gbm* into two signals: Figure 7.2d clearly identifies a reduction in the N-Acetyl Aspartate peak, as compared to the normal tissue; this is a clear sign of tumour proliferative

tissue. Similarly, the Creatine and Choline metabolites are also identified, the concentration of the latter being highly increased, showing the malignancy of the tumour type being analysed. Interestingly, there is an inverted peak at 1.3*ppm*, corresponding to Lactate, a compound frequently seen in high-grade malignant tumours. The second retrieved source (Figure 7.2e) nicely complements the first one by capturing the mobile lipids at 1.3 and 0.9*ppm*, a compound often indicating necrosis and hypoxia.

The third row deals with the analysis of *met*, presenting a single source to represent the tumour type (Figure 7.2g). Such a simple model, despite being a good candidate for data reconstruction, it is a very poor model in terms of interpretability: it basically reflects the shape of the average *met* spectrum, clearly capturing Choline, Creatine and mobile lipid metabolites, emphasizing their uncertainty about their amplitude.

The *ac*2 tumour type is represented in the last row of the figure by means of two sources: the first one (Figure 7.2i) identifies Choline and Creatine peaks, the ratio among them being lower than in the case of high-grade tumours; while the second (Figure 7.2j) captures the Lactate inverse peak as well as the not-well-known signal at the left-end of the spectrum.

As we have stressed throughout the thesis, the interpretability of the obtained results is at least as important as the quantitative suitability of the results themselves. In this respect, it is clear that marginal likelihood should be just part of the heuristic to determine the number of sources to extract if priority is given to interpretability. In a second experiment, we thus decided to extract three sources for each of the tumour types being analysed: *gbm*, *met* and *ac*2, disregarding the marginal likelihood *recommendation*. The obtained sources can be seen in Figure 7.3, and they are to be compared with Figure 7.2.

In the case of *gbm* tumour types, there is no major improvement regarding interpretability on the process of moving from two to three sources: the signal in Figure 7.2e is perfectly conserved in Figure 7.3a; while the source in Figure 7.2d is respected in Figure 7.3b. The new signal in Figure 7.3c can be considered as a mostly negative noise, which is extracted out of the two real generating signals; however, it is of little help in terms of interpretability.

A more interesting result can be found in the *met* case: in this experiment, passing from one to three sources implies a decomposition of a signal barely capturing the average tumour type (Figure 7.2g) into three meaningful sources: Figure 7.3d representing the mobile lipids contribution, Figure 7.3e retrieving the Lactate compound as a negative peak and Figure 7.3f with the Choline and Creatine peaks clearly overlooking the signal. This is a clear example of the divergence between the results aiming at reconstructing the instances out of a set of sources and interpreting such sources. It is also an example of an extracted negative source that could only have been captured by a method that allows negative-valued sources.

Finally, *ac*2 tumour type slightly benefits from adding a third source to the decomposition in terms of interpretability: the first source in Figure 7.2i remains in Figure 7.3g, while the signal in Figure 7.2j is mostly replicated in Figure 7.3i with the exception of the magnified Lactate inverse peak and, to some extent, part of the Choline and Creatine contributions, which are expressed in Figure 7.3h.

The obtained results exemplify how the proposed method currently discussed is a powerful tool for extracting the different types of tissue conforming each tumour type, being especially relevant for knowledge discovery tasks.

7.3.3 Discussion

The two methods derived in this chapter, namely *Probabilistic CNMF* via MAP and full *Bayesian SNMF* using Gibbs sampling, stem from two probabilistic frameworks that allow unsupervised decomposition of real-valued observations into a matrix of real-valued sources and a non-negative mixing matrix. The first matrix contains basic self-explanatory signals and the second one corresponds to the additive contribution of each source to conform every observation. Both techniques benefit from some of the properties provided by the probabilistic paradigm, such as automatic control of regularisation to avoid overfitting and the incorporation of prior information to compensate for some limitations due to small sample sizes.

Nonetheless, given the different formulation and resolution strategies they present, the applicability of each one of them is pretty different: fast *Probabilistic CNMF* is very useful in binary discriminative settings, where encountered sources correspond to *tumour-type* representatives and the mixing matrix unavoidably expresses the degree of tumour type mixture in each of the measured voxels. This phenomenon is a direct consequence of the convex formulation in the objective function.

Conversely, the more time-consuming full *Bayesian SNMF* is especially suited to retrieve the existent *tissue-type* sources that are part of each of the tumour types. Its applicability could be of high value for nosologic images [87], where a colour map of the brain based on tissue delimitation is constructed. Full *Bayesian SNMF* comes with extra features, such as a strategy to determine the most suitable number of sources to represent the observed data, as well as an explicit quantification of estimation uncertainty in the form of a credible interval bounding the retrieved signals, which is highly relevant for domains where only a small number of samples is available.

7.4 Conclusions

The derived *Probabilistic CNMF* via MAP and the full *Bayesian SNMF* using Gibbs sampling are two different unsupervised approaches to decompose a set of observations into a matrix of generated signals and a matrix representing the composition of each signal to conform the observed instances. A comparison and contrast of the two techniques have been carried out in the previous section. Their applicability to the analysis in neuro-oncology by means of SV-^{1}H-MRS data, where varying generating sources contribute to the signal retrieved by the scanner, has proven successful. Now, we revisit the explicit technical requirements that motivated the development of such techniques, as expressed in Section 7.1, together with the preconditions shared by all source separation strategies in our domain (i.e., enumerated in Section 6.1):

1. *It must be able to identify the underlying sources present in the retrieved signal*: each column in matrix **S** contains an estimated source.

2. *It needs to assess the contribution of each source to the signal*: matrix **H** contains the positive coefficients representing the source contribution to every instance.

3. *Both the sources and their contributions must be easily interpretable*: source matrix **S** is provided in the vector unit length, while the instance coefficients in **H** are enforced to be in the range $(0, 1)$, so that outcomes become easily interpretable.

4. *The solution must naturally deal with both negative and positive values*: this is addressed by imposing sources to be a convex combination of instances or understanding their values as random variables sampled from a normal distribution.

5. *Ratios between values of metabolites at certain frequencies must be preserved*: this is the rationale behind specifically dealing with negative values instead of shifting the whole spectrum.

6. *Distances between values of metabolites at specific frequencies must be kept*: the same reasoning as in previous statement applies here.

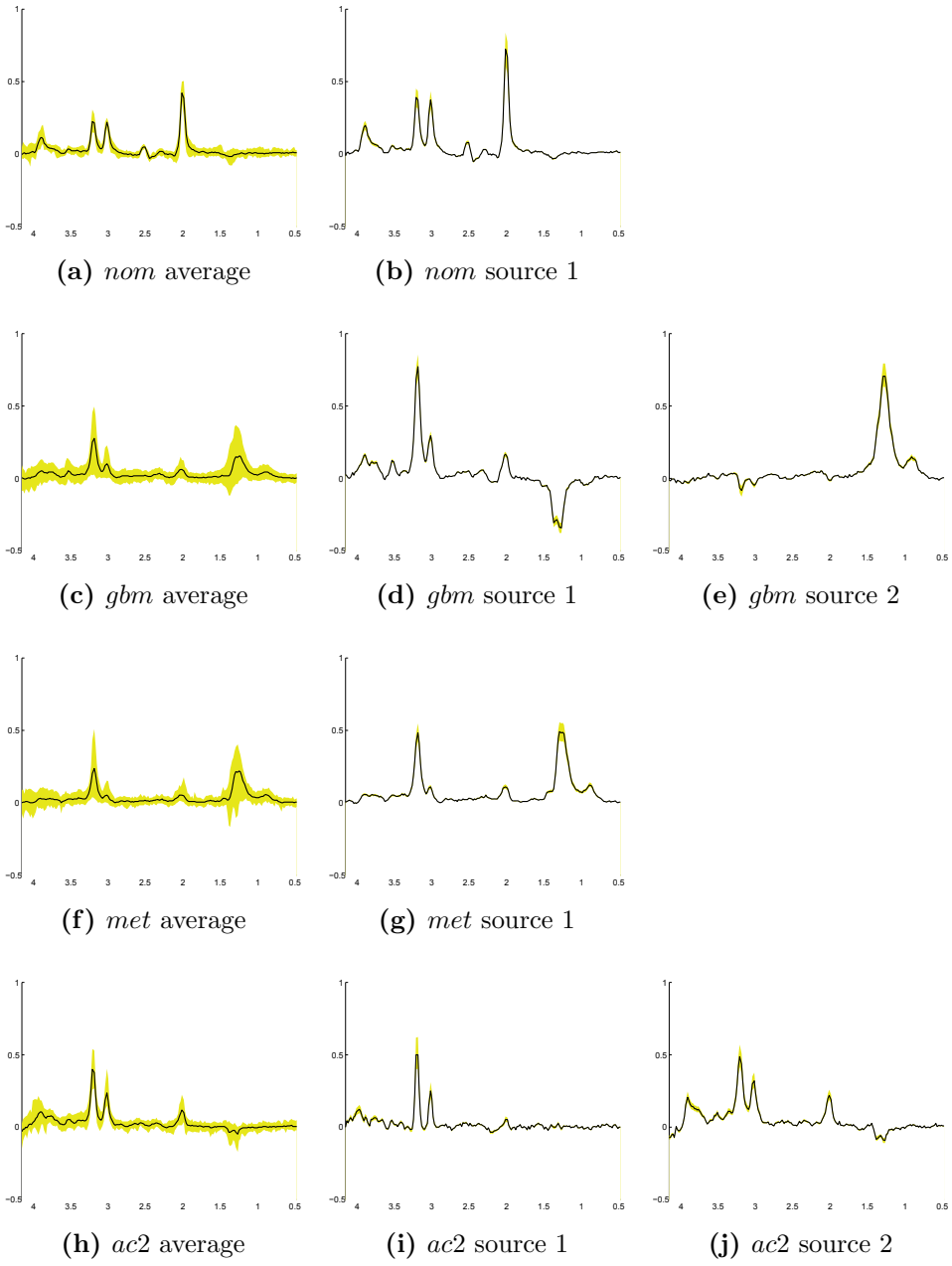
7. *The possibility to incorporate prior knowledge on sources and their contributions*: this knowledge is captured by the prior distributions.

8. *Automatic control of regularisation hyperparameters*: overfitting avoidance is ensured through the regularisation provided by the prior distributions.

9. *Appropriately handle uncertainty and provide an interpretable measure of confidence for the retrieved sources*: in the full Bayesian approach, each derived source comes with a credible interval as a byproduct of Gibbs sampling.

10. *Suitable selection of the most appropriate number of underlying sources*: Chib's method to easily estimate the marginal likelihood from Gibbs draws has been derived for the proposed model. Marginal likelihood can be directly used to either select the model employing the most

adequate number of reconstructing sources, or rank models according to such criterion.



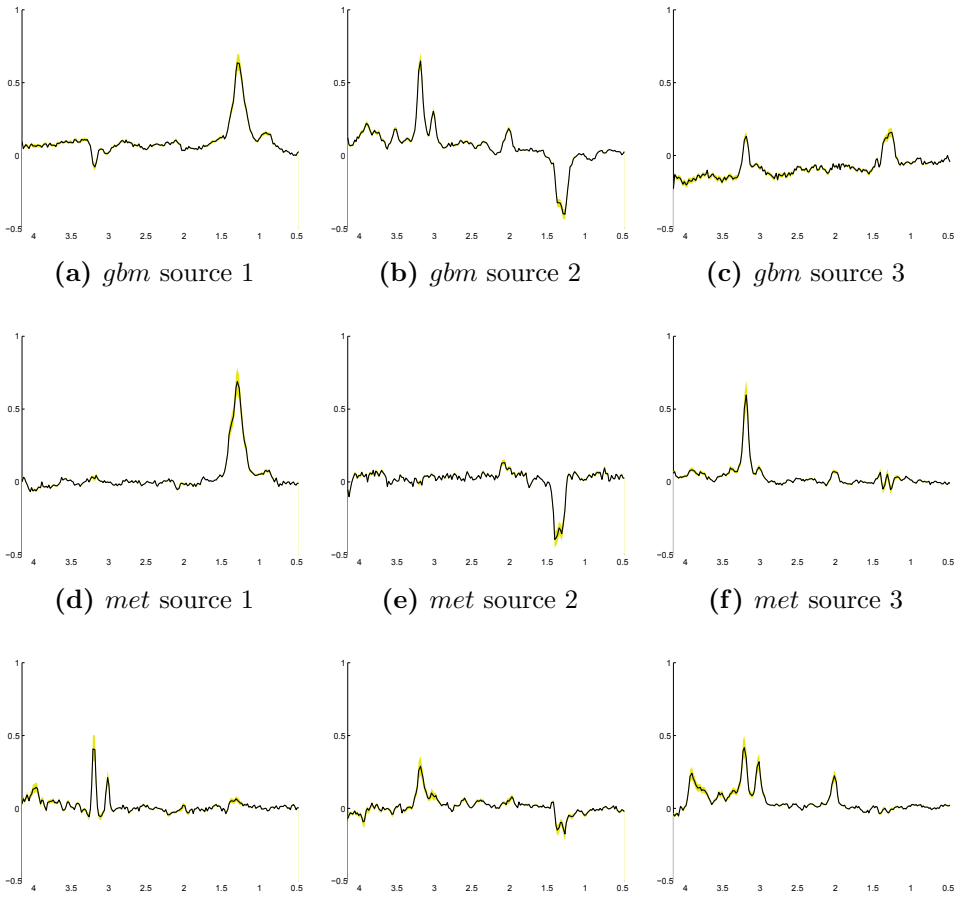
(a) *nom* average **(b)** *nom* source 1

(c) *gbm* average **(d)** *gbm* source 1 **(e)** *gbm* source 2

(f) *met* average **(g)** *met* source 1

(h) *ac*2 average **(i)** *ac*2 source 1 **(j)** *ac*2 source 2

Figure 7.2: Sources identified by Bayesian SNMF after model selection using data acquired at LTE - Each row corresponds to a single tumour type: the first column being the average spectrum, and the other ones the retrieved sources from our method. The black solid line represents the mean, while the shadowed region conforms the 90% credible interval. Y-axes represent unit-free metabolite concentrations and X-axes represent frequency as measured in parts per million (ppm).



(a) *gbm* source 1 **(b)** *gbm* source 2 **(c)** *gbm* source 3

(d) *met* source 1 **(e)** *met* source 2 **(f)** *met* source 3

(g) *ac*2 source 1 **(h)** *ac*2 source 2 **(i)** *ac*2 source 3

Figure 7.3: Three-source decomposition of LTE SV-^{1}H-MRS according to Bayesian SNMF - Each row corresponds to a single tumour type; each column presents one out of the three retrieved sources from our method. The black solid line represents the mean, while the shadowed region conforms the 90% credible interval. Y-axes represent unit-free metabolite concentrations and X-axes represent frequency as measured in parts per million (ppm).

Chapter 8

Conclusions and future work

8.1 Summary

Brain cancer is an extremely disturbing condition due to the damage it can cause to the affected organ as well as the poor prognosis that certain types of this pathology present. An early and accurate diagnosis is crucial to improve the quality of life of the patients and increase survival rates. Current state of the art techniques for obtaining a rigorous diagnostic outcome involve the utilisation of invasive techniques, biopsy being the gold standard.

The risk associated to resorting to this kind of procedures has increased the awareness of the need to find alternative strategies that are able to provide indirect measurements for diagnostic purposes, causing little or even no damage to the patient. In this respect, NMR has become the leading non-invasive measurement technique in clinical practise. MRI is a suitable tool for general tumour location, but it lacks definition and does not help to distinguish between metastatic tumours and those which have their origin in the own brain tissue. Its spectroscopy-based counterpart, MRS, though, can help to disambiguate uncertain cases due to its metabolic profiling capabilities. Together, they are able to provide fine-resolution measurements of biochemical compositions within a delimited area.

Nonetheless, the often complex and difficult to interpret output that MRS systems generate hinders their practical implementation in daily medical practise: a major shortcoming that has of late being tackled with the

aid of statistical and artificial intelligence-based solutions.

In spite of the many milestones recently achieved in the field, there is still room for improvement, for instance in discrimination among the most aggressive tumours, or in the determination and influence that distinct tissues have in the vicinity of most common tumoural areas.

The hypothesis that specific biomarkers match particular frequencies in the MR spectrum motivates the use of advanced feature selection techniques to tackle the aforementioned problems, which can be coupled with ensemble methods for the sake of obtaining models achieving the high degree of specialisation required to capture the patterns of relevance shown by different tumour types. Moreover, the mixture of signals retrieved in an MR measurement encourages the use of source separation approaches both supervised and unsupervised to not only determine, but also quantify the number of distinguishable tissues contributing to the measurement.

Breadth Ensemble Learning has been designed in Chapter 4 to improve the discrimination of aggressive tumours; in turn, *Recursive Logistic Instance Weighting* has been developed in Chapter 5 to increase the stability of feature selection algorithms when faced with this same problem. In Chapter 6, the *Discriminant Convex Non-negative Matrix Factorisation* procedure has been derived to determine tissue-type representatives and estimate their proportion in the most common tumoural areas. Finally, Chapter 7 contributes to the field by providing probabilistic versions of *Convex* and *Semi Non-negative Matrix Factorisation* strategies to better estimate the correct number of tissues present in the analysed sample and handle uncertainty in a principled manner.

8.2 Conclusions

In the following, we present the main conclusions of this thesis:

- The difficult problem consisting in properly classifying heterogeneous SV-^{1}H-MRS data as belonging to either the *glioblastoma* or *metastasis* families of tumours can successfully be addressed using an ensemble learning technique that is built in breadth, aiming at improving the

overall ensemble discriminative capability: see Table 4.2 for comparative results.

- A key element for its success entails a wise subdivision of the input space that feeds each base learner, by projecting the data to a lower dimensionality feature space that greedily best increases the ensemble performance (wrapper-like), given that random feature selection has been shown to be suboptimal: Table 4.2 supports this conclusion by comparing our strategy against Bagging, Boosting (random selection of instances) and Random Forests (random selection of instances and features).

- A second important point to consider concerns the use of strong base learners (i.e., LDA), since weak learners (the usual choice in ensemble settings) perform unacceptably in this domain: this statement is supported by Table 4.1, where best results were obtained by strong LDA and LS-SVM as compared to weak NB and CART.

- Reliability in the eyes of domain experts can be increased by consistently providing a similar set of relevant biomarkers over different runs of the algorithm through the use of stabilising FSS strategies. A module prior to FSS able to rate instances according to their typicality, coupled to a modification of traditional FSS techniques to deal with these instance-rates can accomplish our goal: Figure 5.7 supports this, as it reveals that our pre-processing method outperforms traditional RelievedF-RFE in terms of stability in almost all iterations.

- Moreover, unstable FSS algorithms (e.g., those of the Relief family) are the ones most benefiting from stability improvement strategies as the one presented in the thesis: in that sense, we agree with [75], who analysed the stability of SVM-RFE.

- The final remark on this topic is that stability might come at the price of accuracy loss, and in most situations we are facing a trade-off

between stability and accuracy: this can be appreciated when contrasting Figure 5.6 and Table 5.2, especially for Breast and Parkinson datasets, where this phenomenon is evident.

- Accurately identifying the interpretable latent sources representing biological tissues, of which the measured NMR signal is composed can be faithfully performed using either SNMF or CNMF, which are BSS techniques able to deal with a variety of constraints imposed by our domain data. The former technique is able to retrieve tissue specific sources (e.g., Figure 7.3), while the latter is more suitable to extract tumour-type specific signals, that best resemble the class averages (e.g., Figure 7.1).

- Retrieving tumour-type signals can be aided by including class-specific information to CNMF. This novel technique has proven to be highly valuable in analysing difficult problems, such as, for instance, the discrimination between high-grade *glioblastomas* and *metastases*: such statement is backed up by the results reported in Table 6.3, where all those problems whose extracted sources correlate less than 0.8 to the class mean spectrum are shown to be improved through the use of our method.

- The proposed method has the ability to reconstruct the data in the original data space, where each reconstructed data point contains more discriminative power than its original (observed) counterpart: this effect of data *cleaning* is shown in Figure 6.3.

- Formulating NMF variants from a probabilistic perspective adds a set of ingredients that improve the obtained results. Including prior information is of great help in our domain, where few data are available. Priors also play a role to avoid overfitting, by automatically controlling the regularisation parameter. These are incorporated in the *Probabilistic CNMF* technique that is able to balance between obtaining reliable tumour-type sources with acceptable accuracy capabilities, as can be appreciated in Table 7.1.

- Unfortunately, CNMF can not be formulated for a full Bayesian treatment, but only the constraint-relaxed SNMF can benefit from this formulation. Nonetheless, in the *Bayesian SNMF* technique, the obtained sources are coherent with tissue-specific signals and are sometimes very different from tumour-type averages (e.g., Figure 7.2). Furthermore, a credible interval is also provided to aid in the radiologists' decision making.

- Another advantage of *Bayesian SNMF* is the possibility to analytically determine the best minimum number of sources required to reconstruct the observed data (Table 7.3). However, we realised that the best number of sources to reconstruct the data might not necessarily agree with the number of sources to best interpret the tissues present in a voxel. This statement can be clearly appreciated by comparing Figures 7.2 and 7.3.

8.3 Open problems and potential extensions of this research

Throughout this thesis, we have answered many of the questions that were identified in Chapter 1. Nonetheless, research on the different topics has raised a set of new questions and future research lines that have not been addressed, due to either time or scope constraints. Here, we list some of them:

- The computational time burden of the proposed ensemble solution might become a limitation for its real usage. Therefore, an in-depth study on up-to-date strategies to effectively parallelise its computation, both when evaluating new feature candidates and during base learners training, should be carried out. A possible approach would entail the implementation of the algorithm under the Map-Reduce paradigm [171].

- Most of the literature on ensemble learning advises the use of unstable classifiers as base learners, since the achieved diversity contributes

to reducing variance, hence improving overall ensemble performance. However, we have empirically shown that our ensemble solution on the studied neuro-oncology domain performs best when stable classifiers are employed. We hypothesise that this phenomenon is observed because our solution reduces bias, which turns out to be very high due to the heterogeneity of data. A new line of research could include the derivation of a theoretical analysis on the bias-variance decomposition of the proposed ensemble solution to formally validate such hypothesis.

- Analysing the success of the proposed stability method for RelievedF-RFE, we hypothesise that FSS techniques based on *hypothesis-margin* (e.g., the Relief family of algorithms) are the most suitable when instances are highly heterogeneous (e.g., in *glioblastoma* vs. *metastasis* discrimination), given that they naturally cluster in different locations of the input space, creating neighbourhoods that can be better dealt with this type of margin. A study to validate this hypothesis might become a nice contribution to the field.

- We have developed a rating function for instances to stabilise feature selection strategies assuming that those instances have been generated from a multivariate Gaussian distribution, but this might be far from optimal. Future research should include the assessment of various distributions where the notion of *typicality* might be very different from the one employed here.

- So far we have used the rated instances to stabilise feature selection algorithms, but nothing prevents our technique to be used as a preprocessing step to stabilise learning algorithms for classification or regression purposes. One simple approach consists in using the already modified version of SVM to deal with instance weighting for classification; or modifying the cost function of other existing learning algorithms in a similar way.

- Regarding the derived supervised version of CNMF, a plausible step forward includes the automated estimation of the most adequate num-

ber of sources, which does not need to be coupled with the number of classes being discriminated. In this sense, training a classifier on the lower-dimensional mapping of instances (i.e., stored in $\mathbf{H}$) should be evaluated.

- Class-specific information in DCNMF has been captured by incorporating Fisher Linear Discriminants (a classification algorithm widely used in the domain) to the CNMF cost function, aiming at providing not only reliable underlying sources, but also the extent of contribution that each source applies in generating the measured data. Although identification of sources has largely been improved by means of discriminative knowledge, contribution of each source did not follow as expected. We could tackle this issue by proposing the modification of the cost function, where the scatter matrices are not calculated on the projected instances (i.e., on $\mathbf{H}$), but in the projection itself (i.e., $\mathbf{S^{-1}X}$), in order to influence the bases that span the mapping in the subspace. This idea was first proposed in [136].

- Sticking to the same technology (CNMF), we could evolve the definition of the cost function by replacing the Fisher Linear Discriminant term by the maximum sample margin as formalised in the linear SVM, in a similar fashion as [137] did for NMF.

- The qualitative study using *Bayesian SNMF* opens a promising path towards a new set of interpretable techniques for the analysis of brain tumours using ^{1}H-MRS data. Next steps in this domain will include a quantitative study of the technique in a diagnostic setting where the role of mixing matrix $\mathbf{H}$ will be actively evaluated.

- Another very important contribution to the *Bayesian SNMF* domain should consider the incorporation of class-specific knowledge to the formulation as prior information, similar to the DCNMF but from a probabilistic perspective.

- Finally, it is worth to mention that, in spite of the fact that all developed techniques had the brain tumour diagnosis from SV-^{1}H-MRS

data in mind, nothing prevents them to be used in other domains. Therefore, a nice extension of this thesis would be the adoption of the proposed algorithms to other fields where data match the features necessary to benefit from them.

8.4 List of publications

- Albert Vilamala, Lluís A. Belanche and Alfredo Vellido (2012). *Classifying malignant brain tumours from ^{1}H-MRS data using Breadth Ensemble Learning.* **International Joint Conference on Neural Networks (IJCNN)**. Brisbane, Australia, pp.2803-2810 .

- Albert Vilamala, Paulo J.G. Lisboa, Sandra Ortega-Martorell, Alfredo Vellido (2013). *Discriminant Convex Non-negative Matrix Factorisation for the classification of Human Brain Tumours.* **Pattern Recognition Letters**, 34(14), 1734-1747.

- Albert Vilamala, Lluís A. Belanche (2014). *Improving stability of Feature Selection for Brain Tumour Diagnosis using ^{1}H-MRS data.* 2^{nd} **International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2014)**. Granada, Spain, pp. 1254-1265.

- Albert Vilamala, Lluís A. Belanche, Alfredo Vellido (2014). *A MAP approach for Convex Non-negative Matrix Factorisation in the Diagnosis of Brain Tumors.* **2014 International Workshop on Pattern Recognition in Neuroimaging (PRNI 2014)**. Tübingen, Germany.

References

[1] World Health Organization. Cancer. Fact sheet N297. http://www.who.int/mediacentre/factsheets/fs297/en/, 2015. [Online; Accessed: March 2015]. 1

[2] Generalitat de Catalunya. Departament de salut. El cancer a Catalunya 1993-2020. http://cancer.gencat.cat, 2012. [Online; Accessed: March 2015]. 1

[3] J. Borràs, J. Borràs, R. Gispert, and A. Izquierdo. El impacto del cáncer en Cataluña. *Medicina Clínica*, **131**(1):2–3, 2008. 1

[4] Central Brain Tumor Registry of the United States. Fact sheet. http://www.cbtrus.org/factsheet/factsheet.html. [Online; Accessed: March 2015]. 2

[5] J. Luts. *Classification of Brain Tumors Based on Magnetic Resonance Spectroscopy*. PhD thesis, Katholieke Universiteit Leuven, Belgium, 2010. 3

[6] K. Tsuchiya, A. Fujikawa, M. Nakajima, and K. Honya. Differentiation between solitary brain metastasis and high-grade glioma by diffusion tensor imaging. *British Journal of Radiology*, **78**(930):533–537, 2005. 3

[7] W. Wang, C. Steward, and P. Desmond. Diffusion tensor imaging in glioblastoma multiforme and brain metastases: The role of p, q, l, and fractional anisotropy. *American Journal of Neuroradiology*, **30**(1):203–208, 2009. 3

[8] A. Server, R. Josefsen, B. Kulle, J. Mæ hlen, T. Schellhorn, O. Gadmar, T. Kumar, M. Haakonsen, C. Langberg, and P. H. Nakstad. Proton magnetic resonance spectroscopy in the distinction of high-grade cerebral gliomas from single metastatic brain tumors. *Acta Radiologica*, **51**(3):316–325, 2010. 3

[9] L. Blanchet, P. Krooshof, G. Postma, A. Idema, B. Goraj, A. Heerschap, and L. Buydens. Discrimination between metastasis and glioblastoma multiforme based on morphometric analysis of MR images. *American Journal of Neuroradiology*, **32**(1):67–73, 2011. 3

[10] A. Vellido, J. Martin-Guerrero, and P. Lisboa. Making machine learning models interpretable. In *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 163–172, 2012. 3

[11] P. J. G. Lisboa, A. Vellido, R. Tagliaferri, F. Napolitano, M. Ceccarelli, J. D. Martín-Guerrero, and E. Biganzoli. Application notes: data mining in cancer research. *Computational Intelligence Magazine*, **5**(1):14–18, 2010. 3, 52, 54

[12] ACGT Consortium. Advancing clinico genomic trials on cancer. http://acgt.ercim.eu/. [Online; Accessed: April 2015]. 4

[13] ASSIST Consortium. Assist. http://assist.ee.auth.gr. [Online; Accessed: April 2015]. 4

[14] Biopattern Network of Excellence. Computational intelligence for biopattern analysis in support of eHealthcare. http://www.biopattern.org. [Online; Accessed: April 2015]. 4

[15] MLPM Network. Machine learning for personalized medicine. http://www.mlpm.eu/. [Online; Accessed: April 2015]. 4

[16] Epigene Informatics. Machine learning approaches to epigenomic research. http://cordis.europa.eu/project/rcn/102798_en.html. [Online; Accessed: May 2015]. 4

[17] Metoxia consortium. Metastatic tumours facilitated by hypoxic tumour micro-environments. http://www.metoxia.uio.no. [Online; Accessed: April 2015]. 4

[18] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, and D. Watson. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. *Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA)*, **19**:22–33, 2006. 4, 28

[19] eTumour Consortium. eTumour: Web accessible MR decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data. http://ibime.webs.upv.es/?page_id=36. [Online; Accessed: April 2015]. 5, 29

[20] H. González-Vélez, M. Mier, M. Julià-Sapé, T. N. Arvanitis, J. M. García-Gómez, M. Robles, P. H. Lewis, S. Dasmahapatra, D. Dupplaw, A. Peet, C. Arús, B. Celda, S. Van Huffel, and M. Lluch-Ariet. Healthagents: Distributed multi-agent brain tumor diagnosis and prognosis. *Journal of Applied Intelligence*, **30**(3):191–202, 2009. 5

[21] F. Aligué. CAD para la detección precoz del cáncer de mama (ii). *Mundo Electrónico*, (406):42–47, 2009. 5

[22] A. Oliver. *Automatic mass segmentation in mammographic images*. PhD thesis, Universitat de Girona, 2007. 5

[23] F. F. González-Navarro. *Feature Selection in Cancer Research: Microarray Gene Expression and in vivo ^{1}H-MRS Domains*. PhD thesis, Universitat Politècnica de Catalunya, 2011. 5, 6, 70

[24] C. J. Arizmendi. *Signal Processing Techniques for Brain Tumour Diagnosis from Magnetic Resonance Spectroscopy Data*. PhD thesis, Universitat Politècnica de Catalunya, 2012. 6, 29

[25] S. Ortega-Martorell. *On the Use of Advanced Pattern Recognition Techniques for the Analysis of MRS and MRSI Data in Neuro-oncology*. PhD thesis, Universitat Autònoma de Barcelona, 2012. 6, 24, 99, 107, 109, 117

[26] A. Pérez-Ruiz, M. Julià-Sapé, G. Mercadal, I. Olier, C. Majós, and C. Arús. The INTERPRET decision-support system version 3.0 for evaluation of magnetic resonance spectroscopy data from human brain tumours and other abnormal brain masses. *BMC Bioinformatics*, **11**(1):581, 2010. 6

[27] D. Louis, H. Ohgaki, O. Wiestler, W. Cavenee, P. Burger, A. Jouvet, B. Scheithauer, and P. Kleihues. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, **114**(2):97–109, 2007. 16, 18

[28] National Cancer Institute. What you need to know about brain tumors. http://www.cancer.gov/cancertopics/wyntk/brain. [Online; Accessed: June 2014]. 17, 18

[29] Q. T. Ostrom, H. Gittleman, P. Farah, A. Ondracek, Y. Chen, Y. Wolinsky, N. E. Stroup, C. Kruchko, and J. S. Barnholtz-Sloan. CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010. *Neuro-Oncology*, **15**(suppl 2):ii1–ii56, 2013. 18, 19, 20

[30] M. C. Preul, Z. Caramanos, R. Leblanc, J. G. Villemure, and D. L. Arnold. Using pattern analysis of in vivo proton MRSI data to improve the diagnosis and surgical management of patients with brain tumors. *NMR in Biomedicine*, **11**(4-5):192–200, 1998. 21

[31] A. Boss, S. Bisdas, A. Kolb, M. Hofmann, U. Ernemann, C. D. Claussen, C. Pfannenberg, B. J. Pichler, M. Reimold, and L. Stegger. Hybrid PET/MRI of intracranial masses: Initial experiences and

comparison to PET/CT. *Journal of Nuclear Medicine*, **51**(8):1198–1205, 2010. 22

[32] V. Govindaraju, K. Young, and A. A. Maudsley. Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR in Biomedicine*, **13**(3):129–153, 2000. 25, 66, 91

[33] N. P. Davies, M. Wilson, K. Natarajan, Y. Sun, L. MacPherson, M.-A. Brundler, T. N. Arvanitis, R. G. Grundy, and A. C. Peet. Non-invasive detection of glycine as a biomarker of malignancy in childhood brain tumours using in-vivo^{1}H-MRS at 1.5 tesla confirmed by ex-vivo high-resolution magic-angle spinning NMR. *NMR in Biomedicine*, **23**(1):80–87, 2010. 27

[34] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(12):6745–6750, 1999. 30

[35] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**:531–537, 1999. 30

[36] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**(2):203–209, 2002. 30

[37] G. J. Gordon, R. V. Jensen, L. li Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62**:4963–4967, 2002. 30

[38] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871):530–536, 2002. 30

[39] D. Talantov, A. Mazumder, J. X. Yu, T. Briggs, Y. Jiang, J. Backus, D. Atkins, and Y. Wang. Novel genes associated with malig-

nant melanoma but not benign melanocytic lesions. *Clinical Cancer Research*, **11**(20):7234–7242, 2005. 30

[40] C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance, L. R. Sudarsky, D. G. Standaert, J. H. Growdon, R. V. Jensen, and S. R. Gullans. Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(3):955–960, 2007. 30

[41] Y. Lai, B. Wu, L. Chen, and H. Zhao. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, **20**(17):3146–3155, 2004. 30

[42] Y. Han and L. Yu. A Variance Reduction Framework for Stable Feature Selection. *Statistical Analysis and Data Mining*, **5**:428–445, 2012. 31, 49, 89, 90, 92, 93

[43] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2nd edition, 2010. 34, 35, 69

[44] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, **31**(8):651–666, 2010. 35

[45] T. Kohonen. *Self-Organizing Maps*. Springer, 3 edition, 2000. 35

[46] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1):29–36, 1982. 37

[47] J. M. Lobo, A. Jiménez-Valverde, and R. Real. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**(2):145–151, 2008. 37

[48] D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, **77**(1):103–123, 2009. 37

[49] C. Barber, D. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hull. *ACM Transaction on Mathematical Software*, **22**(4):469–483, 1996. 37

[50] R. Kohavi and D. H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283. Morgan Kaufmann, 1996. 38

[51] G. Seni and J. Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*, **2**. Morgan and Claypool Publishers, 2010. 40

[52] L. Breiman. Bagging predictors. *Machine Learning*, **24**(2):123–140, 1996. 40

[53] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pages 148–156, 1996. 41, 45

[54] T. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(8):832 –844, 1998. 41, 60

[55] L. Breiman. Random forests. *Machine Learning*, **45**(1):5–32, 2001. 42

[56] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 42, 62

[57] P. Langley. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall Symposium on Relevance*, pages 140–144, New Orleans, LA, USA, 1994. AAAI Press. 43

[58] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994. 44, 45, 64, 79

[59] M. Ben-Bassat. *Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation*, **2**, pages 773–791. North Holland, 1982. 45

[60] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1999. 45

[61] K. Kira and L. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. 45, 79

[62] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1-2):273–324, 1997. 45, 46, 79

[63] S. D. Stearns. On selecting features for pattern classifiers. In *Proceedings of the 3rd International Conference on Pattern Recognition (ICPR 1976)*, pages 71–75, Coronado, CA, 1976. 45

[64] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, **15**(11):1119–1125, 1994. 45

[65] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2):121–167, 1998. 45

[66] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth statistics/probability series. Wadsworth International Group, 1984. 45, 62, 67

[67] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(6):559–572, 1901. 46

[68] C. Jutten and J. Herault. Blind separation of sources, part 1: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24**(1):1–10, 1991. 47

[69] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**(2):111–126, 1994. 47, 102, 139

[70] P. Turney. Technical note: Bias and the quantification of stability. *Machine Learning*, **20**(1–2):23–33, 1995. 47

[71] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, **2**:499–526, 2002. 48

[72] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 218–225, 2005. 49, 81

[73] L. I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, AIAP'07, pages 390–395. ACTA Press, 2007. 49, 81

[74] P. Somol and J. Novovičová. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(11):1921–1939, 2010. 49, 82

[75] Y. Saeys, T. Abeel, and Y. Peer. Robust feature selection using ensemble feature selection techniques. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 313–325, Berlin, Heidelberg, 2008. Springer-Verlag. 49, 83, 94, 167

[76] S. Loscalzo, L. Yu, and C. Ding. Consensus group stable feature selection. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 567–576, New York, NY, USA, 2009. ACM. 49, 84

[77] M. Tipping. Bayesian inference: An introduction to principles and practice in machine learning. In O. Bousquet, U. von Luxburg, and G. Ratsch, editors, *Advanced Lectures on Machine Learning*, **3176** of *Lecture Notes in Computer Science*, pages 41–62. Springer Berlin Heidelberg, 2004. 50

[78] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, pages 773–795, 1995. 51

[79] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. 51, 137, 138

[80] W. Negendank. Studies of human tumors by MRS: A review. *NMR in Biomedicine*, **5**(5):303–324, 1992. 52

[81] M. C. Preul, Z. Caramanos, D. L. Collins, J. G. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, and D. L. Arnold. Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy. *Nature Medicine*, **2**(3):323–325, 1996. 52

[82] G. Hagberg. From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods. *NMR in Biomedicine*, **11**(4-5):148–156, 1998. 52, 109

[83] P. J. G. Lisboa, S. P. J. Kirby, A. Vellido, Y. Y. B. Lee, and W. El-Deredy. Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection. *NMR in Biomedicine*, **11**(4-5):225–234, 1998. 53

[84] W. Hollingworth, L. Medina, R. Lenkinski, D. Shibata, B. Bernal, D. Zurakowski, B. Comstock, and J. Jarvik. A systematic literature review of magnetic resonance spectroscopy for the characterization of brain tumors. *American Journal of Neuroradiology*, **27**(7):1404–1411, 2006. 53

[85] P. Lisboa and A. F. G. Taktak. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, **19**(4):408–415, 2006. 53

[86] N. Sibtain, F. Howe, and D. Saunders. The clinical value of proton magnetic resonance spectroscopy in adult brain tumours. *Clinical Radiology*, **62**(2):109 – 119, 2007. 53

[87] F. De Edelenyi, C. Rubin, F. Estève, S. Grand, M. Décorps, V. Lefournier, J. F. Le Bas, and C. Rémy. A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images. *Nature Medicine*, **6**:1287–1289, 2000. 53, 159

[88] A. R. Tate, C. Majós, A. Moreno, F. Howe, J. Griffiths, and C. Arús. Automated classification of short echo time in in vivo ^{1}H brain tumor spectra: A multicenter study. *Magnetic Resonance in Medicine*, **49**(1):29–36, 2003. 53, 109

[89] K. S. Opstad, C. Ladroue, B. A. Bell, J. R. Griffiths, and F. A. Howe. Linear discriminant analysis of brain tumour ^{1}H-MR spectra: a comparison of classification using whole spectra versus metabolite quantification. *NMR in Biomedicine*, **20**(8):763–770, 2007. 53

[90] S. W. Provencher. Automatic quantitation of localized in vivo ^{1}H spectra with LCModel. *NMR in Biomedicine*, **14**(4):260–264, 2001. 53

[91] K. Opstad, M. Murphy, P. Wilkins, B. A. Bell, J. Griffiths, and F. Howe. Differentiation of metastases from high-grade gliomas using short echo time ^{1}H spectroscopy. *Journal of Magnetic Resonance Imaging*, **20**(2):187–192, 2004. 53

[92] L. Lukas. *Least Squares Support Vector Machines Classification Applied To Brain Tumour Recognition Using Magnetic Resonance.* PhD thesis, Katholieke Universiteit Leuven, 2003. 53, 109

[93] A. W. Simonetti, W. J. Melssen, M. van der Graaf, G. J. Postma, A. Heerschap, and L. M. C. Buydens. A chemometric approach for brain tumor classification using magnetic resonance imaging and spectroscopy. *Analytical Chemistry*, **75**(20):5352–5361, 2003. 54

[94] J. M. García-Gómez, S. Tortajada, C. Vidal, M. Julià-Sapé, J. Luts, A. Moreno-Torres, S. Van Huffel, C. Arús, and M. Robles. The influence of combining two echo times in automatic brain tumour classification by magnetic resonance spectroscopy. *NMR in Biomedicine*, **21**(10):1112–1125, 2008. 54, 66

[95] A. Vellido, E. Romero, M. Julià-Sapé, C. Majós, A. Moreno-Torres, J. Pujol, and C. Arús. Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel ^{1}H-MRS. *NMR in Biomedicine*, **25**(6):819–828, 2012. 54, 70

[96] A. Vellido, E. Biganzoli, and P. J. G. Lisboa. Machine learning in cancer research: implications for personalised medicine. In *Proceedings of the 16th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 55–64, 2008. 54

[97] T. Dietterich. Machine-learning research: four current directions. *AI Magazine*, **18**:97–136, 1997. 57

[98] L. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley-Interscience, 2004. 58, 62

[99] M. Grabisch and J. Nicolas. Classification by fuzzy integral: Performance and tests. *Fuzzy Sets and Systems*, **65**(2-3):255 – 271, 1994. 58

[100] D. Wolpert. Stacked generalization. *Neural Networks*, **5**:241–259, 1992. 59

[101] S. Puuronen, V. Terziyan, and A. Tsymbal. A dynamic integration algorithm for an ensemble of classifiers. In Z. Ras and A. Skowron, editors, *Foundations of Intelligent Systems*, **1609** of *Lecture Notes in Computer Science*, pages 592–600. Springer Berlin / Heidelberg, 1999. 59

[102] B. Parmanto, P. W. Munro, and H. Doyle. Improving committee diagnosis with resampling techniques. In *Advances in Neural Information Processing Systems 8*, pages 882–888. MIT Press, 1996. 60

[103] D. Opitz. Feature selection for ensembles. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, AAAI '99/IAAI '99, pages 379–384, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence. 60

[104] N. C. Oza and K. Tumer. Input decimation ensembles: Decorrelation through dimensionality reduction. In *Proceedings of the Second International Workshop on Multiple Classifier Systems*, MCS '01, pages 238–247, London, UK, 2001. Springer-Verlag. 60

[105] A. R. Tate, J. Underwood, D. M. Acosta, M. Julià-Sapé, C. Majós, A. Moreno-Torres, F. A. Howe, M. van der Graaf, V. Lefournier, M. M. Murphy, A. Loosemore, C. Ladroue, P. Wesseling, J. Luc Bosson, M. E. Cabañas, A. W. Simonetti, W. Gajewicz, J. Calvar, A. Capdevila, P. R. Wilkins, B. A. Bell, C. Rémy, A. Heerschap, D. Watson, J. R. Griffiths, and C. Arús. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR in Biomedicine*, **19**(4):411–434, 2006. 62

[106] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*, **514**. Wiley-Interscience, 2004. 62

[107] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, **9**(3):293–300, 1999. 62

[108] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, **10**(3):61–74, 1999. 62

[109] L. Breiman. Arcing classifiers. *Annals of Statistics*, **26**(3):801–824, 1998. 72

[110] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. 2 edition. 76, 85, 133

[111] K. Crammer, R. Gilad-bachrach, A. Navot, and N. Tishby. Margin analysis of the LVQ algorithm. In *Advances in Neural Information Processing Systems 2002*, pages 462–469. MIT press, 2002. 77

[112] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1–3):389–422, 2002. 78, 79

[113] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proceedings of the European conference on machine learning on Machine Learning*, ECML-94, pages 171–182, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc. 79

[114] K. Dunne, P. Cunningham, and F. Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CD-2002-28, Dept. of Computer Science, Trinity College, 2002. 81

[115] S. Alelyani, Z. Zhao, and H. Liu. A dilemma in assessing stability of feature selection algorithms. In *2011 IEEE 13th International Conference on High Performance Computing and Communications (HPCC)*, pages 701–707, 2011. 81

[116] P. Křížek, J. Kittler, and V. Hlaváč. Improving stability of feature selection methods. In *Proceedings of the 12th international conference on Computer analysis of images and patterns*, CAIP'07, pages 929–936, Berlin, Heidelberg, 2007. Springer-Verlag. 82

[117] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1992. 83

[118] L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 803–811, New York, NY, USA, 2008. ACM. 83

[119] M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall Monographs on Statistics & Applied Probability)*. Chapman and Hall, 1995. 83

[120] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(8):790–799, 1995. 83

[121] A. Woźnica, P. Nguyen, and A. Kalousis. Model mining for robust feature selection. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 913–921, New York, NY, USA, 2012. ACM. 84

[122] C. Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(6):437–456, 2012. 85

[123] Y. Han. *Stable Feature Selection: Theory and Algorithms*. PhD thesis, State University of New York at Binghamton, 2012. 85

[124] Y. Huang, P. J. G. Lisboa, and W. El-Deredy. Tumour grading from magnetic resonance spectroscopy: a comparison of feature extraction with variable selection. *Statistics in Medicine*, **22**(1):147–164, 2003. 99

[125] C. Ladroue, F. Howe, J. Griffiths, and A. Tate. Independent component analysis for automated decomposition of in vivo magnetic resonance spectra. *Magnetic Resonance in Medicine*, **50**(4):697–703, 2003. 99

[126] H. Han and X.-L. Li. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. *BMC Bioinformatics*, **12**(1):57, 2011. 99

[127] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755):788–791, 1999. 102, 107

[128] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562, 2001. 102

[129] A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In *Independent Component Analysis and Blind Signal Separation*, **3889** of *Lecture Notes in Computer Science*, pages 32–39. Springer Berlin Heidelberg, 2006. 103

[130] R. Zdunek and A. Cichocki. Non-negative matrix factorization with quasi-newton optimization. In L. Rutkowski, R. Tadeusiewicz, L. Zadeh, and J. urada, editors, *Artificial Intelligence and Soft Computing ICAISC 2006*, **4029** of *Lecture Notes in Computer Science*, pages 870–879. Springer Berlin Heidelberg, 2006. 103

[131] C. J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, **19**:2756–2779, 2007. 103

[132] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**:45–55, 2010. 103, 147, 196

[133] Y. Wang, Y. Jia, C. Hu, and M. Turk. Non-negative matrix factorization framework for face recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, **19**(4):1–17, 2005. 104

[134] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks*, **17**(3):683 –695, 2006. 104, 109

[135] S.-Y. Lee, H.-A. Song, and S.-i. Amari. A new discriminant NMF algorithm and its application to the extraction of subtle emotional differences in speech. *Cognitive Neurodynamics*, **6**(6):525–535, 2012. 105

[136] I. Kotsia, S. Zafeiriou, and I. Pitas. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Transactions on Information Forensics and Security*, **2**(3):588 –595, 2007. 105, 109, 171

[137] O. Zoidi, A. Tefas, and I. Pitas. Multiplicative update rules for concurrent nonnegative matrix factorization and maximum margin classification. *IEEE Transactions on Neural Networks and Learning Systems*, **24**(3):422–434, 2013. 106, 171

[138] H. Lee, J. Yoo, and S. Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, **17**(1):4–7, 2010. 106

[139] M. Ochs, R. Stoyanova, F. Arias-Mendoza, and T. Brown. A new method for spectral decomposition using a bilinear Bayesian approach. *Journal of Magnetic Resonance*, **137**(1):161 – 176, 1999. 106

[140] P. Sajda, S. Du, and L. C. Parra. Recovery of constituent spectra using non-negative matrix factorization. In *Optical Science and Technology, SPIE's 48th Annual Meeting*. 107

[141] P. Sajda, S. Du, T. Brown, R. Stoyanova, D. Shungu, X. Mao, and L. Parra. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *IEEE Transactions on Medical Imaging*, **23**(12):1453–1465, 2004. 107

[142] S. Du, X. Mao, P. Sajda, and D. C. Shungu. Automated tissue segmentation and blind recovery of ^{1}H-MRS imaging spectral patterns of normal and diseased human brain. *NMR in Biomedicine*, **21**(1):33–41, 2008. 107

[143] Y. Su, S. B. Thakur, K. Sasan, S. Du, P. Sajda, W. Huang, and L. C. Parra. Spectrum separation resolves partial-volume effect of MRSI as demonstrated on brain tumor scans. *NMR in Biomedicine*, **21**(10):1030–1042, 2008. 107

[144] A. Croitor Sava, D. Sima, M. Martinez-Bisbal, B. Celda, and S. Van Huffel. Non-negative blind source separation techniques for tumor tissue typing using HR-MAS signals. In *Engineering in Medicine and Biology*

Society (EMBC), 2010 Annual International Conference of the IEEE, pages 3658–3661, 2010. 107

[145] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**(12):1495–1502, 2007. 107

[146] S. Ortega-Martorell, H. Ruiz, A. Vellido, I. Olier, E. Romero, M. Julià-Sapé, J. D. Martín, I. H. Jarman, C. Arús, and P. J. G. Lisboa. A novel semi-supervised methodology for extracting tumor type-specific MRS sources in human brain data. *PLoS ONE*, **8**(12):e83773, 2013. 108

[147] S.-I. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, **47**(5):1701–1711, 2001. 108

[148] Y. Li, D. M. Sima, S. V. Cauter, A. R. Croitor Sava, U. Himmelreich, Y. Pi, and S. Van Huffel. Hierarchical non-negative matrix factorization (hNMF): a tissue pattern differentiation method for glioblastoma multiforme diagnosis using MRSI. *NMR in Biomedicine*, **26**(3):307–319, 2013. 108

[149] S. Ortega-Martorell, I. Olier, M. Julià-Sapé, C. Arús, and P. J. G. Lisboa. Automatic relevance source determination in human brain tumors using Bayesian NMF. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 99–104, 2014. 108

[150] T. Laudadio, A. Sava, Y. Li, N. Sauwen, D. Sima, and S. Van Huffel. NMF in MR spectroscopy. In G. R. Naik, editor, *Non-negative Matrix Factorization Techniques*, Signals and Communication Technology, pages 161–177. Springer Berlin Heidelberg, 2016. 108

[151] K. Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Computational Biology*, **4**(7):e1000029, 2008. 108

[152] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1):1–38, 1977. 114

[153] S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, **37**(11):2217–2232, 2004. 117

[154] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009. 127

[155] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**:267–288, 1994. 128

[156] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2007. 128

[157] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, New York, NY, USA, 2008. ACM. 131, 132

[158] W. R. Gilks. *Markov Chain Monte Carlo In Practice*. Chapman and Hall/CRC, 1999. 133

[159] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2004. 134, 135

[160] J. Niemi. Metropolis-Hastings algorithm. http://www.jarad.me/stat544/2013/03/metropolis-hastings-algorithm/. [Online; Accessed: March 2015]. 134

[161] Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, 2007. 137

[162] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, **37**(2):183–233, 1999. 137

[163] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, **1**(1-2):1–305, 2008. 137

[164] A. Gelman and X.-L. Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**:163–185, 1998. 138

[165] S. Moussaoui, D. Brie, O. Caspary, and A. Mohammad-Djafari. A Bayesian method for positive source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, **5**, pages V–485–8, 2004. 138

[166] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling. *IEEE Transactions on Signal Processing*, **54**(11):4133–4145, 2006. 139

[167] M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In T. Adali, C. Jutten, J. Romano, and A. Barros, editors, *Independent Component Analysis and Signal Separation*, **5441** of *Lecture Notes in Computer Science*, pages 540–547. Springer Berlin Heidelberg, 2009. 140

[168] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, **48**(3):259–302, 1986. 140

[169] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, (4):1–17, 2009. 140

[170] S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, **90**(432):1313–1321, 1995. 153

[171] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, **51**(1):107–113, 2008. 169

[172] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia, 2007. 204

[173] M. I. Jordan. The Conjugate Prior for the Normal Distribution. Technical report, University of California, Berkeley, 2010. 207

Appendix A

Mathematical derivations of the Discriminant Convex Non-negative Matrix Factorisation optimisation function

The first of the appendices corresponds to the mathematical derivations that lead to the update rules for matrices $\mathbf{H}$, $\mathbf{W}$ and $\mathbf{q}$.

A.1 Update rule for mixing matrix H

For solving the objective function described in Eq. 6.8, an update rule for the mixing matrix $\mathbf{H}$ is derived.

First step (following the CNMF definition) is to ensure that the mixing matrix $\mathbf{H}$ is always non-negative (i.e., each value $\mathbf{H}_{ik}$ must always be greater than or equal to 0). This constraint is enforced by adding Lagrangian multipliers $\boldsymbol{\beta}_{ik}$ to the objective function. That is:

$$\Omega' = Tr\Big[\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{H}^\top + (2\alpha - 2)\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{H}^\top$$

A. MATHEMATICAL DERIVATIONS OF THE DISCRIMINANT CONVEX NON-NEGATIVE MATRIX FACTORISATION OPTIMISATION FUNCTION

$$- \quad 2\alpha\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top + \frac{\alpha}{N}\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top - \boldsymbol{\beta}\mathbf{H}^\top\Big].$$

Then, the gradient of the objective function with respect to $\mathbf{H}$ is calculated in order to obtain the update rule for $\mathbf{H}$. This gradient must equal 0 at convergence:

$$\begin{aligned}\frac{\partial\Omega'}{\partial\mathbf{H}} &= 2\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H} + (2\alpha-2)\mathbf{W}^\top\mathbf{X}^\top\mathbf{X} - 4\alpha\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}} \\ &+ \frac{2\alpha}{N}\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N - \boldsymbol{\beta} = 0.\end{aligned}$$

From the Karush-Kuhn-Tucker (KKT) complementary slackness condition, we obtain the following fixed point equation that the solution must satisfy at convergence:

$$\begin{aligned}&\Big(2\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H} + (2\alpha-2)\mathbf{W}^\top\mathbf{X}^\top\mathbf{X} - 4\alpha\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}} \\ &+\frac{2\alpha}{N}\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\Big)_{ik}\mathbf{H}_{ik} = \boldsymbol{\beta}_{ik}\mathbf{H}_{ik} = 0.\end{aligned}$$

This equation holds when either the first or the second factor equals zero. Similarly, the following equation holds if and only if the previous one does. Such transformation does not affect the current derivation, but it will help to ensure convergence, as will be seen in next section.

$$\begin{aligned}&\Big(2\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H} + (2\alpha-2)\mathbf{W}^\top\mathbf{X}^\top\mathbf{X} - 4\alpha\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}} \\ &+\frac{2\alpha}{N}\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\Big)_{ik}\mathbf{H}_{ik}^2 = 0.\end{aligned}$$

If a matrix $\mathbf{A}$ contains negative values, it can be transformed into a subtraction of two non-negative matrices $\mathbf{A}^+$ and $\mathbf{A}^-$:

$$\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^- = \left(\frac{|\mathbf{A}| + \mathbf{A}}{2}\right) - \left(\frac{|\mathbf{A}| - \mathbf{A}}{2}\right),$$

where $|\mathbf{A}|$ is the matrix containing the absolute value of each element in $\mathbf{A}$.

Following this identity, under the constraint that the only matrix eligible to be negative in our equation is $\mathbf{X}$, we decompose $(\mathbf{X}^\top\mathbf{X})$ into $(\mathbf{X}^\top\mathbf{X})^+ -$

$(\mathbf{X}^\top\mathbf{X})^-$. After applying this transformation and factorising the equation to obtain positive summands at both sides of the equality, we obtain:

$$\begin{aligned}
&\Big[\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H} + (1-\alpha)\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^- + 2\alpha\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\tilde{\mathbf{M}} \\
&+\frac{\alpha}{N}\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{J}_N\Big]_{ik}\mathbf{H}_{ik}^2 \\
= \\
&\Big[\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H} + (1-\alpha)\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+ + 2\alpha\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\tilde{\mathbf{M}} \\
&+\frac{\alpha}{N}\mathbf{W}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{J}_N\Big]_{ik}\mathbf{H}_{ik}^2.
\end{aligned}$$

Finally, the update rule for the mixing matrix $\mathbf{H}$ (expressed in Eq. 6.9) is obtained by solving the last equation by $\mathbf{H}$, which satisfies the fixed point equation at convergence, $\mathbf{H}^\infty = \mathbf{H}^{t+1} = \mathbf{H}^t$.

A.2 Update rule for unmixing matrix W

Following the same reasoning as with the mixing matrix $\mathbf{H}$, an update rule for the unmixing matrix $\mathbf{W}$ is derived, which corresponds to the second necessary piece to solve the objective function described in Eq. 6.8.

First, the constraints that enforce non-negativity of matrix $\mathbf{W}$ are set by Lagrangian multipliers $\boldsymbol{\gamma}_{ik}$ included in the previously defined objective function:

$$\begin{aligned}
\Omega' &= Tr\Big[\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{H}^\top\mathbf{W}^\top + (2\alpha-2)\mathbf{X}^\top\mathbf{X}\mathbf{H}^\top\mathbf{W}^\top \\
&- 2\alpha\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top\mathbf{W}^\top + \frac{\alpha}{N}\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top\mathbf{W}^\top - \boldsymbol{\gamma}\mathbf{W}^\top\Big].
\end{aligned}$$

Afterwards, we calculate the gradient of the objective function with respect to $\mathbf{W}$, which must equal 0 at convergence:

$$\begin{aligned}
\frac{\partial\Omega'}{\partial\mathbf{W}} &= 2\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{H}^\top + (2\alpha-2)\mathbf{X}^\top\mathbf{X}\mathbf{H}^\top - 4\alpha\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top \\
&+ \frac{2\alpha}{N}\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top - \boldsymbol{\gamma} = 0.
\end{aligned}$$

A. MATHEMATICAL DERIVATIONS OF THE DISCRIMINANT CONVEX NON-NEGATIVE MATRIX FACTORISATION OPTIMISATION FUNCTION

From the KKT complementary slackness condition, the following fixed point equation that the solution must satisfy at convergence is obtained:

$$\Big(2\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{H}^\top + (2\alpha-2)\mathbf{X}^\top\mathbf{X}\mathbf{H}^\top - 4\alpha\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top + \frac{2\alpha}{N}\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top\Big)_{ik}\mathbf{W}_{ik} = \boldsymbol{\gamma}_{ik}\mathbf{W}_{ik} = 0.$$

The equation holds when either the first or the second factor equals zero. Again, the following equation will hold if and only if the previous one does, ensuring convergence:

$$\Big(2\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{H}^\top + (2\alpha-2)\mathbf{X}^\top\mathbf{X}\mathbf{H}^\top - 4\alpha\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top + \frac{2\alpha}{N}\mathbf{X}^\top\mathbf{X}\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top\Big)_{ik}\mathbf{W}_{ik}^2 = 0.$$

Decomposing $(\mathbf{X}^\top\mathbf{X})$ into $(\mathbf{X}^\top\mathbf{X})^+ - (\mathbf{X}^\top\mathbf{X})^-$ and factorising to obtain positive summands at both sides of the equation we get:

$$\begin{aligned}
&\Big[(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{H}^\top + (1-\alpha)(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top + 2\alpha(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top \\
&+\frac{\alpha}{N}(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top\Big]_{ik}\mathbf{W}_{ik}^2 \\
=\ &\Big[(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{H}^\top + (1-\alpha)(\mathbf{X}^\top\mathbf{X})^+\mathbf{H}^\top + 2\alpha(\mathbf{X}^\top\mathbf{X})^+\mathbf{W}\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top \\
&+\frac{\alpha}{N}(\mathbf{X}^\top\mathbf{X})^-\mathbf{W}\mathbf{H}\mathbf{J}_N\mathbf{H}^\top\Big]_{ik}\mathbf{W}_{ik}^2.
\end{aligned}$$

Last, the resulting update rule for the unmixing matrix $\mathbf{W}$ (described in Eq. 6.10) can be obtained, satisfying the fixed point equation. That is, at convergence, $\mathbf{W}^\infty = \mathbf{W}^{t+1} = \mathbf{W}^t$.

A.3 Update rule for vector q in the prediction phase

Last block in this section presents the derivation of an update rule for each of the rows $\mathbf{q}$ in the mixing matrix $\mathbf{Q}$ for prediction, following the same procedure as for $\mathbf{H}$ and $\mathbf{W}$ matrices, necessary to solve the objective function described in Eq. 6.11.

Once more, the constraints that enforce non-negativity of matrix $\mathbf{q}$ are set by Lagrangian multipliers $\boldsymbol{\eta}_{ik}$, applied to the previous function:

$$\begin{aligned}\bar{\Omega}' &= Tr\Big[-(2-2\alpha)\mathbf{S}^\top\mathbf{Z}\mathbf{V}^\top\mathbf{q}^\top + \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{B}}\mathbf{q}^\top \\ &+ \mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{C}}\mathbf{q}^\top - \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{E}}\mathbf{q}^\top - \mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{F}}\mathbf{q}^\top - \boldsymbol{\eta}\mathbf{q}^\top\Big].\end{aligned}$$

Then, the gradient of the objective function with respect to $\mathbf{q}$, which must equal 0 at convergence, is calculated:

$$\frac{\partial\bar{\Omega}'}{\partial\mathbf{q}} = (2\alpha-2)\mathbf{S}^\top\mathbf{Z}\mathbf{V}^\top + \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{B}} + 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{C}} - \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{E}} - 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{F}} - \boldsymbol{\eta} = 0.$$

From the KKT complementary slackness condition, the following fixed point equation that the solution must satisfy at convergence is obtained:

$$\left((2\alpha-2)\mathbf{S}^\top\mathbf{Z}\mathbf{V}^\top + \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{B}} + 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{C}} - \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{E}} - 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{F}}\right)_{ik}\mathbf{q}_{ik} = \boldsymbol{\eta}_{ik}\mathbf{q}_{ik} = 0.$$

Such equality holds whenever either the first or the second factor equals zero. Similarly, the following equation will hold if and only if the previous one does:

$$\left((2\alpha-2)\mathbf{S}^\top\mathbf{Z}\mathbf{V}^\top + \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{B}} + 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{C}} - \mathbf{S}^\top\mathbf{S}\mathbf{H}\bar{\mathbf{E}} - 2\mathbf{S}^\top\mathbf{S}\mathbf{q}\bar{\mathbf{F}}\right)_{ik}\mathbf{q}_{ik}^2 = 0.$$

By decomposing $(\mathbf{S}^\top\mathbf{Z})$ into $(\mathbf{S}^\top\mathbf{Z})^+ - (\mathbf{S}^\top\mathbf{Z})^-$, $(\mathbf{S}^\top\mathbf{S})$ into $(\mathbf{S}^\top\mathbf{S})^+ - (\mathbf{S}^\top\mathbf{S})^-$ and factorising the equation to obtain positive summands at both sides of the equality:

A. MATHEMATICAL DERIVATIONS OF THE DISCRIMINANT CONVEX NON-NEGATIVE MATRIX FACTORISATION OPTIMISATION FUNCTION

$$
\begin{aligned}
&\Bigg[2(1-\alpha)(\mathbf{S}^\top\mathbf{Z})^-\mathbf{V}^\top + (\mathbf{S}^\top\mathbf{S})^+\mathbf{H}\bar{\mathbf{B}} + 2(\mathbf{S}^\top\mathbf{S})^+\mathbf{q}\bar{\mathbf{C}} \\
&+(\mathbf{S}^\top\mathbf{S})^-\mathbf{H}\bar{\mathbf{E}} + 2(\mathbf{S}^\top\mathbf{S})^-\mathbf{q}\bar{\mathbf{F}}\Bigg]_{ik}\mathbf{q}_{ik}^2 \\
= &\Bigg[2(1-\alpha)(\mathbf{S}^\top\mathbf{Z})^+\mathbf{V}^\top + (\mathbf{S}^\top\mathbf{S})^-\mathbf{H}\bar{\mathbf{B}} + 2(\mathbf{S}^\top\mathbf{S})^-\mathbf{q}\bar{\mathbf{C}} \\
&+(\mathbf{S}^\top\mathbf{S})^+\mathbf{H}\bar{\mathbf{E}} + 2(\mathbf{S}^\top\mathbf{S})^+\mathbf{q}\bar{\mathbf{F}}\Bigg]_{ik}\mathbf{q}_{ik}^2,
\end{aligned}
$$

which leads to the update rule for the vector $\mathbf{q}$ described in Eq. 6.12 in the main text, satisfying $\mathbf{q}^\infty = \mathbf{q}^{t+1} = \mathbf{q}^t$ at convergence.

Appendix B

Discriminant Convex Non-negative Matrix Factorisation: proof of convergence

The objective function presented in Eq. 6.8 is not convex for matrices $\mathbf{H}$ and $\mathbf{W}$ simultaneously, meaning that it unavoidably converges to local minimum. However, this function is convex with respect to each matrix separately. We prove the convergence of the alternating update algorithm by defining an appropriate convex auxiliary function and finding its global minimum. Convergence of the predictive update rule (Eq. 6.12) will also be proved.

A function $Z(\mathbf{L}, \tilde{\mathbf{L}})$ is an auxiliary function of $\Omega(\mathbf{L})$ if it satisfies

$$Z(\mathbf{L}, \tilde{\mathbf{L}}) \geq \Omega(\mathbf{L}), Z(\mathbf{L}, \mathbf{L}) = \Omega(\mathbf{L})$$

for any $\mathbf{L}, \tilde{\mathbf{L}}$. Let us define

$$\mathbf{L}^{(t+1)} = \arg\min_{\mathbf{L}} Z(\mathbf{L}, \mathbf{L}^{(t)}).$$

By construction, we have

$$\Omega(\mathbf{L}^{(t)}) = Z(\mathbf{L}^{(t)}, \mathbf{L}^{(t)}) \geq Z(\mathbf{L}^{(t+1)}, \mathbf{L}^{(t)}) \geq \Omega(\mathbf{L}^{(t+1)}).$$

B. DISCRIMINANT CONVEX NON-NEGATIVE MATRIX FACTORISATION: PROOF OF CONVERGENCE

In the subsequent blocks, appropriate convex auxiliary functions $Z(\mathbf{L}, \tilde{\mathbf{L}})$ for the mixing ($\mathbf{H}$ and $\mathbf{q}$) and unmixing ($\mathbf{W}$) matrices will be defined. They will help to prove convergence.

B.1 Proof of convergence for the H update rule

From Eq. 6.8, where $\tilde{\mathbf{X}} = \mathbf{XWH}$, we substitute $\mathbf{A} \leftarrow \mathbf{W}^\top\mathbf{X}^\top\mathbf{XW}$, $\mathbf{B} \leftarrow \mathbf{W}^\top\mathbf{X}^\top\mathbf{X}$, $\mathbf{L} \leftarrow \mathbf{H}$, $\sigma \leftarrow (1-\alpha)$; separating positive and negative values according to $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$ and $\mathbf{B} = \mathbf{B}^+ - \mathbf{B}^-$, we then obtain:

$$\begin{aligned}\Omega(\mathbf{L}) &= Tr[\mathbf{A}^+\mathbf{L}\mathbf{L}^\top - \mathbf{A}^-\mathbf{L}\mathbf{L}^\top - 2\sigma\mathbf{B}^+\mathbf{L}^\top + 2\sigma\mathbf{B}^-\mathbf{L}^\top - 2\alpha\mathbf{A}^+\mathbf{L}\tilde{\mathbf{M}}\mathbf{L}^\top \\ &+ 2\alpha\mathbf{A}^-\mathbf{L}\tilde{\mathbf{M}}\mathbf{L}^\top + \frac{\alpha}{N}\mathbf{A}^+\mathbf{L}\mathbf{J}_N\mathbf{L}^\top - \frac{\alpha}{N}\mathbf{A}^-\mathbf{L}\mathbf{J}_N\mathbf{L}^\top],\end{aligned}$$

out of which upper and lower bounds for positive and negative summands, respectively, can be found.

Using the inequality $a \leq \frac{(a^2+b^2)}{2b}$, which holds for any $a, b > 0$, and setting $a \leftarrow \mathbf{B}^-\mathbf{L}^\top$, $b \leftarrow \mathbf{B}^-\tilde{\mathbf{L}}^\top$, we obtain an upper bound for the fourth term:

$$Tr(2\sigma\mathbf{B}^-\mathbf{L}^\top) = 2\sigma\sum_{ik}\mathbf{B}^-_{ik}\mathbf{L}_{ik} \leq \sigma\sum_{ik}\mathbf{B}^-_{ik}\frac{\mathbf{L}^2_{ik} + \tilde{\mathbf{L}}^2_{ik}}{\tilde{\mathbf{L}}_{ik}}.$$

For the remaining three positive summands, the following inequality [132] will be used:

$$Tr(\mathbf{S}^\top\mathbf{ASB}) \leq \sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(\mathbf{A}\tilde{\mathbf{S}}\mathbf{B})_{ip}\mathbf{S}^2_{ip}}{\tilde{\mathbf{S}}_{ip}}.$$

An upper bound for the first term will be found by setting $\mathbf{A} \leftarrow \mathbf{A}^+$, $\mathbf{B} \leftarrow \mathbf{I}$ and $\mathbf{S} \leftarrow \mathbf{L}$:

$$Tr(\mathbf{A}^+\mathbf{L}\mathbf{L}^\top) \leq \sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}})_{ik}\mathbf{L}^2_{ik}}{\tilde{\mathbf{L}}_{ik}}.$$

Similarly, by setting $\mathbf{A} \leftarrow \mathbf{A}^-$, $\mathbf{B} \leftarrow \tilde{\mathbf{M}}$ and $\mathbf{S} \leftarrow \mathbf{L}$, we get an upper bound for the sixth term:

$$Tr(2\alpha\mathbf{A}^-\mathbf{L}\tilde{\mathbf{M}}\mathbf{L}^\top) \leq 2\alpha\sum_{ik}\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}\mathbf{L}^2_{ik}}{\tilde{\mathbf{L}}_{ik}}.$$

Finally, an upper bound for the seventh term can be obtained by setting $\mathbf{A} \leftarrow \mathbf{A}^+$, $\mathbf{B} \leftarrow \mathbf{J}_N$ and $\mathbf{S} \leftarrow \mathbf{L}$:

$$Tr\left(\frac{\alpha}{N}\mathbf{A}^+\mathbf{L}\mathbf{J}_N\mathbf{L}^\top\right) \leq \frac{\alpha}{N}\sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}}.$$

We now turn our attention to obtain lower bounds for the negative summands in the equation. For that, we use the inequality that states $z \geq 1 + \log z$ for any $z > 0$. By setting $z \leftarrow \frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}$, the following equation is obtained:

$$\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} \geq 1 + \log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}.$$

Then, if each side of this inequality is multiplied by the third term of our equation, and after simplification, we get:

$$Tr(2\sigma\mathbf{B}^+\mathbf{L}^\top) = 2\sigma\sum_{ik}\mathbf{B}_{ik}^+\mathbf{L}_{ik} \geq 2\sigma\sum_{ik}\mathbf{B}_{ik}^+\tilde{\mathbf{L}}_{ik}\left(1 + \log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right).$$

Likewise, setting $z \leftarrow \frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}$ leads to lower bounds for the second, fifth and last summands:

$$Tr(\mathbf{A}^-\mathbf{L}\mathbf{L}^\top) \geq \sum_{ikj}\mathbf{A}_{ij}^-\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right),$$

$$Tr(2\alpha\mathbf{A}^+\mathbf{L}\tilde{\mathbf{M}}\mathbf{L}^\top) \geq 2\alpha\sum_{ikjl}\mathbf{A}_{ij}^+\tilde{\mathbf{M}}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right),$$

and

$$Tr\left(\frac{\alpha}{N}\mathbf{A}^-\mathbf{L}\mathbf{J}_N\mathbf{L}^\top\right) \geq \frac{\alpha}{N}\sum_{ikjl}\mathbf{A}_{ij}^-\mathbf{J}_{N_{kl}}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right).$$

Therefore, an auxiliary function that bounds our objective is obtained by locating all bounds together:

$$\begin{aligned}
Z(\mathbf{L},\tilde{\mathbf{L}}) &= \sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} - \sum_{ikj}\mathbf{A}_{ij}^-\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right) \\
&- 2\sigma\sum_{ik}\mathbf{B}_{ik}^+\tilde{\mathbf{L}}_{ik}\left(1 + \log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right) + \sigma\sum_{ik}\mathbf{B}_{ik}^-\frac{\mathbf{L}_{ik}^2 + \tilde{\mathbf{L}}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} \\
&- 2\alpha\sum_{ikjl}\mathbf{A}_{ij}^+\tilde{\mathbf{M}}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right) + 2\alpha\sum_{ik}\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} \\
&+ \frac{\alpha}{N}\sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} - \frac{\alpha}{N}\sum_{ikjl}\mathbf{A}_{ij}^-\mathbf{J}_{N_{kl}}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}\left(1 + \log\frac{\mathbf{L}_{ik}\mathbf{L}_{jk}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jk}}\right).
\end{aligned}$$

In order to find the minimum of $Z(\mathbf{L}, \tilde{\mathbf{L}})$, the gradient is calculated:

$$\begin{aligned}\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial \mathbf{L}_{ik}} &= 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} - 2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} - 2\sigma\frac{\mathbf{B}^{+}_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} + 2\sigma\frac{\mathbf{B}^{-}_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &- 4\alpha\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} + 4\alpha\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ \frac{2\alpha}{N}\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} - \frac{2\alpha}{N}\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}.\end{aligned}$$

The Hessian matrix containing the second derivatives is defined as

$$\frac{\partial^2 Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial \mathbf{L}_{ik}\partial \mathbf{L}_{jl}} = \boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{kl}\mathbf{Y}_{ik},$$

being a diagonal matrix with positive entries, where

$$\begin{aligned}\mathbf{Y}_{ik} &= 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}})_{ik}}{\tilde{\mathbf{L}}_{ik}} + 2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 2\sigma\frac{\mathbf{B}^{+}_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 2\sigma\frac{\mathbf{B}^{-}_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ 4\alpha\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 4\alpha\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ \frac{2\alpha}{N}\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}}{\tilde{\mathbf{L}}_{ik}} + \frac{2\alpha}{N}\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}}.\end{aligned}$$

Hence, $Z(\mathbf{L}, \tilde{\mathbf{L}})$ is a convex function of $\mathbf{L}$. Then, the global minimum is obtained by setting $\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial \mathbf{L}_{ik}} = 0$ and solving for $\mathbf{L}$. Rearranging terms, we obtain

$$\begin{aligned}\mathbf{L}_{ik} &= \arg\min_{\mathbf{L}} Z(\mathbf{L},\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}_{ik}\sqrt{\frac{\breve{\mathbf{B}}_{\mathbf{L}_{ik}}}{\breve{\mathbf{V}}_{\mathbf{L}_{ik}}}} \\ \breve{\mathbf{B}}_{\mathbf{L}_{ik}} &= (\mathbf{A}^{-}\tilde{\mathbf{L}})_{ik} + \sigma\mathbf{B}^{+}_{ik} + 2\alpha(\mathbf{A}^{+}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik} + \frac{\alpha}{N}(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik} \\ \breve{\mathbf{V}}_{\mathbf{L}_{ik}} &= (\mathbf{A}^{+}\tilde{\mathbf{L}})_{ik} + \sigma\mathbf{B}^{-}_{ik} + 2\alpha(\mathbf{A}^{-}\tilde{\mathbf{L}}\tilde{\mathbf{M}})_{ik} + \frac{\alpha}{N}(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{J}_N)_{ik}.\end{aligned}$$

Changing back to $(\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}) \leftarrow \mathbf{A}$, $(\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}) \leftarrow \mathbf{B}$, $\mathbf{H} \leftarrow \mathbf{L}$, $(1-\alpha) \leftarrow \sigma$, we retrieve the update rule for $\mathbf{H}$ (Eq. 6.9).

B.2 Proof of convergence for the W update rule

Now, we follow the same approach as the one used in $\mathbf{H}$ with the purpose of proving the convergence of the update rule for $\mathbf{W}$. More precisely, from Eq. 6.8, bearing in mind that $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}\mathbf{H}$, we substitute $\mathbf{A} \leftarrow \mathbf{X}^\top\mathbf{X}$,

$\mathbf{B} \leftarrow \mathbf{H}\mathbf{H}^\top$, $\mathbf{C} \leftarrow \mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^\top$, $\mathbf{D} \leftarrow \mathbf{H}\mathbf{J}_N\mathbf{H}^\top$, $\mathbf{L} \leftarrow \mathbf{W}$, $\sigma \leftarrow (1-\alpha)$; separating positive and negative values according to $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, we obtain:

$$\begin{aligned}\Omega(\mathbf{L}) &= Tr[\mathbf{A}^+\mathbf{L}\mathbf{B}\mathbf{L}^\top - \mathbf{A}^-\mathbf{L}\mathbf{B}\mathbf{L}^\top - 2\sigma\mathbf{A}^+\mathbf{L}\mathbf{H} + 2\sigma\mathbf{A}^-\mathbf{L}\mathbf{H} - 2\alpha\mathbf{A}^+\mathbf{L}\mathbf{C}\mathbf{L}^\top \\ &+ 2\alpha\mathbf{A}^-\mathbf{L}\mathbf{C}\mathbf{L}^\top + \frac{\alpha}{N}\mathbf{A}^+\mathbf{L}\mathbf{D}\mathbf{L}^\top - \frac{\alpha}{N}\mathbf{A}^-\mathbf{L}\mathbf{D}\mathbf{L}^\top].\end{aligned}$$

We then find an auxiliary function for $\Omega(\mathbf{L})$, using the same inequalities as in the previous section to obtain each upper and lower bound. The auxiliary function now becomes

$$\begin{aligned}Z(\mathbf{L},\tilde{\mathbf{L}}) &= \sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{B})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} - \sum_{ijkl}\mathbf{A}_{ij}^-\mathbf{B}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}\left(1+\log\frac{\mathbf{L}_{ik}\mathbf{L}_{jl}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}}\right) \\ &- 2\sigma\sum_{ik}(\mathbf{A}^+\mathbf{H}^\top)_{ik}\tilde{\mathbf{L}}_{ik}\left(1+\log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right) + \sigma\sum_{ik}(\mathbf{A}^-\mathbf{H}^\top)_{ik}\frac{\mathbf{L}_{ik}^2+\tilde{\mathbf{L}}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} \\ &- 2\alpha\sum_{ijkl}\mathbf{A}_{ij}^+\mathbf{C}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}\left(1+\log\frac{\mathbf{L}_{ik}\mathbf{L}_{jl}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}}\right) + 2\alpha\sum_{ik}\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\mathbf{C})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} \\ &+ \frac{\alpha}{N}\sum_{ik}\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{D})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}} - \frac{\alpha}{N}\sum_{ijkl}\mathbf{A}_{ij}^-\mathbf{D}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}\left(1+\log\frac{\mathbf{L}_{ik}\mathbf{L}_{jl}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}}\right).\end{aligned}$$

We calculate the gradient in order to find the minimum of $Z(\mathbf{L},\tilde{\mathbf{L}})$:

$$\begin{aligned}\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial\mathbf{L}_{ik}} &= 2\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{B})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} - 2\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\mathbf{B})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} \\ &- 2\sigma\frac{(\mathbf{A}^+\mathbf{H}^\top)_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} + 2\sigma\frac{(\mathbf{A}^-\mathbf{H}^\top)_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &- 4\alpha\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{C})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}} + 4\alpha\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\mathbf{C})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ \frac{2\alpha}{N}\frac{(\mathbf{A}^+\tilde{\mathbf{L}}\mathbf{D})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}} - \frac{2\alpha}{N}\frac{(\mathbf{A}^-\tilde{\mathbf{L}}\mathbf{D})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}.\end{aligned}$$

The Hessian matrix containing second derivatives,

$$\frac{\partial^2 Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial\mathbf{L}_{ik}\partial\mathbf{L}_{jl}} = \boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{kl}\mathbf{Y}_{ik},$$

is a diagonal matrix with positive entries, where

$$\begin{aligned}\mathbf{Y}_{ik} &= 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{B})_{ik}}{\tilde{\mathbf{L}}_{ik}} + 2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{B})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 2\sigma\frac{(\mathbf{A}^{+}\mathbf{H}^{\top})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 2\sigma\frac{(\mathbf{A}^{-}\mathbf{H}^{\top})_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ 4\alpha\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{C})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}} + 4\alpha\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{C})_{ik}}{\tilde{\mathbf{L}}_{ik}} \\ &+ \frac{2\alpha}{N}\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{D})_{ik}}{\tilde{\mathbf{L}}_{ik}} + \frac{2\alpha}{N}\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{D})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}^2_{ik}}.\end{aligned}$$

Hence, $Z(\mathbf{L},\tilde{\mathbf{L}})$ is a convex function of $\mathbf{L}$. Then, the global minimum can be found by setting $\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial \mathbf{L}_{ik}} = 0$ and solving for $\mathbf{L}$. Rearranging terms, we obtain:

$$\begin{aligned}\mathbf{L}_{ik} &= \arg\min_{\mathbf{L}} Z(\mathbf{L},\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}_{ik}\sqrt{\frac{\breve{\mathbf{B}}_{\mathbf{L}_{ik}}}{\breve{\mathbf{V}}_{\mathbf{L}_{ik}}}} \\ \breve{\mathbf{B}}_{\mathbf{L}_{ik}} &= (\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{B})_{ik} + \sigma(\mathbf{A}^{+}\mathbf{H}^{\top})_{ik} + 2\alpha(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{C})_{ik} + \frac{\alpha}{N}(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{D})_{ik} \\ \breve{\mathbf{V}}_{\mathbf{L}_{ik}} &= (\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{B})_{ik} + \sigma(\mathbf{A}^{-}\mathbf{H}^{\top})_{ik} + 2\alpha(\mathbf{A}^{-}\tilde{\mathbf{L}}\mathbf{C})_{ik} + \frac{\alpha}{N}(\mathbf{A}^{+}\tilde{\mathbf{L}}\mathbf{D})_{ik}.\end{aligned}$$

Changing back to $(\mathbf{X}^{\top}\mathbf{X}) \leftarrow \mathbf{A}$, $(\mathbf{H}\mathbf{H}^{\top}) \leftarrow \mathbf{B}$, $(\mathbf{H}\tilde{\mathbf{M}}\mathbf{H}^{\top}) \leftarrow \mathbf{C}$, $(\mathbf{H}\mathbf{J}_N\mathbf{H}^{\top}) \leftarrow \mathbf{D}$,$\mathbf{W} \leftarrow \mathbf{L}$, $(1-\alpha) \leftarrow \sigma$, the update rule for $\mathbf{W}$ described by Eq. 6.10 in the main text can be retrieved.

B.3 Proof of convergence for the q update rule

Last proof of convergence for Eq. 6.12 is done using the same procedure as in $\mathbf{H}$ and $\mathbf{W}$. That is, from Eq. 6.11, we substitute $\mathbf{A} \leftarrow \mathbf{S}^{\top}\mathbf{S}$, $\mathbf{B} \leftarrow \mathbf{S}^{\top}\mathbf{Z}$, $\mathbf{L} \leftarrow \mathbf{q}$, $\sigma \leftarrow (1-\alpha)$; separating positive and negative values according to $\mathbf{A} = \mathbf{A}^{+} - \mathbf{A}^{-}$. We obtain:

$$\begin{aligned}\Omega'(\mathbf{L}) &= Tr[-2\sigma\mathbf{B}^{+}\mathbf{V}^{\top}\mathbf{L}^{\top} + 2\sigma\mathbf{B}^{-}\mathbf{V}^{\top}\mathbf{L}^{\top} + \mathbf{A}^{+}\mathbf{H}\bar{\mathbf{B}}\mathbf{L}^{\top} - \mathbf{A}^{-}\mathbf{H}\bar{\mathbf{B}}\mathbf{L}^{\top} + \mathbf{A}^{+}\mathbf{L}\bar{\mathbf{C}}\mathbf{L}^{\top} \\ &- \mathbf{A}^{-}\mathbf{L}\bar{\mathbf{C}}\mathbf{L}^{\top} - \mathbf{A}^{+}\mathbf{H}\bar{\mathbf{E}}\mathbf{L}^{\top} + \mathbf{A}^{-}\mathbf{H}\bar{\mathbf{E}}\mathbf{L}^{\top} - \mathbf{A}^{+}\mathbf{L}\bar{\mathbf{F}}\mathbf{L}^{\top} + \mathbf{A}^{-}\mathbf{L}\bar{\mathbf{F}}\mathbf{L}^{\top}].\end{aligned}$$

Using the same inequalities as in previous sections to obtain upper and

lower bounds, an auxiliary function for $\Omega'(\mathbf{L})$ is defined:

$$\begin{aligned}
Z(\mathbf{L},\tilde{\mathbf{L}}) &= -2\sigma\sum_{ik}(\mathbf{B}^{+}\mathbf{V}^{\top})_{ik}\tilde{\mathbf{L}}_{ik}\left(1+\log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right)+\sigma\sum_{ik}(\mathbf{B}^{-}\mathbf{V}^{\top})_{ik}\frac{\mathbf{L}_{ik}^2+\tilde{\mathbf{L}}_{ik}^2}{\tilde{\mathbf{L}}_{ik}}\\
&+ \sum_{ik}(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{B}})_{ik}\frac{\mathbf{L}_{ik}^2+\tilde{\mathbf{L}}_{ik}^2}{2\tilde{\mathbf{L}}_{ik}}-\sum_{ik}(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{B}})_{ik}\tilde{\mathbf{L}}_{ik}\left(1+\log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right)\\
&+ \sum_{ik}\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}}-\sum_{ijkl}\mathbf{A}_{ij}^{-}\bar{\mathbf{C}}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}\left(1+\log\frac{\mathbf{L}_{ik}\mathbf{L}_{jl}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}}\right)\\
&- \sum_{ik}(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{E}})_{ik}\tilde{\mathbf{L}}_{ik}\left(1+\log\frac{\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\right)+\sum_{ik}(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{E}})_{ik}\frac{\mathbf{L}_{ik}^2+\tilde{\mathbf{L}}_{ik}^2}{\tilde{\mathbf{L}}_{ik}}\\
&- \sum_{ijkl}\mathbf{A}_{ij}^{+}\bar{\mathbf{F}}_{kl}\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}\left(1+\log\frac{\mathbf{L}_{ik}\mathbf{L}_{jl}}{\tilde{\mathbf{L}}_{ik}\tilde{\mathbf{L}}_{jl}}\right)+\sum_{ik}\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}\mathbf{L}_{ik}^2}{\tilde{\mathbf{L}}_{ik}}.
\end{aligned}$$

With the purpose to find the minimum of $Z(\mathbf{L},\tilde{\mathbf{L}})$, we calculate the gradient:

$$\begin{aligned}
\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial\mathbf{L}_{ik}} &= -2\sigma\frac{(\mathbf{B}^{+}\mathbf{V}^{\top})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}+2\sigma\frac{(\mathbf{B}^{-}\mathbf{V}^{\top})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}+\frac{(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{B}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\\
&- \frac{(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{B}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}+2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}-2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}\\
&- \frac{(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{E}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}+\frac{(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{E}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}\\
&- 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}}+2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}\mathbf{L}_{ik}}{\tilde{\mathbf{L}}_{ik}}.
\end{aligned}$$

The Hessian matrix containing second derivatives is

$$\frac{\partial^2 Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial\mathbf{L}_{ik}\partial\mathbf{L}_{jl}}=\boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{kl}\mathbf{Y}_{ik},$$

being a diagonal matrix with positive entries, such that

$$\begin{aligned}
\mathbf{Y}_{ik} &= 2\sigma\frac{(\mathbf{B}^{+}\mathbf{V}^{\top})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}^2}+2\sigma\frac{(\mathbf{B}^{-}\mathbf{V}^{\top})_{ik}}{\tilde{\mathbf{L}}_{ik}}+\frac{(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{B}})_{ik}}{\tilde{\mathbf{L}}_{ik}}+\frac{(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{B}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}^2}\\
&+ 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik}}{\tilde{\mathbf{L}}_{ik}}+2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}^2}+\frac{(\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{E}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}^2}+\frac{(\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{E}})_{ik}}{\tilde{\mathbf{L}}_{ik}}\\
&+ 2\frac{(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}\tilde{\mathbf{L}}_{ik}}{\mathbf{L}_{ik}^2}+2\frac{(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}}{\tilde{\mathbf{L}}_{ik}},
\end{aligned}$$

which means that $Z(\mathbf{L},\tilde{\mathbf{L}})$ is a convex function of $\mathbf{L}$. The global minimum can then be obtained by setting $\frac{\partial Z(\mathbf{L},\tilde{\mathbf{L}})}{\partial\mathbf{L}_{ik}}=0$ and solving for $\mathbf{L}$. Rearranging

terms, we obtain:

$$\begin{aligned}
\mathbf{L}_{ik} &= \arg\min_{\mathbf{L}} Z(\mathbf{L}, \tilde{\mathbf{L}}) = \tilde{\mathbf{L}}_{ik} \sqrt{\frac{\breve{\mathbf{B}}_{\mathbf{L}_{ik}}}{\breve{\mathbf{V}}_{\mathbf{L}_{ik}}}} \\
\breve{\mathbf{B}}_{\mathbf{L}_{ik}} &= 2\sigma(\mathbf{B}^{+}\mathbf{V}^{\top})_{ik} + (\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{B}})_{ik} + 2(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik} + (\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{E}})_{ik} + 2(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik} \\
\breve{\mathbf{V}}_{\mathbf{L}_{ik}} &= 2\sigma(\mathbf{B}^{-}\mathbf{V}^{\top})_{ik} + (\mathbf{A}^{+}\mathbf{H}\bar{\mathbf{B}})_{ik} + 2(\mathbf{A}^{+}\tilde{\mathbf{L}}\bar{\mathbf{C}})_{ik} + (\mathbf{A}^{-}\mathbf{H}\bar{\mathbf{E}})_{ik} + 2(\mathbf{A}^{-}\tilde{\mathbf{L}}\bar{\mathbf{F}})_{ik}.
\end{aligned}$$

Finally, changing back to $(\mathbf{S}^{\top}\mathbf{S}) \leftarrow \mathbf{A}$, $(\mathbf{S}^{\top}\mathbf{Z}) \leftarrow \mathbf{B}$, $\mathbf{q} \leftarrow \mathbf{L}$, $(1-\alpha) \leftarrow \sigma$, the update rule for $\mathbf{q}$ described in Eq. 6.12 is retrieved.

Appendix C

Mathematical derivations for the Bayesian Semi Non-negative Matrix Factorisation Gibbs sampler

This section corresponds to the detailed mathematical derivations of the equations required to build the Gibbs sampler as expressed in Section 7.3.2.1. We start by reminding the posterior density we aim at approximating:

$$p\left(\mathbf{S},\mathbf{H},\sigma^2 \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\sigma^2\right)\cdot p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right)\cdot p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right)\cdot p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right), \quad \text{(C.1)}$$

where

$$\begin{aligned} p\left(\mathbf{X} \mid \mathbf{S},\mathbf{H},\sigma^2\right) &= \prod_{d=1}^{D}\prod_{n=1}^{N} p\left(\mathbf{X}_{d,n} \mid (\mathbf{SH})_{d,n},\sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{DN}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{d=1}^{D}\sum_{n=1}^{N}\left(\mathbf{X}_{d,n}-(\mathbf{SH})_{d,n}\right)^2\right\} \quad \text{(C.2)} \end{aligned}$$

is the Gaussian likelihood function; and the priors are expressed as normally distributed for $\mathbf{S}$:

$$\begin{aligned} p\left(\mathbf{S} \mid \boldsymbol{\theta}_S\right) &= \prod_{d=1}^{D}\prod_{k=1}^{K} p\left(\mathbf{S}_{d,k} \mid \mu_o,\sigma_o^2\right) \\ &= \left(2\pi\sigma_o^2\right)^{-DK/2} \exp\left\{-\frac{1}{2\sigma_o^2}\sum_{d=1}^{D}\sum_{k=1}^{K}\left(\mathbf{S}_{d,k}-\mu_o\right)^2\right\}; \quad \text{(C.3)} \end{aligned}$$

exponentially distributed for $\mathbf{H}$:

$$p\left(\mathbf{H} \mid \boldsymbol{\theta}_H\right) = \prod_{k=1}^{K}\prod_{n=1}^{N} p\left(\mathbf{H}_{k,n} \mid \lambda_o\right) = \lambda_o^{KN} \exp\left\{-\lambda_o \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{H}_{k,n}\right\}; \quad \text{(C.4)}$$

and as sampled from an inverse Gamma for σ^2:

$$p\left(\sigma^2 \mid \boldsymbol{\theta}_\sigma\right) = p\left(\sigma^2 \mid \alpha_o, \beta_o\right) = \frac{\beta_o^{\alpha_o}}{\Gamma(\alpha_o)}\left(\sigma^2\right)^{-\alpha_o-1} \exp\left\{-\frac{\beta_o}{\sigma^2}\right\}. \quad \text{(C.5)}$$

Notice that all prior distributions have been appropriately chosen to encode prior knowledge and to be conjugate priors of the likelihood.

Given that drawing a sequence of samples from the conditional posterior densities of the model parameters converges to the joint posterior, in the following sections, we derive each of the conditional posteriors in turn.

C.1 Conditional posterior density of S

In order to derive the conditional posterior density of $\mathbf{S}$, following the same procedure as in [172], we retrieve Eq. C.1 and get rid of those parameters that are independent from $\mathbf{S}$, hence reducing to the multiplication of Eqs. C.2 and C.3. Let $p(\mathbf{S}_{d,k} \mid \boldsymbol{\theta}_A) \overset{def}{=} p(\mathbf{S}_{d,k} \mid \mathbf{X}, \mathbf{S}_{\setminus(d,k)}, \mathbf{H}, \sigma^2)$, then:

$$\begin{aligned}
p(\mathbf{S}_{d,k} \mid \boldsymbol{\theta}_A) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\mathbf{X}_{d,n} - \sum_{k=1}^{K}\mathbf{S}_{d,k}\mathbf{H}_{k,n}\right)^2 - \frac{1}{2\sigma_o^2}\left(\mathbf{S}_{d,k} - \mu_o\right)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[\left(\mathbf{X}_{d,n} - \sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right) - \mathbf{S}_{d,k}\mathbf{H}_{k,n}\right]^2\right. \\
&\quad \left. -\frac{1}{2\sigma_o^2}\left[\mathbf{S}_{d,k} - \mu_o\right]^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[\left(\mathbf{X}_{d,n} - \sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)^2 + \mathbf{S}_{d,k}^2\mathbf{H}_{k,n}^2\right.\right. \\
&\quad \left.\left. -2\mathbf{S}_{d,k}\mathbf{H}_{k,n}\left(\mathbf{X}_{d,n} - \sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)\right]\right. \\
&\quad \left. -\frac{1}{2\sigma_o^2}\left[\mathbf{S}_{d,k}^2 + \mu_o^2 - 2\mathbf{S}_{d,k}\mu_o\right]^2\right\}
\end{aligned}$$

Factorising, we obtain:

$$\begin{aligned} p(\mathbf{S}_{d,k} \mid \boldsymbol{\theta}_A) \quad \propto \quad & \exp\left\{ -\frac{\mathbf{S}_{d,k}^2}{2}\left[\frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}{\sigma^2}+\frac{1}{\sigma_o^2}\right]\right. \\ & +\mathbf{S}_{d,k}\left[\frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}+\frac{\mu_o}{\sigma_o^2}\right] \\ & \left.-\frac{1}{2}\left[\frac{\sum_{n=1}^{N}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)^2}{\sigma^2}+\frac{\mu_o^2}{\sigma_o^2}\right]\right\} \end{aligned}$$

The resulting distribution of multiplying two Gaussians is another Gaussian. Therefore, our posterior should match the following form:

$$\begin{aligned} p(\mathbf{S}_{d,k} \mid \boldsymbol{\theta}_A) \quad \propto \quad & \exp\left\{-\frac{1}{2\sigma_p^2}\left(\mathbf{S}_{d,k}-\mu_p\right)^2\right\} = \exp\left\{-\frac{1}{2\sigma_p^2}\left(\mathbf{S}_{d,k}^2-2\mathbf{S}_{d,k}\mu_p+\mu_p^2\right)\right\} \\ \propto \quad & \exp\left\{-\frac{\mathbf{S}_{d,k}^2}{2}\left[\frac{1}{\sigma_p^2}\right]+\mathbf{S}_{d,k}\left[\frac{\mu_p}{\sigma_p^2}\right]-\frac{\mu_p^2}{2\sigma_p^2}\right\} \end{aligned}$$

We proceed by *completing the square*; matching the first term results in:

$$\begin{aligned} -\frac{\mathbf{S}_{d,k}^2}{2}\left[\frac{1}{\sigma_p^2}\right] &= -\frac{\mathbf{S}_{d,k}^2}{2}\left[\frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}{\sigma^2}+\frac{1}{\sigma_o^2}\right] \\ \frac{1}{\sigma_p^2} &= \frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}{\sigma^2}+\frac{1}{\sigma_o^2}=\frac{\sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}^2}{\sigma_o^2\sigma^2}+\frac{\sigma^2}{\sigma^2\sigma_o^2} \\ \frac{1}{\sigma_p^2} &= \frac{\sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}^2+\sigma^2}{\sigma_o^2\sigma^2} \\ \sigma_p^2 &= \frac{\sigma_o^2\sigma^2}{\sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}^2+\sigma^2}. \end{aligned} \tag{C.6}$$

Now, matching the second term leads to:

$$\begin{aligned} \mathbf{S}_{d,k}\left[\frac{\mu_p}{\sigma_p^2}\right] &= \mathbf{S}_{d,k}\left[\frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}+\frac{\mu_o}{\sigma_o^2}\right] \\ \frac{\mu_p}{\sigma_p^2} &= \frac{\sum_{n=1}^{N}\mathbf{H}_{k,n}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}+\frac{\mu_o}{\sigma_o^2} \\ \frac{\mu_p}{\sigma_p^2} &= \frac{\sigma_o^2\sum_{n=1}^{N}\mathbf{H}_{k,n}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)+\mu_o\sigma^2}{\sigma_o^2\sigma^2} \end{aligned}$$

Replacing σ_p^2 according to Eq. C.6:

$$\begin{aligned}
\mu_p &= \frac{\sigma_o^2 \sum_{n=1}^{N} \mathbf{H}_{k,n} \left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right) + \mu_o \sigma^2}{\sigma_o^2 \sigma^2} \times \frac{\sigma_o^2 \sigma^2}{\sigma_o^2 \sum_{n=1}^{N} \mathbf{H}_{k,n}^2 + \sigma^2} \\
\mu_p &= \frac{\sigma_o^2 \sum_{n=1}^{N} \mathbf{H}_{k,n} \left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right) + \mu_o \sigma^2}{\sigma_o^2 \sum_{n=1}^{N} \mathbf{H}_{k,n}^2 + \sigma^2} \\
\mu_p &= \sigma_p^2 \left[\frac{\sum_{n=1}^{N} \mathbf{H}_{k,n} \left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right)}{\sigma^2} + \frac{\mu_o}{\sigma_o^2} \right].
\end{aligned}$$

C.2 Conditional posterior density of H

The conditional posterior density of $\mathbf{H}$ will be derived following the same approach as in previous section: from Eq. C.1 we only keep the parameters directly related to $\mathbf{H}$, hence reducing to the multiplication of Eqs. C.2 and C.4. Let $p(\mathbf{H}_{k,n} \mid \boldsymbol{\theta}_B) \stackrel{def}{=} p(\mathbf{H}_{k,n} \mid \mathbf{X}, \mathbf{S}, \mathbf{H}_{\setminus(k,n)}, \sigma^2)$, then:

$$\begin{aligned}
p(\mathbf{H}_{k,n} \mid \boldsymbol{\theta}_B) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{d=1}^{D} \left(\mathbf{X}_{d,n} - \sum_{k=1}^{K} \mathbf{S}_{d,k} \mathbf{H}_{k,n} \right)^2 - \lambda_o \mathbf{H}_{k,n} \right\} \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{d=1}^{D} \left[\left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right) - \mathbf{S}_{d,k} \mathbf{H}_{k,n} \right]^2 \right. \\
&\quad \left. -\lambda_o \mathbf{H}_{k,n} \right\} \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{d=1}^{D} \left[\left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right)^2 + \mathbf{S}_{d,k}^2 \mathbf{H}_{k,n}^2 \right. \right. \\
&\quad \left. \left. -2\mathbf{S}_{d,k} \mathbf{H}_{k,n} \left(\mathbf{X}_{d,n} - \sum_{k' \neq k} \mathbf{S}_{d,k'} \mathbf{H}_{k',n} \right) \right] - \lambda_o \mathbf{H}_{k,n} \right\}
\end{aligned}$$

Factorising, we get:

$$p(\mathbf{H}_{k,n} \mid \boldsymbol{\theta}_B) \propto \exp\left\{ -\frac{\mathbf{H}_{k,n}^2}{2} \left[\frac{\sum_{d=1}^{D} \mathbf{S}_{d,k}^2}{\sigma^2} \right] \right.$$

$$+\mathbf{H}_{k,n}\left[\frac{\sum_{d=1}^{D}\mathbf{S}_{d,k}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}-\lambda_o\right]$$
$$-\frac{1}{2}\left[\frac{\sum_{d=1}^{D}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)^2}{\sigma^2}\right]\Bigg\}$$

The resulting distribution of multiplying a Gaussian by an exponential distribution is a rectified normal distribution of the form:

$$\begin{aligned}
p(\mathbf{H}_{k,n}\mid\boldsymbol{\theta}_B) &\propto \exp\left\{-\frac{1}{2\sigma_p^2}\left[\mathbf{H}_{k,n}-\mu_p\right]^2-\lambda_p\mathbf{H}_{k,n}\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_p^2}\left[\mathbf{H}_{k,n}^2-2\mathbf{H}_{k,n}\mu_p+\mu_p^2\right]-\lambda_p\mathbf{H}_{k,n}\right\}\\
&\propto \exp\left\{-\frac{\mathbf{H}_{k,n}^2}{2}\left(\frac{1}{\sigma_p^2}\right)+\mathbf{H}_{k,n}\left(\frac{\mu_p}{\sigma_p^2}-\lambda_p\right)-\frac{1}{2}\left(\frac{\mu_p^2}{\sigma_p^2}\right)\right\}
\end{aligned}$$

We *complete the square* by matching the first term:

$$\begin{aligned}
-\frac{\mathbf{H}_{k,n}^2}{2}\left(\frac{1}{\sigma_p^2}\right) &= -\frac{\mathbf{H}_{k,n}^2}{2}\left(\frac{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2}{\sigma^2}\right)\\
\sigma_p^2 &= \frac{\sigma^2}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2}
\end{aligned}\tag{C.7}$$

Matching the second term:

$$\mathbf{H}_{k,n}\left(\frac{\mu_p}{\sigma_p^2}-\lambda_p\right)=\mathbf{H}_{k,n}\left[\frac{\sum_{d=1}^{D}\mathbf{S}_{d,k}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}-\lambda_o\right]$$

Assuming $\lambda_p=\lambda_o$ and replacing σ_p^2 according to Eq. C.7:

$$\begin{aligned}
\mu_p &= \frac{\sum_{d=1}^{D}\mathbf{S}_{d,k}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sigma^2}\times\frac{\sigma^2}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2}\\
\mu_p &= \frac{\sum_{d=1}^{D}\mathbf{S}_{d,k}\left(\mathbf{X}_{d,n}-\sum_{k'\neq k}\mathbf{S}_{d,k'}\mathbf{H}_{k',n}\right)}{\sum_{d=1}^{D}\mathbf{S}_{d,k}^2}
\end{aligned}$$

C.3 Conditional posterior density of σ^2

Finally, the conditional posterior density of σ^2 will be derived according to [173]: starting from Eq. C.1, we keep the parameters directly related to σ^2;

which reduces the aforementioned equation to the product of Eqs. C.2 and C.5. Let $p(\sigma^2 \mid \boldsymbol{\theta}_C) \stackrel{def}{=} p(\sigma^2 \mid \mathbf{X}, \mathbf{S}, \mathbf{H})$, then:

$$\begin{aligned} p(\sigma^2 \mid \boldsymbol{\theta}_C) &\propto \left(2\pi\sigma^2\right)^{-DN/2} \times \frac{\beta_o^{\alpha_o}}{\Gamma(\alpha_o)} \left(\sigma^2\right)^{-\alpha_o-1} \\ &\quad \exp\left\{-\frac{1}{2\sigma^2}\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n}\right]^2 - \frac{\beta_o}{\sigma^2}\right\} \\ &\propto \frac{\beta_o^{\alpha_o}}{2\pi\Gamma(\alpha_o)} \times \frac{(\sigma^2)^{-\alpha_o-1}}{\sigma^{DN}} \\ &\quad \exp\left\{-\frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n}\right]^2}{2\sigma^2} - \frac{\beta_o}{\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-\left(\alpha_o + \frac{DN}{2}\right)-1} \\ &\quad \exp\left\{-\frac{\beta_o + \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n}\right]^2}{\sigma^2}\right\} \end{aligned} \quad \text{(C.8)}$$

The resulting distribution of multiplying a Gaussian by an inverse Gamma distribution results to another inverse Gamma, of the form:

$$p(\sigma^2 \mid \boldsymbol{\theta}_C) \propto \left(\sigma^2\right)^{-\alpha_p-1} \exp\left\{-\frac{\beta_p}{\sigma^2}\right\}$$

Matching the parameters with those in Eq. C.8, we obtain:

$$\begin{aligned} \alpha_p &= \alpha_o + \frac{DN}{2}; \\ \beta_p &= \beta_o + \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}\left[\mathbf{X}_{d,n} - (\mathbf{SH})_{d,n}\right]^2. \end{aligned}$$