2016-01-16

LFDA: A Probabilistic Graphical Model for the Study of Excitation Emission Matrices

Oscar Luis Martinez
University of Miami, oscarlmartineza@gmail.com

UNIVERSITY OF MIAMI

LFDA: A PROBABILISTIC GRAPHICAL MODEL FOR THE STUDY OF EXCITATION EMISSION MATRICES

By

Oscar Luis Martinez

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

May 2016

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

LFDA: A PROBABILISTIC GRAPHICAL MODEL FOR THE STUDY OF EXCITATION EMISSION MATRICES

Oscar Luis Martinez

Approved:

Miroslav Kubat, Ph.D.
Associate Professor of Electrical and Computer Engineering

Mei-Ling Shyu, Ph.D.
Professor and Associate Chair of Electrical and Computer Engineering

Kamal Premaratne, Ph.D.
Victor P Clarke Professor of Electrical and Computer Engineering

Dean of the Graduate School

Xiaodong Cai, Ph.D.
Professor of Electrical and Computer Engineering

James N. Wilson, Ph.D.
Associate Professor of Chemistry

MARTINEZ, OSCAR LUIS
LFDA: A Probabilistic Graphical Model for the Study of Excitation Emission Matrices

(Ph.D., Electrical and Computer Engineering)
(May 2016)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Miroslav Kubat.
No. of pages in text. (108)

Traditional classification techniques assume samples are described by vectors of features. However, in some domains samples are gathered by measuring a variable with respect to two or more other variables: for a given value of x and y measure z. In such domains, samples are more naturally described by matrices or by higher dimensional arrays.

We present a novel latent Dirichlet allocation (LDA)-based approach for modeling and analyzing fluorescent spectroscopy excitation-emission Matrices (EEMs) and other three way datasets. We introduce parallels between topic modeling and three-way arrays which allow us to create adaptations to use LDA-based methods in latent fluorophore studies. The proposed framework views the EEMs as being generated from an underlying hidden pool of flourophore compounds, and provides a latent flourophore-space representation of an EEM. We show that this LDA-based model can increase classification performance, especially when paired with parallel factor analysis (PARAFAC) which may be regarded as perhaps the most popular and widely used tool for dealing with EEMs. Our experiments show that the proposed LDA-based algorithm is in some cases more robust than PARAFAC to certain types of noise and data disturbances. We also observe that pairing this LDA-based method with PARAFAC leads to an improvement in classification performance and to added robustness at high peak-signal-to-noise-ration (PSNR) values.

We also present an extended graphical model that incorporates the effect of *outside* variables that may affect fluorescent expression of certain compounds. The extended model

offers further insight into the interaction between these variables and the latent fluorophore components while facilitating the model building process.

The performance of machine learning algorithms is known to be impaired if the representation of the individual classes in the training set is imbalanced, i.e., one class outnumbering the other class(es). Such is the case for several experiments in this proposal. Many approaches to deal with this problem have been developed, none of them totally satisfactory. Here we propose *membership-based minority oversampling (MeMO),* as yet another possible solution, and explores, experimentally, the conditions under which it outperforms earlier attempts.

Finally we introduce a Dempster-Shafer based fusion model that is intended to adaptively merge the PARAFAC and LDA-based models when their outputs are being used for classification purposes.

Contents

List of Figures vi

List of Tables viii

CHAPTER 1 Introduction 1

1.1 Motivation 2

1.2 Excitation Emission Matrices (EEMs) 3

1.3 Multi-way Analysis 7

1.3.1 The PARAFAC Model 8

1.3.2 Fitting a PARAFAC Model 10

CHAPTER 2 Probabilistic Modeling 14

2.1 Probabilistic Graphical Models 14

2.2 Inference 18

2.2.1 Exponential Family 18

2.2.2 Conjugate Priors 19

2.2.3 Variational Methods 20

2.2.4 MCMC and Gibbs Sampling 21

CHAPTER 3 The Proposed Approach 24

3.1 LDA for EEMs 24

3.2 Generative Process 27

3.3 Parameter Estimation 30
3.3.1 Sampling the Conditional Distribution 30
3.3.2 MCMC Gibbs Sampling Algorithm 31
3.3.3 Gibbs Sampling Derivation 33
3.4 Latent Fluorescent Dirichlet Allocation (LFDA) 38
3.4.1 Model Definition 38
3.4.2 Model Derivation 39
3.5 Classification Enhancement Technique (CET) 47

CHAPTER 4 Minority Oversampling **50**
4.1 Previous Work 51
4.1.1 Random Under-Sampling and Oversampling Methods 51
4.1.2 Clustering Methods 52
4.1.3 SMOTE : Synthetic Minority Oversampling Technique 53
4.1.4 Data Cleaning Methods 54
4.2 The Proposed Approach: MeMO 54
4.2.1 The Basic Concept 54
4.2.2 Example 56

CHAPTER 5 Non-parametric Regression **59**
5.1 Overview 60
5.2 Non-parametric Regression for Microtox Prediction 62

CHAPTER 6 Fusion-based Classification **63**
6.1 Dempster-Shafer Theory 63
6.1.1 Belief and Plausability 65
6.1.2 Evidence Combination 65
6.2 DS-Model 67

CHAPTER 7 Results **70**

7.1 MeMO 70

7.1.1 Testbeds 70

7.1.2 Classification Results 71

7.2 Classification Using PARAFAC and LDA 75

7.2.1 Performance Criteria 75

7.2.2 Testbeds 77

7.2.3 Results 81

7.3 Classification Using PARAFAC and LFDA 88

7.4 Regression Using PARAFAC and LFDA 88

7.4.1 Microtox® Assessment Process 90

7.4.2 DS Performance Measures 96

7.4.3 LFDA Drawbacks 99

CHAPTER 8 Conclusion and Future Work **100**

8.1 Conclusion 100

8.2 Future Work 101

Bibliography **102**

List of Figures

1.1 EEM showing zeroth and first order scatter as prominent diagonal lines. . . 6

1.2 EEM after removal of zeroth and first order scatter. 6

1.3 A three-way dataset where the matrix samples are stacked on top of each other. 8

1.4 The matricizing process. 11

2.1 Graphical model representation of nodes and edges. 15

2.2 Plate notation. 16

2.3 Graphical model: mixture of Gaussians 17

3.1 EEM discretization. 26

3.2 The LDA graphical model. 28

3.3 The LFDA graphical model. 39

3.4 Ground truth structure of 4 fluorophores. 45

3.5 Samples generated from the underlying structure. 46

3.6 48

4.1 Point distribution for initial synthetic dataset. 57

4.2 Point distribution after minority class oversampling. 58

6.1 Plot of the function β with $\sigma = 5$. 68

7.1 ROC curve for Breast Cancer Wisconsin dataset (UCI) 72

7.2 ROC curve for Pima Indians Diabetes dataset (UCI) 73
7.3 ROC curve for LandSat dataset (UCI) 74
7.4 ROC curve for synthetic dataset 76
7.5 An EEM sample subjected to different types of noise. 80
7.6 F_1 vs. PSNR for fluorophore dataset 85
7.7 F1 vs. PSNR for synthetic dataset. 87
7.8 ROC curves with variables $\mathbf{x}_{LDA}$. 92
7.9 ROC curves with variables $\mathbf{x}_{PA}$. 94
7.10 ROC curves with variables $\mathbf{x}_{LFDA}$. 95

List of Tables

2.1 Common Exponential Family Distributions 19

7.1 Test beds 71

7.2 Area under the curve of ROC curves for different degree of imbalances with *MeMO* and *SMOTE*. 75

7.3 Micro Averaging and Macro averaging of Precision and Recall 77

7.4 Description of Data Sets 78

7.5 Major Fluorescent Components 78

7.6 LDA F_1 Prediction performance at different levels of imbalance 82

7.7 LDA F_1 Prediction performance at different levels of imbalance 82

7.8 LDA F_1 Prediction performance at different levels of imbalance 82

7.9 Cancer Dataset: Performance Values 83

7.10 Fluorophore Dataset: Performance values 84

7.11 Synthetic Dataset: Performance Values Under Noise Created by Shifting Peak Locations 88

7.12 Synthetic Dataset: Performance Values Under Type 3 Noise 89

7.13 Water Dataset: Performance Values 97

7.14 Friedman Nemenyi Post-hoc Test 98

7.15 AUC for NP regression 99

Chapter 1

Introduction

Although machine learning applications typically employ samples, which are described attributes expressed in vector form, in some domains the samples are more naturally expressed as matrices. This is the case of fluorescence spectroscopy where each example is an excitation-emission matrix (EEM) with columns representing excitation wavelengths, rows representing emission wavelengths, and each excitation-emission pair containing the fluorescence corresponding to both.

Representing the matrices as vectors allows the use of traditional machine learning algorithms for their classification and study. On the surface, this seems to be quite easy: simply concatenate the rows of the matrix, thus obtaining a vector of $n \cdot m$ attributes, where n and m are the numbers of rows and columns, respectively. This, however, will likely obscure critically important information about spectral interrelationships. Furthermore, in chemomtrics and in psychometrics, where matrix-like samples are common, the objective of the analysis is at times more subtle than simple classification. In certain applications it is more important to understand the underlying structure of the data than to classify samples according to their attributes.

In an attempt to overcome this weakness, some authors prefer to decompose their matrix datasets by the use of parallel factor analysis (PARAFAC) [1, 2, 3, 4]. This approach allows the analyst to represent the entire data set in terms of two loading matrices and one

score matrix (plus residuals). The rows of the score matrix offer a representation of the samples in the original dataset in terms of the two loading matrices. This approach allows the analyst to understand the underlying structure of the data, represented in the loading matrices, but it also offers a vector description of each sample, represented by the score matrix. Nevertheless, this tool suffers from high sensitivity to outliers, forcing analysts to discard samples which might offer other insight into the data.

Seeking an improvement, we developed a method that uses PARAFAC in collaboration with a probabilistic graphical model based on latent Dirichlet allocation (LDA). Our graphical model allows direct comparisons with studies in literature given that it also infers the underlying or *hidden* structure of the data and the model is human readable. Furthermore, the graphical model is capable of incorporating external variables that can affect the underlying structure of each the matrix. This last advantage can help the analyst understand not only the hidden structure of the data but also how this structure is affected by external stimuli.

This dissertation is ordered in the following way: Chapter 1 introduces excitation-emission matrices and multi-way analysis, chapter 2 explains graphical models and parameter estimation methods used in the field, chapter 3 describes the proposed graphical model methods to study EEMs and addresses the matrix quantization problem, chapter 4 introduces a new oversampling techniques used for dataset imalances that can be used in multi-way arrays, chapter 5 describes Non-Parametric regression and its advantages over traditional approches, chapter 6 explains Dempster-Shafer theory and belief fusion, chapter 7 contains experimental results for the proposed methods and finally, chapter 8 summarizes the results and proposes possible avenues for future work.

1.1 Motivation

Many publications and studies in the area of fluorescence spectroscopy are based on decomposition methods that offer insight about the family of fluorophores present in a sam-

ple. Decomposition methods like PARAFAC are human readable and offer an easily interpretable output that can be shared and understood across studies and research groups. Human readable decomposition outputs show the analyst the basic fluorescent groups present in a set of measurements, allowing an analyst to quickly determine the type of compounds related to said fluorescent groups. Techniques based on neural networks or in direct prediction from the spectra do not offer a human readable output and their results cannot be compared to other studies in literature. Thus, a technique that can offer a human interpretable output, while also being backward compatible to the PARAFAC decomposition, is highly desirable.

The PARAFAC model, calculated through an alternating least square algorithm, is known to have a fit that is very sensitive to outliers. This disadvantage can add time consuming model validation procedures to the model building process. A model that can more adequately handle noise and outliers would reduce the time necessary for model validation and simplify the classification task.

Three-way datasets that do not follow a tri-linear pattern cannot appropriately be modeled using PARAFAC. It would be advantageous for a model to have the ability to fit data which follows a tri-linear pattern while also offering the flexibility to model datasets that do not. In several applications such as chemometrics, the tri-linear nature of the data can sometimes be affected by external parameters such as pH or temperature. The PARAFAC model has no direct way of handling this variations and therefore falls short when the data acquisition was not carefully calibrated to account for such variations. A model that can incorporate these external variables could offer a deeper understanding of their effect on the data while also simplifying the model building process.

1.2 Excitation Emission Matrices (EEMs)

An EEM is a matrix obtained through the concatenation of several fluorescence emission scans collected at periodical excitation wavelengths. It can be visualized as a three-

dimensional surface where each point has a location given by an excitation-emission pair and a height specified by a corresponding fluorescent value. The peak locations correspond to types of fluorescent substances and their heights correspond to their concentrations.

EEMs have been used to characterize dissolved organic matter (DOM) in water and soil [5, 6, 7], to study water fingerprinting [2, 8, 9] and, in general, to study fluorescent organic substances. Fluorescence spectroscopy can measure concentrations down to parts per billion [10], offering excellent sensitivity, simplicity and low cost as compared to other analytical techniques.

In metabonomic diagnostics, the response of metabolites in biological systems is measured and used to predict disease. Over the past three decades the use of EEMs as a metabonomic tool to analyse human blood plasma has been well documented [11, 12, 13]. Leiner *et al.* [14] used deviations of tryptophan fluorescence of human blood serum to detect gynecological malignancies. Madhuri *et al.* performed several studies using fluorescence spectroscopic data of blood plasma to detect liver decease [15], oral malignancies [16] and cancer [12, 17]. Most of these studies used maximum valued excitation-emission pairs or ratio variables to describe peak locations. The feature extraction was therefore performed in a supervised manner through expert analysis of spectral regions known for their fluorescent composition. The location of selected peaks was subsequently used to perform linear discriminant analysis or to establish a correlation between peak locations and class labels. Although informative, these approaches require expert domain knowledge and use only a few wavelength pairs, while ignoring the full excitation-emission spectrum.

Recently, researchers have begun to use the whole spectral approach taking advantage of the information of the entire EEM by introducing modern chemometric techniques for data analysis. Lawaetz *et al.*[18] used EEM data on human blood plasma to detect colorectal cancer and Hudson *et al.*[6] used EEM data for organic matter characterization. These

studies applied parallel factor analysis (PARAFAC) to factorize the three-way EEM data and to detect the chemical differences between samples.

As mentioned before, the PARAFAC model is highly sensitive to outliers and noise. Consequently, an alternative approach to replace or supplement PARAFAC as a fluorescent decomposition and pattern recognition method is highly desirable. In this paper we propose a novel approach, using a probabilistic graphical model based on latent Dirichlet allocation to transform the EEMs fluorescent information into an underlying fluorophore-like space. Section 1.3 presents traditional multi-way analysis methods and introduces the PARAFAC model, which is currently the state of the art technique used to study EEMs.

EEM measurements are usually pretreated for fluorescent artifacts before analysis. When taking fluorescent measurements, small particles in the samples cause light to reflect and deviate from its path producing what is known as scattering. This phenomenon produces distinctive fluorescent patterns in the samples that are completely unrelated to their actual fluorescent properties. Rayleigh (elastic) and Raman (inelastic) scattering peaks, along with their corresponding harmonic reflections, occur in all samples, even those not exhibiting fluorescence[19]. The scattering forms distinctive diagonal lines across the fluorescent landscape as can be seen in fig. 1.1.

Elastic scattering has no energy loss; thus, the scattered excitation wavelength is identical to the detected emission wavelength. Inelastic scattering has some energy loss; therefore, the detected emission wavelength is slightly longer than the original excitation wavelength. The zeroth harmonics of these scattering peaks can be seen as a diagonal lines that pass through the matrix around the values where *emission wavelength = excitation wavelength*. The first harmonics occur around the area of the matrix where *emission wavelength = 2*excitation wavelength*.

The intensity of the scattering will vary depending on solvent type and on the concentration of particles in the solution [19]. Therefore, it is common practice to remove the

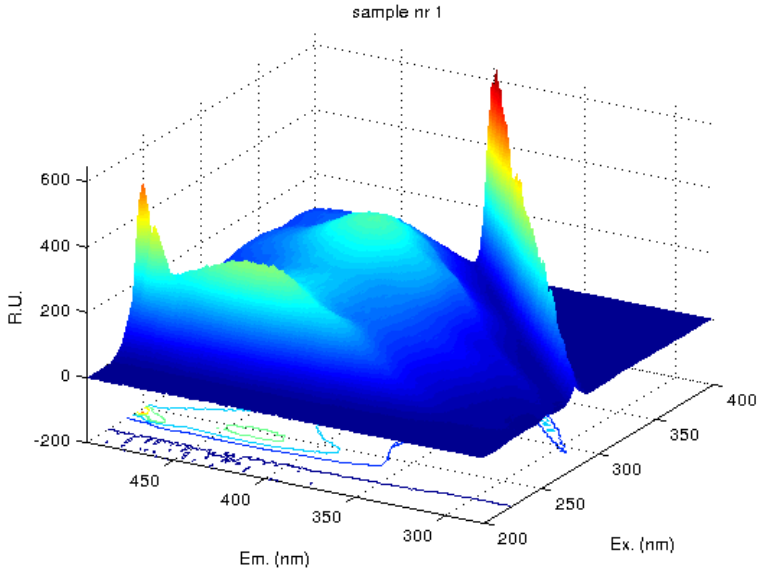


Figure 1.1: EEM showing zeroth and first order scatter as prominent diagonal lines which usually have higher magnitude than other fluorescent features.

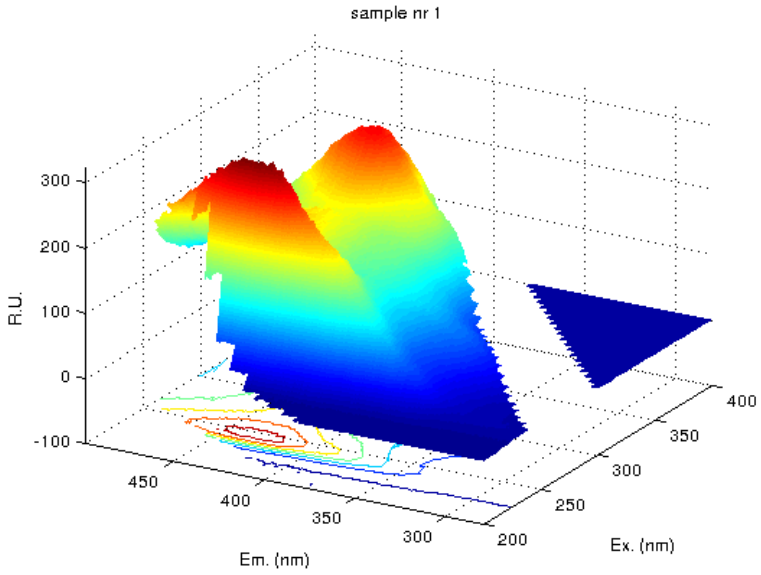


Figure 1.2: EEM after removal of zeroth and first order scatter. Note that removed entries are set to missing. Fluorescent features that were obscured by scattering become more evident.

scattering from the matrices by setting the matrix entries in those areas to missing as shown in fig. 1.2.

1.3 Multi-way Analysis

Multi-way datasets are comprised of data samples having groups of variables measured in a crossed fashion [20]. Two-way data describes a traditional dataset that can be organized into a matrix (two-dimensional structure), where the rows are the samples and the columns are the measured variables. Three-way data describes a dataset that can be organized into a cube (three-dimensional structure), where the measured variables are entries in a matrix and the samples are slices of the cube. In the same way, n-way data refers to a dataset that can be organized into a hypercube (n-dimensional structure). In psychometrics this type of data can be created by measuring a set of variables on a group of patients at different time points. In chemometrics the clearest example of a three-way dataset is given by EEMs, but two-way data measured with different control variables (time, pH or location) can also be expressed as an n-way array.

Parallel factor analysis (PARAFAC) and other decomposition methods such as Tucker3 and two-way PCA have been the traditional methods of choice to analyze the multi-way nature of EEM data in chemometrics. Given the correct number of factors or components, PARAFAC scores represent relative concentrations while the loadings represent emission and excitation spectra of the fluorophores in a sample [21].

Let $\mathbf{X} \in \Re^{D \times R \times C}$ denote a three-way array and let x_{drc} be an entry of $\mathbf{X}$ in the d^{th} sample, at the r^{th} emission wavelength and at the c^{th} excitation wavelength. Such an array can be analysed using two-way analysis methods by taking two-way data *"slices"* and concatenating them into a new array $\mathbf{Y} \in \Re^{D \times RC}$. However, such an approach cannot properly capture the underlying structure of the high-dimensional array because two-way analysis methods suffer from rotational freedom [22]. Furthermore, two-way PCA methods tend to use the excess degrees of freedom to model noise or model the systematic variation redun-

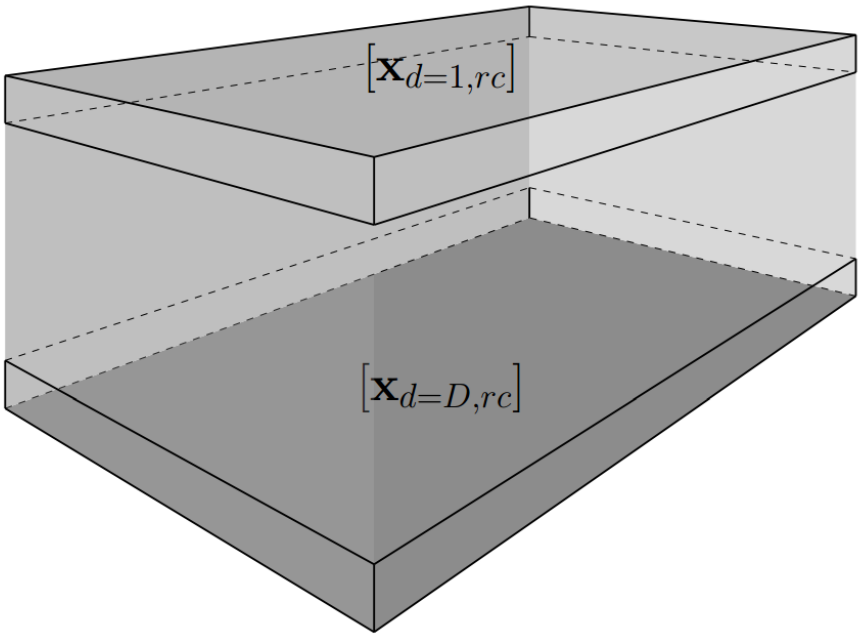


Figure 1.3: A three-way dataset where the matrix samples $\mathbf{x}_{d=1,rc}, \mathbf{x}_{d=2,rc}, \cdots, \mathbf{x}_{d=D,rc}$ are stacked on top of each other. A measured value at location (r,c) will be at the same location for any sample $k \in \{1, \cdots, D\}$.

dantly [20]. This limitation is the main reason for the popularity of generalized *multiway data analysis techniques* in high-dimensional array decomposition.

Leyard Tucker proposed an extension of bilinear factor analysis to higher-order datasets [23], which led to a series of multi-way analysis models. This made Tucker one of the initial contributors to "modern" multi-way analysis techniques, later called *Turker models* or *N-mode principal component analysis*.

1.3.1 The PARAFAC Model

Among the current methods for multi-way analysis, the most popular model proposed in [24] is Parallel factor analysis (PARAFAC). The PARAFAC model has been widely used in chemometrics for EEM analysis[1]. Like *Tucker3*, this is also an extension of bilinear factor models to multilinear data. Adopting ideas of Parallel Proportional Profiles [26], the model

[1] A similar idea has been proposed by Carroll & Chang [25] which they called Canonical Decomposition (CANDECOMP)

assumes that samples with same factors, but under different conditions, will have different scaling depending only on the conditions.

The The T-component PARAFAC model for our three way data array $\mathbf{X}$, is a decomposition of $\mathbf{X}$ expressed as a linear combination of rank-1 tensors:

$$\mathbf{X} = \sum_{t=1}^{T} \mathbf{u}_t \circ \mathbf{v}_t \circ \mathbf{w}_t + \mathbf{E}, \tag{1.1}$$

where $\mathbf{u}_t = \{u_{dt}\}$, $\mathbf{v}_t = \{v_{rt}\}$ and $\mathbf{w}_t = \{w_{ct}\}$ indicate the t-th column of component matrices $\mathbf{U} \in \mathbb{R}^{D\times T}$, $\mathbf{V} \in \mathbb{R}^{R\times T}$, and $\mathbf{W} \in \mathbb{R}^{C\times T}$, respectively; $\mathbf{E} = \{e_{drc}\} \in \mathbb{R}^{D\times R\times C}$ represent the 'residuals'. The symbol $\circ$ denotes the vector outer product: $\mathbf{u}_t \circ \mathbf{v}_t \circ \mathbf{w}_t$ generates the matrix $\mathbf{Y} = \{y_{drc}\} \in \mathbb{R}^{D\times R\times C}$, where $y_{drc} = \sum_t u_{dt} v_{rt} w_{ct}$. Thus, eq. (1.1) can be expressed for a single element of $\mathbf{X}$ as

$$x_{drc} = \sum_{t=1}^{T} u_{dt} v_{rt} w_{ct} + e_{drc}, \tag{1.2}$$

where u_{dt}, v_{rt}, and w_{ct} are referred to as the *core-values.* These core-values in three-way analysis can be compared to the singular values of two-way analysis that one gets via the singular value decomposition (SVD). The core-values g_{ttt} are explicitly shown in eq. (1.3) as scaling parameters, such that the *core-array* is a diagonal three-way array $\mathbf{G} \in \mathbb{R}^{T\times T\times T}$, with diagonal entries $g_{111}, \ldots, g_{TTT}$:

$$x_{drc} = \sum_{t=1}^{T} g_{ttt} \grave{u}_{dt} \grave{v}_{rt} \grave{w}_{ct} + e_{drc}. \tag{1.3}$$

Here, we use $\grave{\mathbf{u}}_t = \{\grave{u}_{dt}\}$, $\grave{\mathbf{v}}_t = \{\grave{v}_{rt}\}$ and $\grave{\mathbf{w}}_t = \{\grave{w}_{ct}\}$ to differentiate the notation from the components in eq. (1.2). Other than that, there is no difference between the two sets of variables.

The aim in PARAFAC is to find unique estimates for the $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ component matrices (up to permutation, sign and scaling indeterminacy)[21]. This is the reason PARAFAC does not have rotational freedom; the calculated component matrices in a PARAFAC model cannot be changed without changing the residuals $\mathbf{E}$. This property is also known as the *in-*

trinsic axes property. Thus, unlike Principal Component Analysis (PCA), PARAFAC provides *unique* basis vector orientations. This property makes PARAFAC a popular choice, especially in the chemometrics community. If we pick the correct number of factors T, a PARAFAC analysis provides a unique model for three-way EEM data such that the model's factors correspond directly to chemical mixture measurements. However, determining the optimal number of factors for real data is not trivial and different techniques like core consistency [18] must be used to optimize the analysis.

Like PARAFAC, Tucker models also decompose $\mathbf{X}$ into three component matrices $\mathbf{U} \in \Re^{D\times P}$, $\mathbf{V} \in \Re^{R\times Q}$ and $\mathbf{W} \in \Re^{C\times T}$. Using the same notation, eq. (1.4) shows the factorization with a *Tucker3 model*.

$$x_{drc} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{t=1}^{T} g_{pqt}u_{dp}v_{rq}w_{ct} + e_{drc} \tag{1.4}$$

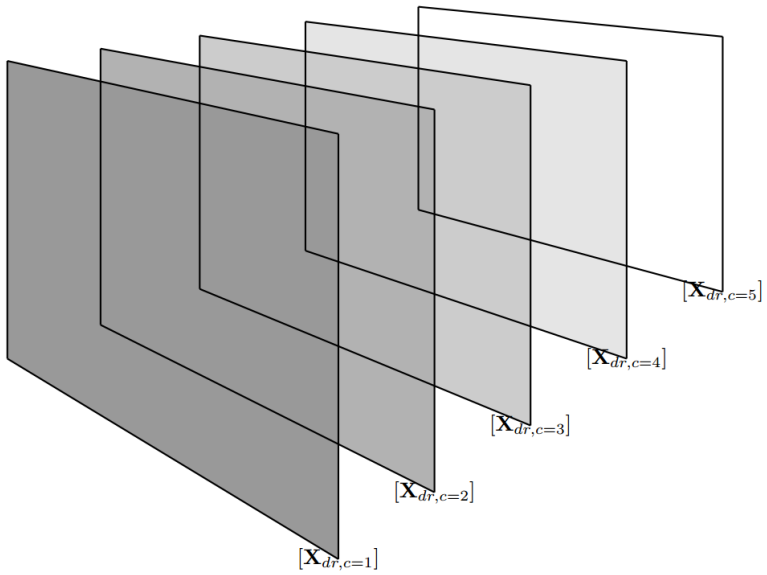
The main difference in *Tucker3 models* resides in the *core-array* $\mathbf{G} \in \Re^{P\times Q\times T}$, with elements g_{pqt}, as opposed to the diagonal core-array of PARAFAC. This makes the *Tucker3 model* more flexible as it allows interaction between *factors*. This appears to be a good feature when analyzing three-way datasets. Nonetheless, this is the reason PARAFAC is a more popular model in the chemometric community. The flexibility added by the filled elements g_{pqt} limits the *Tucker model's* ability to provide unique component matrices. Consequently, the relationship between factors and chemical measurements in a *Tucker model* is complicated and difficult to interpret, unlike the direct and intuitive relationship that results from the PARAFAC model.

1.3.2 Fitting a PARAFAC Model

The PRAFAC model can also be written using the Khatri-Rao product as follows:

$$\mathbf{X}_{(D\times RC)} = \mathbf{U}(\mathbf{V}\odot\mathbf{W})' + \mathbf{E} \tag{1.5}$$

where $\mathbf{X}_{(D\times RC)}$ is the unfolded array of size $D \times RC$, otherwise known as the matricization of the tree-way array in the first mode. An example of this process is shown in fig. 1.4. This notation simplifies the notation in the algorithm formulation.



(a) Array slices along the third mode of a three-way array $\mathbf{X}$ with dimension $C = 5$

$[\mathbf{X}_{dr,c=1}]$	$[\mathbf{X}_{dr,c=2}]$	$[\mathbf{X}_{dr,c=3}]$	$[\mathbf{X}_{dr,c=4}]$	$[\mathbf{X}_{dr,c=5}]$

(b) Matricizing X along the first mode.

Figure 1.4: The matricizing process in the first mode creates a matrix $\mathbf{X}_{(D\times RC)} \in \Re^{D\times RC}$, the same process in the second mode creates a matrix $\mathbf{X}_{(R\times DC)} \in \Re^{R\times DC}$ and in the third mode a matrix $\mathbf{X}_{(C\times RD)} \in \Re^{C\times RD}$.

Most algorithms used to fit PARAFAC models are based on alternating least squares, although some faster algorithms are based on a generalized eigenvalue problem to calculate an approximate solution [21]. Here, we describe an alternating least squares approach which is easy to implement and guarantees convergence.

Using the notation in eq. (1.5) a least squares loss function $\mathscr{L}$ can be written as:

$$\mathscr{L} = \min_{\mathbf{U},\mathbf{V},\mathbf{W}} \|\mathbf{X}_{(D\times RC)} - \mathbf{U}(\mathbf{V}\odot\mathbf{W})'\|^2 \tag{1.6}$$

The ALS algorithm finds an estimate for **U** given **V** and **W**, then an estimate for **V** given **U** and **W** and an estimate for **W** given **U** and **V**. The symmetry of the model allows the updates to be formulated in the same way for the three modes by simply shifting the role of each matrix [21]. In order to formulate the problem, let us define three different ways in which the three-way array **X** will be matricized depending on which estimate $\mathbf{U}, \mathbf{V}$ or **W** is being performed:

$$\mathbf{X}_{mode(\mathbf{B})} = \begin{cases} \mathbf{X}_{(D\times RC)}, & \text{if } \mathbf{B} = \mathbf{U} \\ \mathbf{X}_{(C\times RD)}, & \text{if } \mathbf{B} = \mathbf{W} \\ \mathbf{X}_{(R\times DC)}, & \text{if } \mathbf{B} = \mathbf{V} \end{cases} \tag{1.7}$$

With this definition in place we can rewrite eq. (1.6) in more general terms as:

$$\mathscr{L}_B = \min_{\mathbf{B}} \|\mathbf{X}_{mode(\mathbf{B})} - \mathbf{B}(\mathbf{Z})'\|^2 \tag{1.8}$$

where, **B** takes values **W**, **U** or **V**, and **Z** is given by the following equation:

$$\mathbf{Z} = \begin{cases} \mathbf{V}\odot\mathbf{W}, & \text{if } \mathbf{B} = \mathbf{U} \\ \mathbf{V}\odot\mathbf{U}, & \text{if } \mathbf{B} = \mathbf{W} \\ \mathbf{W}\odot\mathbf{U}, & \text{if } \mathbf{B} = \mathbf{V} \end{cases} \tag{1.9}$$

The solution to eq. (1.8) when **B** has full rank is:

$$\mathbf{B} = \mathbf{X}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \tag{1.10}$$

Thus, the algorithm can be summarized with the following steps

PARAFAC($\mathbf{X}$)

1 $\triangleright$ Initialize $\mathbf{V}$ and $\mathbf{W}$

2 **repeat**

3 **for** $\mathbf{B} \in \{\mathbf{U}, \mathbf{W}, \mathbf{V}\}$

4 **do**

5 $\mathbf{Z} \leftarrow (\mathbf{V} \odot \mathbf{W})$

6 $\mathbf{B} \leftarrow \mathbf{X}_{mode(\mathbf{B})}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$

7 **until** Change in fit is small.

As reported in [20], good initialization guesses for $\mathbf{V}$ and $\mathbf{W}$ can ensure that no local minimum is found. A variety of techniques to speed up convergence and avoid local minima have been proposed in the literature. Some authors [27, 28, 29] propose using approximate solutions obtained from generalized rank annhilation methods or from direct trilinear decomposition, where two samples are used to estimate the modes. On the other hand, Harshman *et al.* [30] propose running several PARAFAC models with random initial values so that repeated convergence to the same model minimizes the chance of finding a local minimum.

Finally, the best number of components for a PARAFAC model can be determined by cross-validation, by performing a core consistency diagnosis (CONCORDIA) [18] or by analysing the residuals of the model. Systematic variation of the residuals tends to indicate that more components are needed to represent the variation in the data. Thus, several models with an increasing number of components are created while studying the behavior of the residuals. If a plot of the residual sum of squares versus the number of components abruptly flattens out for a certain number of components, it is highly probable that the true number of components has been found [20].

Chapter 2

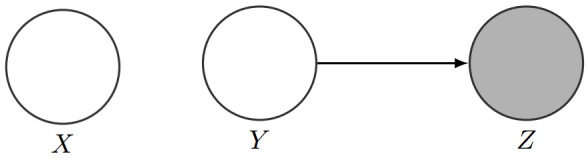
Probabilistic Modeling

The main purpose behind probabilistic modeling is to define, infer and use a statistical model to represent a real world process. Probabilistic models vary greatly depending on their purpose. Therefore, the number, the type and the model's parameters is tailored to represent a specific problem domain.

In order to use a statistical model we must find efficient ways to perform inference and we must also find ways to use such a model to make predictions. Solving the inference problem allows us to find the parameters that best explain the observations. On the other hand, solving the prediction problem allows us to use the model with predetermined parameters to make predictions on new data.

2.1 Probabilistic Graphical Models

Graphical models (GMs) are illustrative diagrams that represent probability distributions. The design and the interpretation of complex probabilistic interactions can be greatly simplified when their GMs are used to represent them. Using GMs can help avoid the countless and tedious algebraic manipulations needed to formulate models manually, while simultaneously, they can offer insights into the model itself by simple graph inspection. As defined in [31] a GM can be described as a group of probability distributions factorizing according to the structure of the underlying graph.



(a) Latent variable X.

(b) Observed variable Z dependent on latent variable Y.

Figure 2.1: In a graphical model nodes represent random variables, edges represent possible dependecies.

In a graphical model, clear nodes represent latent random variables, shaded nodes represent observed random variables and edges represent possible dependencies between variables.[1] In order to enhance readability, plates are used to show substructure replication and denote the replication factor in the lower right corner of the plate.

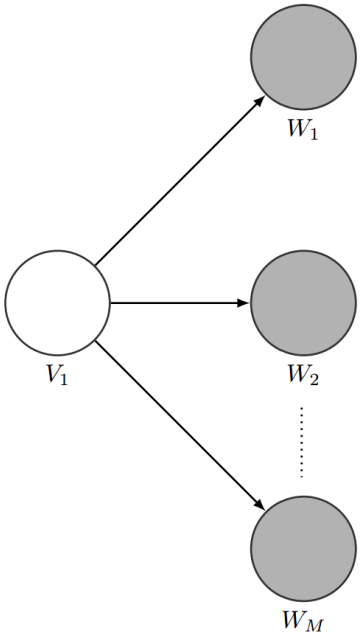
In the following discussion, we will limit our scope to *directed acyclic graphs* (DAGs). A graph $G = (V,E)$ is formed by a set of vertices or nodes V and a set of edges E, where each node $v \in V$ is associated with a random variable x_v distributed according to a probability distribution p_v. Edges in a DAG are directed from parent to child such that the factorization of the graph can be written as follows:

$$p(\mathbf{x}) = \prod_{v \in V} p_v(x_v | x_{i \in parents(v)}) \tag{2.1}$$

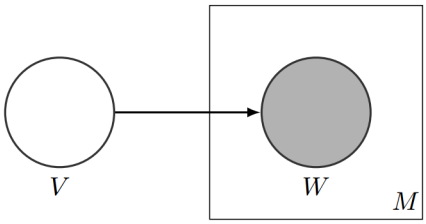
where $parents(v)$ denotes the set of nodes that have a directed edge pointing at v and $p(\mathbf{x})$ is the joint distribution of all variables.

Graphical models can be used to describe latent variable models, where the observed variables interact in complex ways with unobserved variables through a distribution described by the graph. Ultimately, the graph can be viewed as a representation of a generative process that describes the series of stochastic steps by which the data is generated.

[1] D-separation is used to define independence in graphical models. An introduction to this concept can be found in [32]



(a) Observed variables W_i, where $i \in \overline{1,M}$, dependent on latent variable V.



(b) Observed variable W dependent on latent variable V. The plate indicates variable W is repeated M times.

Figure 2.2: Plate notation is used to increase the readability of complex models. Plates indicate repeating structures.

Example. Gaussian mixture models, which can be expressed as latent variable mixture distributions, have discrete latent variables that define the assignment of data points to components of the mixture. From a generative stand point we can view the data set X as being generated one sample at a time. Assuming a data-set of N samples, the process of creating such samples takes the following form:

1. Assign $n \leftarrow 1$ and **BEGIN**

2. ***Latent aspect assignment for each sample:*** Pick a mixture component $z_n \in \overline{1,T}$ from a multinomial probability distribution with parameter $\Pi = \{\pi_t\}$, where π_t is the probability of picking the t-th component.

3. ***Sample creation:*** Generate a sample $x_n \sim \mathcal{N}(\mu_{z_n}, \sigma_{z_n})$. If $n < N$ then $n \leftarrow n+1$ and repeat from item 2, otherwise **STOP.**

These steps describe the hypothetical generative scenario that created the *seen variables*, viz., x_n where $n \in \overline{1,N}$. The variables in the intermediate steps are the *unseen* or *latent* variables of the model that help explain the underlying structure of the data. The graphical model corresponding to this generative process can be see in fig. 2.3.

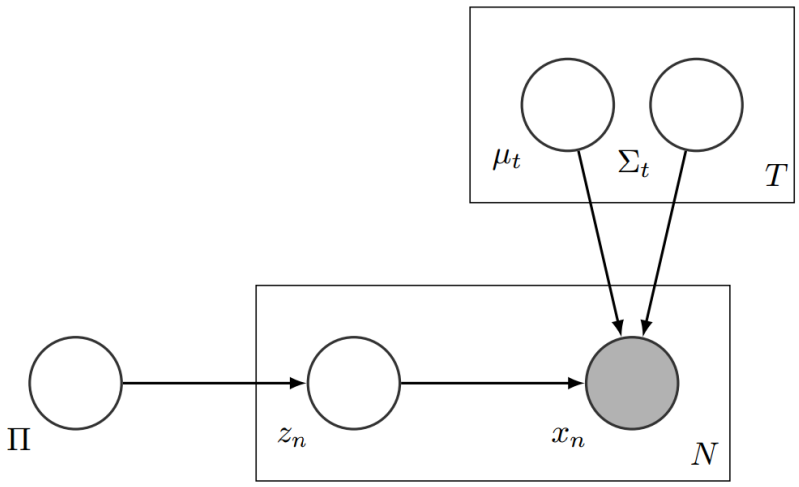


Figure 2.3: Graphical model for a mixture of gaussians

2.2 Inference

Once a graphical model has been conceived, our focus turns to estimating the posterior distribution of the latent variables conditioned on the observations. More formally, assuming a model with observed variables $x_n \in X$ and latent variables $z_n \in Z$ with $n \in \overline{1,N}$, the inference problem lies in finding $p(Z|X)$. This posterior distribution can be written as,

$$p(Z|X) = \frac{p(Z,X)}{p(X)} = \frac{p(X,Z)}{\sum_Z p(X,Z)} \tag{2.2}$$

where the joint distribution $p(X,Z)$ can be factorized according to the graph and $p(X)$ can be obtained by marginalizing the joint distribution $p(X,Z)$ over Z. This sum over all possible values of Z increases exponentially with the possible values of z_n. Thus, if $z_n \in \overline{1,K}$ for $n \in \overline{1,N}$ the sum in eq. (2.2) is over K^N possible values. This marginalization step becomes prohibitive as the number of variables in Z increases. Therefore, exact inference in most models is intractable. Note that the same is true for continuous variables where the summation sign is replaced by an integral.

Thankfully, several approximate inference techniques can be use in order to estimate the posterior $p(Z|X)$. Variational methods approximate the desired distribution $p(Z|X)$ with an approximate posterior $q(Z)$ by minimizing the *Kullback-Leiber divergence* [31]. On the other hand, Markov chain Monte Carlo (MCMC) sampling methods build a Markov chain with a stationary distribution proportional to $p(Z|X)$, and collect samples from the chain after it has converged to the desired distribution.

2.2.1 Exponential Family

The distributions considered in GM literature are in the exponential family. This family of distributions simplifies both variational and MCMC optimization techniques and encompasses many common distributions. The exponential family distributions take the following form:

Table 2.1: Common Exponential Family Distributions

Distribution	η	$h(x)$	$g(\eta)$	$u(x)$
Bernoulli	$ln\left(\frac{\mu}{1-\mu}\right)$	1	$(1+\exp(\eta))^{-1}$	x
Beta	$\begin{bmatrix}\alpha\\ \beta\end{bmatrix}$	$\frac{1}{x(1-x)}$	$\frac{\eta_1\eta_2}{\eta_1+\eta_2}$	$\begin{bmatrix}ln(x)\\ ln(1-x)\end{bmatrix}$
Gaussian	$\begin{bmatrix}\mu/\sigma^2\\ -1/2\sigma^2\end{bmatrix}$	$(2\pi)^{-1/2}$	$(-2\eta_2)^{1/2}\exp\left\{\frac{\eta_1^2}{4\eta_2}\right\}$	$\begin{bmatrix}x\\ x^2\end{bmatrix}$
Multinomial (M states)	$\eta_k = ln\left(\frac{\mu_k}{1-\sum_j \mu_j}\right)$	1	$\left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}$	x

$$p(x|\eta) = h(x)g(\eta)\exp\{\eta^\top u(x)\} \tag{2.3}$$

where x is a scalar or vector that can be continuous or discrete, η are the *natural parameters* of the distribution, $u(x)$ are the *sufficient statistics* and $g(\eta)$ is a normalization coefficient that ensures the distribution satisfies

$$g(\eta)\int h(x)\exp\{\eta^\top u(x)\}dx = 1. \tag{2.4}$$

Several common distributions such as the Gaussian, the Beta and the multinomial distribution are in the exponential family and can be expressed in the form of eq. (2.3). Table 2.1 shows how some common distributions are written as exponential family distributions and a thorough review on the properties of exponential family distributions is given in [33].

2.2.2 Conjugate Priors

In Bayesian decision theory the posterior over the model parameters in eq. (2.2) is usually expressed in terms of the likelihood $p(X|Z)$ and the prior over the model parameters $p(Z)$:

$$p(Z|X) = \frac{p(X|Z)p(Z)}{\sum_Z p(X|Z)p(Z)} \tag{2.5}$$

The likelihood function is usually fixed, as it is determined from the generative process expressed through the graphical model. On the other hand, the prior function is not fixed and its choice can change the form in which the posterior is expressed. Certain distributions produce a posterior, which has the same algebraic form as the prior. Such distributions are known as *conjugate priors* and they produce closed form posteriors that are still part of the exponential family distributions.

Conjugate priors play an important role in approximate inference techniques as they exhibit very useful properties, which simplify parameter optimization in both variational and sampling methods.

2.2.3 Variational Methods

Variational methods have their roots in the calculus of variations, which emanated from the work of Euler and Lagrange in the 18^{th} century [32]. Variational approaches approximate the posterior $p(Z|X)$ with an approximate posterior $q(Z)$. In order to apply variational optimization to the inference problem, let us decompose the log marginal probability of the observations $\ln p(X)$ as follows:

$$
\begin{aligned}
\ln p(X) &= \sum_Z q(Z) \ln p(X), \quad \text{Given that } \sum_Z q(Z) = 1 \\
&= \sum_Z q(Z) \ln \frac{p(Z|X)p(X)}{p(Z|X)} \\
&= \sum_Z q(Z) \ln \frac{p(Z,X)}{p(Z|X)} \\
&= \sum_Z q(Z) \ln p(Z,X) - q(Z) \ln p(Z|X) \\
&= \sum_Z q(Z) \ln p(Z,X) - q(Z) \ln q(Z) - (q(Z) \ln p(Z|X) - q(Z) \ln q(Z)) \\
&= \sum_Z q(Z) \ln \frac{p(X,Z)}{q(Z)} - \sum_Z q(Z) \ln \frac{p(Z|X)}{q(Z)}.
\end{aligned}
$$

Using this decomposition the equation can then be written as

$$\ln p(X) = \mathscr{L}(q(Z)) + KL(q(Z)||p(Z|X)). \tag{2.6}$$

where $KL(q(Z)||p(Z|X))$ is the Kullback-Leibler (KL) divergence between $p(Z|X)$ and $q(Z)$ which satisfies $KL(q(Z)||p(Z|X)) \geq 0$. Thus, $\mathscr{L}(q(Z)) \leq \ln p(X)$, making $\mathscr{L}(q(Z))$ the lower bound of the log marginal $\ln p(X)$. Note that the maximum lower bound is achieved by minimizing the KL divergence, which by definition occurs when $q(Z) = p(Z|X)$.

The problem then becomes finding the optimum $q(Z)$ that minimizes the KL divergence. Since the true posterior is intractable, a restricted but tractable family of distributions is considered. There are several ways of restricting the family of distributions $q(Z)$ giving rise to a wide variety of variational approximations methods in literature.

One of the most widely used methods is known as *mean field variational inference* in which the family of distributions $q(Z)$ is assumed factorizable by assuming that Z can be partitioned into disjoint subsets Z_i , with $i \in \overline{1,M}$, such that:

$$q(Z) = \prod_{i=1}^{M} q(Z_i). \tag{2.7}$$

Using this family of distributions to maximize the lower bound $\mathscr{L}(q(Z))$ gives rise to a set equations that can be used to iteratively estimate a solution [32]. A thorough review of variational methods can be found in [31].

2.2.4 MCMC and Gibbs Sampling

MCMC methods have their beginnings in the Metropolis algorithm [34, 35], which was originally conceived by physicists to estimate the value of complex integrals using a random number generator. This process is known as Monte Carlo integration. The idea is to express the integrals as expectations of some probability distribution and then estimate the expectation by sampling the distribution. In the context of GMs, we are interested in

finding the expectation of some function $f(Z)$ with respect to a distribution $p(Z)$. Note that in eq. (2.5) $f(Z) = p(X|Z)$.

The Gibbs sampler was introduced as a special case of Metropolis-Hastings sampling in image processing [36]. The simplicity of the Gibbs sampler is due to the fact that it only considers univariate conditional distributions instead of the more complex underlying join distributions. Smith *et al.* [37] showed the usefulness of the sampler in Bayesian statistics estimating posterior distributions.

Given a k-dimensional random variable, sampling from a joint distribution would imply sampling a k-dimensional vector in a single pass. Instead, Gibbs sampling generates k random variables sequentially from k univariate conditional distributions [38].

Example. Assuming three variables $\{z_1, z_2, z_3\}$ a sampler would cycle through the variables sampling each one from its conditional distribution. Let z_i^j denote the value of the i-th variable at the j-th step. A sampler running through τ iterations would initialize $\{z_1^1, z_2^1, z_3^1\}$ and perform the following steps:

1. $t \leftarrow 1$ and **BEGIN**
2. Sample z_1^{t+1} from $p(z_1|z_2^t, z_3^t)$
3. Sample z_2^{t+1} from $p(z_2|z_1^{t+1}, z_3^t)$
4. Sample z_3^{t+1} from $p(z_3|z_1^{t+1}, z_2^{t+1})$
5. if $t < \tau$ then $t \leftarrow t+1$ and repeat from item 2, otherwise **STOP.**

This iterative process defines a Markov chain, which after a sufficient *burn-in* period will reach its stationary distribution, which is proportional to the joint distribution of all variables $p(z_1, z_2, z_3)$ as shown in eq. (2.8).

$$p(z_i|z_{\neg i}) = \frac{p(z_i, z_{\neg i})}{p(z_{\neg i})} \propto p(\mathbf{z}) \tag{2.8}$$

where $p(z_{\neg i})$ is constant, with $z_{\neg i}$ denoting all other variables except for z_i and $\mathbf{z} = z_i \cup z_{\neg i}$.

Estimating the correct number of iterations required for the burn-in is one of the major challenges and drawbacks of MCMC algorithms. Most efforts to determine convergence have focused on theoretical solutions and on diagnostic tools. Theoretical solutions try to analyze the transition kernel of the chain in order to estimate the minimum number of iterations needed [39, 40]. On the other hand, diagnostic tools are used in the output of the algorithm in order to determine if the samples collected have reached the stationary distribution [41]. Convergence in the context of MCMC algorithms is notably difficult given that, unlike other iterative algorithms, at convergence the output of an MCMC sampler does not approach a specific number or even a distribution. The output of the sampler are samples from a distribution that are highly correlated from one step of the chain to the next. Several methods have been proposed in literature, but convergence of MCMC algorithms is still an area of active research. A comparative review of different techniques can be found in [42].

Features of the joint distribution can be computed by using the samples obtained from the chain, but several iterations must pass between collected samples in order to minimize correlation. Further details on the Gibbs sampler can be found in [43, 44].

Chapter 3

The Proposed Approach

Latent Dirichlet Allocation (LDA) is known as a topic model, originally introduced in [45]. It is primarily used for discovering the main themes that compose large collections of documents. LDA represents documents as a finite mixture of hidden aspects or latent topics and was initially designed as a probabilistic model of text, casting the goal of discovering themes in collections of documents as a posterior inference problem. The LDA model is modular enough to be nested and tailored for specific problem domains, which has allowed its application in a wide variety of fields. In this paper we tailor LDA for our application domain. We introduce and explain LDA's theoretical background as it relates directly to our domain area.

3.1 LDA for EEMs

As in Section 1.3, we identify the collection of all EEMs in our data set as the three-way array $\mathbf{X} = \{x_{drc}\} \in \mathbb{R}^{D\times R\times C}$. Note that, $d \in \overline{1,D}$, $r \in \overline{1,R}$, and $c \in \overline{1,C}$. For a fixed d value, $\mathbf{x}_d \in \mathbb{R}^{R\times C}$ refers to the d-th EEM sample in the data set. As mentioned before, D such matrices can be obtained by slicing the first dimension of the original data set $\mathbf{X}$.

A discrete EEM $\widehat{\mathbf{x}}_d = \{\widehat{x}_{drc}\} \in \mathbb{N}^{R\times C}$, where $\widehat{x}_{drc} \in \mathbb{N}$, can now be generated as

$$\widehat{x}_{drc} = \left\lfloor \frac{x_{drc}}{\delta} \right\rfloor, \tag{3.1}$$

where δ is the quantization parameter defining the smallest step size with which fluorescence can change. Note that $\widehat{x}_{drc} \times \delta \leq x_{drc}$; equality for all r and c values are achieved when δ equals the sensitivity of the spectrometer. In Section 3.5, we specify an adaptive approach to obtain a location specific quantization value to better retain information crucial to class separability. However, it is worth mentioning that when the objective is quantitative in nature (i.e., when we are interested in not only the presence/absence of compounds, but their concentrations as well), a fixed value of δ should be used to quantize all entries in a matrix.

For the purpose of expressing the EEM data set in terms of an LDA model, we now provide an alternate characterization of the d-th discrete EEM sample $\widehat{\mathbf{x}}_d = \{\widehat{x}_{drc}\}$. Towards this purpose, it is more convenient to identify the elements $\widehat{x}_{drc}$ by indexing them row-wise so that we may drop the triple subscript notation in $\widehat{x}_{drc}$ in favor of the double subscript notation $\widehat{x}_{di}$, where $\widehat{x}_{di} = \widehat{x}_{drc}$, with $i = (r-1) \times C + c$. Let us now characterize $\widehat{\mathbf{x}}_d = \{\widehat{x}_{di}\}$ via a landscape of 'fluorescence blocks' in the following manner: corresponding to each element $\widehat{x}_{di}$, $i \in \overline{1, RC}$, of $\widehat{\mathbf{x}}_d$, pick an $\widehat{x}_{di}$ number of fluorescence blocks, each block having a 'value' or 'weight' of i. The total number of such blocks required to capture the complete d-th discrete EEM sample $\widehat{\mathbf{x}}_d$ is $L_d = \sum_{i=1}^{RC} \widehat{x}_{di}$. We can enumerate these fluorescence blocks of $\widehat{\mathbf{x}}_d$ via $q_{d\ell}$, $\ell \in \overline{1, L_d}$; the weight associated with $q_{d\ell}$ can be expressed via

$$q_{d\ell} = i, \text{ for } \ell = \sum_{j=1}^{i-1} \widehat{x}_{dj} + 1, \ldots, \sum_{j=1}^{i} \widehat{x}_{dj}, \tag{3.2}$$

for $i \in \overline{1, RC}$. Note that, $q_{d\ell} \in \overline{1, RC}$. In essence, this process yields an equivalent characterization of the d-th discrete EEM $\widehat{\mathbf{x}}_d = \{\widehat{x}_{di}\}$ in the form of $\mathbf{q}_d = \{q_{d\ell}\}$, which is a landscape made of fluorescence blocks. Each block has a height of δ and blocks that share the common position i (and hence having the same weight) are stacked on top of each other; the height of the stack represents the fluorescence at the location i.

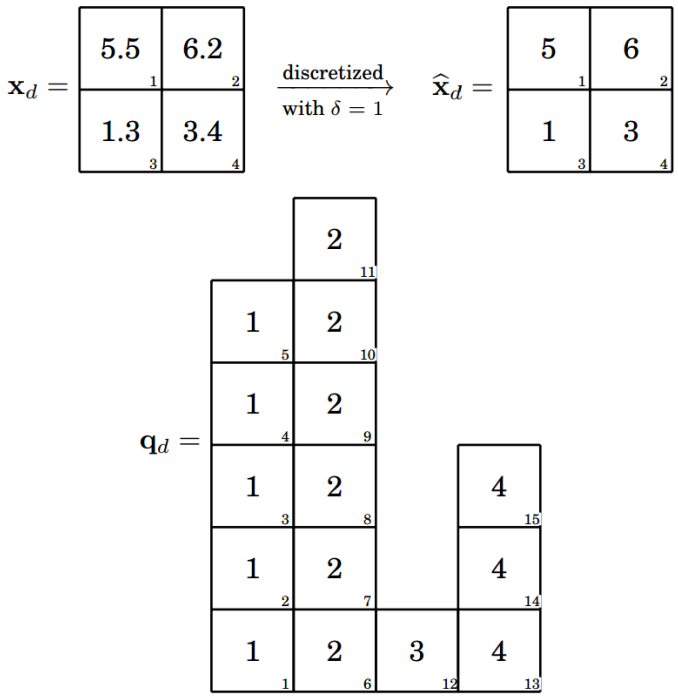


Figure 3.1: The top two figures depict a 2×2 EEM sample $\mathbf{x}_d$ and its discretized version $\widehat{\mathbf{x}}_d = \{\widehat{x}_{di}\}$, $i \in \overline{1,4}$, assuming a quantization parameter $\delta = 1$ (numbers at the middle and bottom-right of each box depict fluorescence and value of i, respectively). The bottom figure depicts the equivalent characterization $\mathbf{q}_d = \{q_{d\ell}\}$, $\ell \in \overline{1,15 = L_d}$ (numbers at the middle and bottom-right of each box depict the values of i and ℓ, respectively).

Example. The quantized matrix $\widehat{\mathbf{x}}_d$ in Figure 3.1 may be described as having 5 blocks at location 1, 6 blocks at location 2, 1 block at location 3, and 3 blocks at location 4, i.e.,

$$q_{d1},\ldots,q_{d5} = 1; \quad q_{d6},\ldots,q_{d11} = 2;$$
$$q_{d12} = 3; \quad q_{d13},\ldots,q_{d15} = 4.$$

Due to the equivalence of the representations $\widehat{\mathbf{x}}_d = \{\widehat{x}_{d\ell}\}$ and $\mathbf{q}_d = \{q_{d\ell}\}$, we will continue to use $\widehat{\mathbf{x}}_d$ to refer to either representation.

3.2 Generative Process

In our implementation, we view the discrete EEM samples $\widehat{\mathbf{x}}_d$, $d \in \overline{1,D}$, as being generated from an underlying pool of T compounds with different fluorescent properties. Henceforth, these 'latent' compounds, indexed via $\{1,\ldots,T\}$, will be referred to as the *latent aspects* of our model. It is worth noting that, in the original LDA implementation in [45] and in most work related to LDA, the latent aspects of a model represent the topics that could be present in a document. More formally, from a generative standpoint, they represent the topic set from which the underlying structure of the document originates.

For each latent aspect enumerated via $t = \overline{1,T}$, let $\phi_t = \{\phi_{ti}\}$, $i \in \overline{1,RC}$, denote a multinomial distribution over the RC EEM matrix entries; for each discrete EEM sample enumerated via $d = \overline{1,D}$, let $\theta_d = \{\theta_{dt}\}$, $t \in \overline{1,T}$, denote a multinomial distribution over the T latent aspects. We assume that ϕ_t and θ_d are drawn from Dirichlet distributions

$$\phi_t = \{\phi_{ti}\} \sim Dir_{RC}(\beta = \{\beta_i\}), \text{ for } t \in \overline{1,T};$$
$$\theta_d = \{\theta_{dt}\} \sim Dir_T(\alpha = \{\alpha_t\}), \text{ for } d \in \overline{1,D},$$

where $Dir_N(\mathbf{v})$ denotes the N-dimensional Dirichlet distribution with parameter $\mathbf{v}$. For example, the probability density function (p.d.f.) $p(\theta_d \,|\, \alpha)$ is given by the Dirichlet distribution

$$p(\theta_d \,|\, \alpha) = \frac{\Gamma\left(\sum_t \alpha_t\right)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{dt}^{\alpha_t - 1}, \tag{3.3}$$

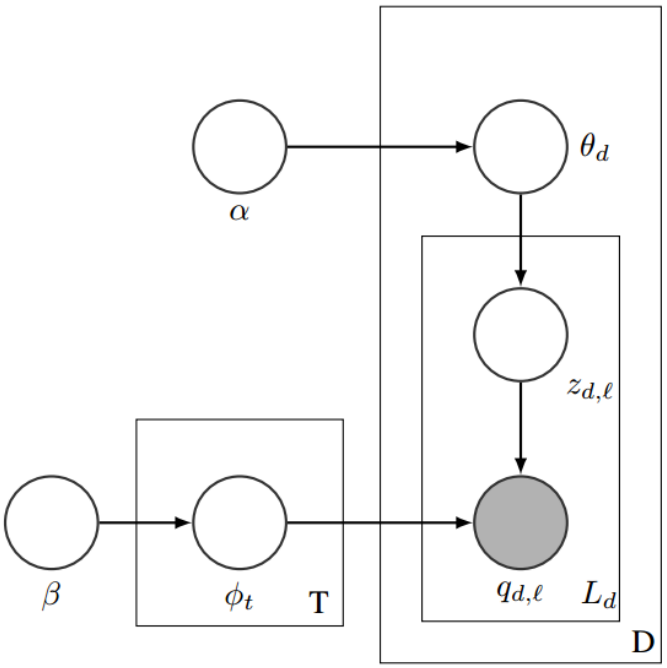


Figure 3.2: The LDA graphical model. Shaded and clear nodes represent observed random variables and latent random variables, respectively. The rectangles represent repeated structures and the arrows represent dependencies.

where $\Gamma(\cdot)$ is the Gamma function and $\alpha = \{\alpha_t\}$ is the T-dimensional vector parameter of the Dirichlet distribution. In a similar manner, the random variable ϕ_t follows a Dirichlet distribution but conditioned on the RC-dimensional vector parameter $\beta = \{\beta_i\}$.

Using $Multi_N(\mathbf{w})$ and $Poisson(\vartheta)$ to denote the N-dimensional multinomial distribution with parameter $\mathbf{w}$ and the Poisson distribution with parameter ϑ, respectively, the generative process of creating the d-th discrete EEM sample takes the following form:

1. ***Latent aspect distributions in samples,*** ϕ_t***:*** For each latent aspect, i.e., for each $t \in \overline{1,T}$, choose a multinomial distribution over the RC EEM matrix entries from $\phi_t \sim Dir_{RC}(\beta)$.

2. ***Latent aspect proportions,*** θ_d***:*** Choose a multinomial distribution over the T latent aspects from $\theta_d \sim Dir_T(\alpha)$.

3. ***Number of fluorescence blocks,*** L_d***:*** Choose the number of fluorescence blocks from $L_d \sim Poisson(\vartheta)$.

4. Consider the fluorescence blocks $\ell = \overline{1, L_d}$. Assign $\ell \leftarrow 1$ and **BEGIN:**

   (a) ***Latent aspect assignment for each block,*** $z_{d\ell}$***:*** Choose a latent aspect from $z_{d\ell} \sim Multi_T(\theta_d)$.

   (b) ***Landscape of fluorescence blocks,*** $\{q_{d\ell}\}$***:*** For this latent aspect $z_{d\ell}$, choose the weight of the associated fluorescence block from $q_{d\ell} \sim Multi_{RC}(\phi_{z_{d\ell}})$.

   (c) ***Fluorescence intensity,*** $\widehat{\mathbf{x}}_d = \{\widehat{x}_{d\ell}\}$***:*** Increase the fluorescence intensity at the location corresponding to $q_{d\ell}$ by incrementing the value of $\widehat{x}_{dq_{d\ell}}$ by 1.

   (d) If $\ell < L_d$, assign $\ell \leftarrow \ell + 1$ and repeat from Step 4a; otherwise, **STOP.**

Note that, as in the original description of the LDA [45], the Poisson assumption in Item 3 is not critical to the formulation of the problem given that it is independent from the data generating variables ϕ_t, θ_d and $z_{d\ell}$. Thus, we ignore it in the discussion to follow.

The graphical model in Figure 3.2 represents this generative process. The shaded nodes represent observed random variables; the clear nodes represent hidden or latent random variables; the arrows represent dependencies between variables; the plates represent repeating structures; and the number at the lower right corner of each plate represents the repetition factor. As can be seen from Figure 3.2, there are T multinomial distributions ϕ_t (one per latent aspect), D latent aspect distributions θ_d (one per EEM sample), and L_d fluorescence blocks in the d-th EEM. More importantly, Figure 3.2 provides a factorization of the joint p.d.f. $p(\mathbf{Z}, \widehat{\mathbf{X}}, \Theta, \Phi \,|\, \alpha, \beta)$ as

$$p(\mathbf{Z}, \widehat{\mathbf{X}}, \Theta, \Phi \,|\, \alpha, \beta) = \left(\prod_{t=1}^{T} p(\phi_t \,|\, \beta) \right) \prod_{d=1}^{D} p(\theta_d \,|\, \alpha) \prod_{\ell=1}^{L_d} p(z_{d\ell} \,|\, \theta_d)\, p(q_{d\ell} \,|\, \phi_{z_{d\ell}}). \tag{3.4}$$

Here, $\Phi = \{\phi_t\}$, $\Theta = \{\theta_d\}$, $\mathbf{Z} = \{\mathbf{z}_d\}$ where $\mathbf{z}_d = \{z_{d\ell}\}$, and $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_d\}$ is represented via $\{q_{d\ell}\}$, with $t \in \overline{1, T}$, $d \in \overline{1, D}$, and $\ell \in \overline{1, L_d}$.

3.3 Parameter Estimation

Given the discrete EEM sample data set $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_d\}$, the utility of the LDA model is to estimate the posterior distributions over the latent aspect assignments $z_{d\ell}$, the latent aspect proportions θ_d, and the latent aspect distributions ϕ_t. This is conducted by drawing samples from the joint p.d.f. $p(\mathbf{Z}, \widehat{\mathbf{X}}, \Theta, \Phi \,|\, \alpha, \beta)$, for which we employ *Gibbs sampling,* a Markov chain-Monte Carlo (MCMC) algorithm. When sampling directly from a joint distribution is difficult or when the joint distribution is unknown, Gibbs sampling provides a strategy to obtain a sequence of random samples from the joint distribution from its conditional distributions (which are typically easier to sample from). One may simplify the procedure even further by noting that the Dirichlet distributions of ϕ_t and θ_d constitute conjugate priors for the multinomial distribution of $z_{d\ell}$. This allows one to use a modification referred to as *collapsed Gibbs sampling* [46] where ϕ_t and θ_d can be marginalized out when sampling the latent aspect assignments $z_{d\ell}$.

3.3.1 Sampling the Conditional Distribution

In essence, our objective is to sample from the conditional distribution of $z_{d\ell}$ corresponding to the ℓ-th fluorescence block of the EEM sample $\widehat{\mathbf{x}}_d$ conditioned on all the other remaining variables, i.e., $p(z_{d\ell} \,|\, z_{\neg(d\ell)}, \widehat{\mathbf{X}})$, where $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_d\}$ and $z_{\neg(d\ell)}$ denotes all other fluorescence blocks except $z_{d\ell}$ itself. Note that

$$p(z_{d\ell} \,|\, z_{\neg(d\ell)}, \widehat{\mathbf{X}}, \alpha, \beta) = \frac{p(\mathbf{Z} \,|\, \widehat{\mathbf{X}}, \alpha, \beta)}{p(z_{\neg(d\ell)} \,|\, \widehat{\mathbf{X}}, \alpha, \beta)}, \tag{3.5}$$

where $\mathbf{Z} = z_{d\ell} \cup z_{\neg(d\ell)}$ is the totality of the fluorescence blocks across all the D EEM samples. The denominator in eq. (3.5), which is devoid of the variable $z_{d\ell}$ being sampled, is essentially a normalizing constant. Note also that

$$\frac{p(\mathbf{Z} \,|\, \widehat{\mathbf{X}}, \alpha, \beta)}{p(z_{\neg(d\ell)} \,|\, \widehat{\mathbf{X}}, \alpha, \beta)} = \frac{p(\mathbf{Z}, \widehat{\mathbf{X}} \,|\, \alpha, \beta)}{p(z_{\neg(d\ell)}, \widehat{\mathbf{X}} \,|\, \alpha, \beta)}. \tag{3.6}$$

Thus, the conditional distribution on the left hand side of eq. (3.5) is proportional to the joint distribution of all latent variables $p(\mathbf{Z}, \widehat{\mathbf{X}} \,|\, \alpha, \beta)$. It is this observation that eventually leads

to the conclusion that sampling from the conditional distribution of all variables iteratively will lead to the distribution we are interested in.

3.3.2 MCMC Gibbs Sampling Algorithm

Given that the Dirichlet distribution is the conjugate prior of the multinomial distribution, we are able to integrate out of the multinomial parameters. Such approach is known as the collapse of the sampler. To find $p(\mathbf{Z}, \widehat{\mathbf{X}} \,|\, \alpha, \beta)$ on the right-hand side of eq. (3.5) from the original joint distribution $p(\mathbf{Z}, \widehat{\mathbf{X}}, \Theta, \Phi \,|\, \alpha, \beta)$ given in eq. (3.4), rearrage the terms and integrate out the multinomial parameters Φ and Θ,

$$
\begin{aligned}
&p(\mathbf{Z}, \widehat{\mathbf{X}} \,|\, \alpha, \beta) \\
&= \int_{\Theta} \int_{\Phi} \left(\prod_{d=1}^{D} p(\theta_d \,|\, \alpha) \prod_{\ell=1}^{L_d} p(z_{d\ell} \,|\, \theta_d) \right) \left(\prod_{t=1}^{T} p(\phi_t \,|\, \beta) \right) \left(\prod_{d=1}^{D} \prod_{\ell=1}^{L_d} p(q_{d\ell} \,|\, \phi_{z_{d\ell}}) \right) d\Phi \, d\Theta.
\end{aligned}
\tag{3.7}
$$

To develop the MCMC Gibbs sampling algorithm, we will utilize the following counting variable:

$$
n_{di}^{(t)} = \sum_{\ell=1}^{L_d} \mathbb{1}_t(z_{d\ell}) \cdot \mathbb{1}_i(q_{d\ell}), \tag{3.8}
$$

where $\mathbb{1}_v(u)$ is the indicator function

$$
\mathbb{1}_v(u) = \begin{cases} 1, & \text{for } u = v; \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}
$$

We will use the sub/superscript $*$ to indicate that the corresponding variable has been 'summed' out. So, for example,

$$
n_{d*}^{(t)} = \sum_{i=1}^{RC} n_{di}^{(t)}; \;\; n_{*i}^{(t)} = \sum_{d=1}^{D} n_{di}^{(t)}; \;\; n_{**}^{(t)} = \sum_{i=1}^{RC} \sum_{d=1}^{D} n_{di}^{(t)}. \tag{3.10}
$$

Note that, $n_{d*}^{(t)}$ is a count of the number of fluorescent blocks which belong to aspect t in the d-th EEM sample $\widehat{\mathbf{x}}_d$; $n_{*i}^{(t)}$ is a count of the number of blocks positioned at i which are assigned to aspect t in the complete set $\widehat{\mathbf{X}}$ of D EEM samples; and $n_{**}^{(t)}$ is a count of all the

blocks which belong to aspect t in the complete data set. Let us also use $\neg n_{d'*}^{(t)}$ and $\neg n_{*i}^{(t)}$ which are defined similar to $n_{d*}^{(t)}$ and $n_{*i}^{(t)}$, except that the counts for block $q_{d'\ell'}$ (i.e., the ℓ'-th block in the d'-th EEM sample) are ignored.

Then, as we demonstrate in section 3.3.3, we can show that the conditional distribution in eq. (3.5) is proportional to

$$p(z_{d'\ell'} \,|\, z_{\neg(d'\ell')}, \widehat{\mathbf{X}}, \alpha, \beta) \propto \frac{\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}}}{\neg n_{d'*}^{(*)} + \sum_{t=1}^{T} \alpha_t} \cdot \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} + \sum_{i=1}^{RC} \beta_i}. \tag{3.11}$$

The MCMC Gibbs sampling algorithm is initialized by assigning random values from $\{1, \cdots, T\}$ to all variables $z_{d\ell}$, thus effectively establishing the values at the first step of the Markov chain. After initialization, the Gibbs sampler iterates through all variables $z_{d\ell}$ several times until the Markov chain approaches the target distribution. Determination of this *burn-in* period is one of the difficulties associated with MCMC algorithms. The interested reader can find several methods to check convergence in [47]. At the completion of the burn-in period, the converged values for the sampled variables are recorded. Now the matrix latent proportions θ_d and the latent aspect distributions ϕ_t can be found by sampling $\mathbf{Z}$ via

$$\theta_d^{(t)} = \frac{n_{d*}^{(t)} + \alpha_t}{n_{d*}^{(*)} + \sum_{t=1}^{T} \alpha_t}; \;\; \phi_t^{(i)} = \frac{n_{*i}^{(t)} + \beta_i}{n_{*i}^{(*)} + \sum_{i=1}^{RC} \beta_i}. \tag{3.12}$$

With symmetrical Dirichlet priors, i.e., $\alpha_t = \alpha$ and $\beta_i = \beta$, we have $\sum_{t=1}^{T} \alpha_t = T\alpha$ and $\sum_{i=1}^{RC} \beta_i = RC\beta$.

The values calculated in eq. (3.12) can be estimated using a single draw from a read out of the sampler or by sampling several times allowing enough iterations between read outs so as to minimize correlation and then obtaining an average of the sampled variables $z_{d\ell}$.

3.3.3 Gibbs Sampling Derivation

Note that eq. (3.7) can be expressed as

$$p(\widehat{\mathbf{X}},\mathbf{Z}\,|\,\alpha,\beta) = \underbrace{\int_{\Theta}\left(\prod_{d=1}^{D} p(\theta_d\,|\,\alpha)\prod_{\ell=1}^{L_d} p(z_{d\ell}\,|\,\theta_d)\right)d\Theta}_{P_1}\cdot\underbrace{\int_{\Phi}\left(\prod_{t=1}^{T} p(\phi_t\,|\,\beta)\right)\left(\prod_{d=1}^{D}\prod_{\ell=1}^{L_d} p(q_{d\ell}\,|\,\phi_{z_{d\ell}})\right)d\Phi}_{P_2}. \tag{3.13}$$

The conjugacy between the Dirichlet priors and the multinomial distributions significantly simplifies the Gibbs sampling algorithm. Substitute the Dirichlet priors and the multinomial distributions into eq. (3.13):

$$P_1 = \int_{\Theta}\prod_{d=1}^{D}\underbrace{\frac{\Gamma\left(\sum_{t=1}^{T}\alpha_t\right)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\left(\prod_{t=1}^{T}\theta_{dt}^{\alpha_t-1}\right)}_{p(\theta_d|\alpha)}\prod_{\ell=1}^{L_d}\underbrace{\theta_{dz_{d\ell}}}_{p(z_{d\ell}\,|\,\theta_d)}\,d\Theta;$$

$$P_2 = \int_{\Phi}\underbrace{\left(\prod_{t=1}^{T}\frac{\Gamma\left(\sum_{i=1}^{RC}\beta_i\right)}{\prod_{i=1}^{RC}\Gamma(\beta_i)}\left(\prod_{i=1}^{RC}\phi_{ti}^{\beta_i-1}\right)\right)}_{p(\phi_t\,|\,\beta)}\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\underbrace{\phi_{z_{d\ell}q_{d\ell}}}_{p(q_{d\ell}\,|\,\phi_{z_{d\ell}})}\,d\Phi. \tag{3.14}$$

Let us now consider the following:

- ***Term*** $\prod_{\ell=1}^{L_d}\theta_{dz_{d\ell}}$ ***in*** P_1***:*** The probability of observing a given set of aspect variables $z_{d\ell}$, $\ell\in\overline{1,L_d}$, in the d-th EEM, where $\theta_{dz_{d\ell}}$ is the $z_{d\ell}$-th component of θ_d (i.e., the probability of aspect $z_{d\ell}$ given θ_d). Then, we observe that exponentiating the probabilities $\theta_{dt}, t\in\overline{1,T}, d\in\overline{1,D}$ to the count variable that indicates how many times the probability appears in the factorization, we get

$$\prod_{\ell=1}^{L_d}\theta_{dz_{d\ell}} = \prod_{t=1}^{T}\theta_{dt}^{n_{d*}^{(t)}}, \tag{3.15}$$

where $n_{d*}^{(t)}$ denotes the count of fluorescent blocks which belong to aspect t in the d-th EEM sample.

- ***Term*** $\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\phi_{z_{d\ell}q_{d\ell}}$ ***in*** P_2***:*** The probability of observing a set of block positions $q_{d\ell}$ given aspect $z_{d\ell}$, where $\phi_{z_{d\ell}q_{d\ell}}$ refers to the $q_{d\ell}$-th component of $\phi_{z_{d\ell}}$. As shown before we can observe that by exponentiating the probabilities $\phi_{ti}, t \in \overline{1,T}, i \in \overline{1,RC}$ to the count variable that indicates how many times the probability appears in the factorization, the term can be rewritten as

$$\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\phi_{z_{d\ell}q_{d\ell}} = \prod_{t=1}^{T}\prod_{i=1}^{RC}\phi_{ti}^{n_{*i}^{(t)}}. \tag{3.16}$$

Substitute eq. (3.15) and eq. (3.16) in eq. (3.14) and switch the integral and the product operators:

$$\begin{aligned}
P_1 &= \int_{\Theta}\prod_{d=1}^{D}\frac{\Gamma\left(\sum_{t=1}^{T}\alpha_t\right)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\left(\prod_{t=1}^{T}\theta_{dt}^{\alpha_t-1}\right)\prod_{t=1}^{T}\theta_{dt}^{n_{d*}^{(t)}}d\Theta; \\
&= \prod_{d=1}^{D}\int_{\theta_d}\frac{\Gamma\left(\sum_{t=1}^{T}\alpha_t\right)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\prod_{t=1}^{T}\theta_{dt}^{n_{d*}^{(t)}+\alpha_t-1}\,d\theta_d; \\
P_2 &= \int_{\Phi}\left(\prod_{t=1}^{T}\frac{\Gamma\left(\sum_{i=1}^{RC}\beta_i\right)}{\prod_{i=1}^{RC}\Gamma(\beta_i)}\left(\prod_{i=1}^{RC}\phi_{ti}^{\beta_i-1}\right)\right)\left(\prod_{t=1}^{T}\prod_{i=1}^{RC}\phi_{ti}^{n_{*i}^{(t)}}\right)d\Phi \\
&= \prod_{t=1}^{T}\int_{\phi_t}\frac{\Gamma\left(\sum_{i=1}^{RC}\beta_i\right)}{\prod_{i=1}^{RC}\Gamma(\beta_i)}\prod_{i=1}^{RC}\phi_{ti}^{n_{*i}^{(t)}+\beta_i-1}\,d\phi_t.
\end{aligned} \tag{3.17}$$

Take the constant terms out of the integrals in eq. (3.17), and noting that the integrals are unnormalized Dirichlet distributions, multiply the inside of the integral by a normalization

factor and the outside by the factor's reciprocal:

$$P_1 = \left(\frac{\Gamma\left(\sum_{t=1}^{T} \alpha_t\right)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \right)^{D} \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(n_{d*}^{(t)} + \alpha_t)}{\Gamma\left(\sum_{t=1}^{T} n_{d*}^{(t)} + \alpha_t\right)} \underbrace{\int_{\theta_d} \frac{\Gamma\left(\sum_{t=1}^{T} n_{d*}^{(t)} + \alpha_t\right)}{\prod_{t=1}^{T} \Gamma(n_{d*}^{(t)} + \alpha_t)} \prod_{t=1}^{T} \theta_{dt}^{n_{d*}^{(t)} + \alpha_t - 1} d\theta_d}_{=1};$$

$$P_2 = \left(\frac{\Gamma\left(\sum_{i=1}^{RC} \beta_i\right)}{\prod_{i=1}^{RC} \Gamma(\beta_i)} \right)^{T} \prod_{t=1}^{T} \frac{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)} + \beta_i)}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)} + \beta_i\right)} \underbrace{\int_{\phi_t} \frac{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)} + \beta_i\right)}{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)} + \beta_i)} \prod_{i=1}^{RC} \phi_{ti}^{n_{*i}^{(t)} + \beta_i - 1} d\phi_t}_{=1}. \tag{3.18}$$

With the multivariate integrals over θ_d and ϕ_t simplifying to yield 1, and the terms containing only α_t and β_i terms treated as constants, we have

$$P_1 \propto \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(n_{d*}^{(t)} + \alpha_t)}{\Gamma\left(\sum_{t=1}^{T} n_{d*}^{(t)} + \alpha_t\right)}; \; P_2 \propto \prod_{t=1}^{T} \frac{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)} + \beta_i)}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)} + \beta_i\right)}. \tag{3.19}$$

Isolate the terms corresponding to the assignments of the ℓ'-th block in the d'-th EEM sample:

$$P_1 \propto \prod_{d \neq d'} \frac{\prod_{t=1}^{T} \Gamma(n_{d*}^{(t)} + \alpha_t)}{\Gamma\left(\sum_{t=1}^{T} n_{d*}^{(t)} + \alpha_t\right)} \cdot \frac{\prod_{t=1}^{T} \Gamma(n_{d'*}^{(t)} + \alpha_t)}{\Gamma\left(\sum_{t=1}^{T} n_{d'*}^{(t)} + \alpha_t\right)};$$

$$P_2 \propto \prod_{t=1}^{T} \frac{\prod_{i \neq q_{d'\ell'}} \Gamma(n_{*i}^{(t)} + \beta_i) \cdot \Gamma(n_{*q_{d'\ell'}}^{(t)} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)} + \beta_i\right)}. \tag{3.20}$$

Incorporate the terms that do not depend on d' and $q_{d'\ell'}$ into the proportionality constants:

$$P_1 \propto \frac{\prod_{t=1}^{T} \Gamma(n_{d'*}^{(t)} + \alpha_t)}{\Gamma\left(\sum_{t=1}^{T} n_{d'*}^{(t)} + \alpha_t\right)}; \; P_2 \propto \prod_{t=1}^{T} \frac{\Gamma(n_{*q_{d'\ell'}}^{(t)} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)} + \beta_i\right)}. \tag{3.21}$$

Let us also define a value $\neg n$ the same way we defined n but without the count for the current ℓ'-th block in matrix $\widehat{\mathbf{x}}_{d'}$. Note that, $\neg n_{\cdot\cdot}^{(\cdot)} = n_{\cdot\cdot}^{(\cdot)}$, when the counts are independent of d' and ℓ'; otherwise, $n_{\cdot\cdot}^{(\cdot)} = \neg n_{\cdot\cdot}^{(\cdot)} + 1$. Next, isolate the terms that depend on $z_{d'\ell'}$:

$$P_1 \propto \frac{\prod_{t \neq z_{d'\ell'}} \Gamma(\neg n_{d'*}^{(t)} + \alpha_t) \cdot \Gamma(\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}} + 1)}{\Gamma\left(1 + \sum_{t=1}^{T} \neg n_{d'*}^{(t)} + \alpha_t\right)};$$

$$P_2 \propto \prod_{t \neq z_{d'\ell'}} \frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(t)} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} \neg n_{*i}^{(t)} + \beta_i\right)} \cdot \frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}} + 1)}{\Gamma\left(1 + \sum_{i=1}^{RC} \neg n_{*i}^{(z_{d'\ell'})} + \beta_i\right)}. \tag{3.22}$$

We also know that $\Gamma(u+1) = u\Gamma(u)$. Use these to get

$$P_1 \propto \frac{\prod_{t \neq z_{d'\ell'}} \Gamma(\neg n_{d'*}^{(t)} + \alpha_t) \cdot \Gamma(\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}}) \cdot (\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d',\ell'}})}{\Gamma\left(\sum_{t=1}^{T} \neg n_{d'*}^{(t)} + \alpha_t\right) \cdot \left(\sum_{t=1}^{T} \neg n_{d'*}^{(t)} + \alpha_t\right)};$$

$$P_2 \propto \prod_{t \neq z_{d'\ell'}} \frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(t)} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} \neg n_{*i}^{(t)} + \beta_i\right)} \cdot \frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} \neg n_{*,i}^{(z_{d'\ell'})} + \beta_i\right)} \cdot \frac{(\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}})}{\left(\sum_{i=1}^{RC} \neg n_{*,i}^{(z_{d'\ell'})} + \beta_i\right)}. \tag{3.23}$$

Fold the terms $\Gamma(\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}})$ and $\frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} \neg n_{*,i}^{(z_{d'\ell'})} + \beta_i\right)}$ back into the products:

$$P_1 \propto \frac{\prod_{t=1}^{T} \Gamma(\neg n_{d'*}^{(t)} + \alpha_t) \cdot (\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d',\ell'}})}{\Gamma\left(\sum_{t=1}^{T} \neg n_{d'*}^{(t)} + \alpha_t\right) \cdot \left(\sum_{t=1}^{T} \neg n_{d'*}^{(t)} + \alpha_t\right)};$$

$$P_2 \propto \prod_{t=1}^{T} \frac{\Gamma(\neg n_{*q_{d'\ell'}}^{(t)} + \beta_{q_{d'\ell'}})}{\Gamma\left(\sum_{i=1}^{RC} \neg n_{*i}^{(t)} + \beta_i\right)} \cdot \frac{(\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}})}{\left(\sum_{i=1}^{RC} \neg n_{*,i}^{(z_{d'\ell'})} + \beta_i\right)}. \tag{3.24}$$

Finally, incorporating the products over t into the proportionality constants, conclude that the conditional distribution in eq. (3.5) is proportional to

$$p(z_{d'\ell'} \mid z_{\neg(d'\ell')}, \widehat{\mathbf{X}}, \alpha, \beta) \propto \frac{\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}}}{\neg n_{d'*}^{(*)} + \sum_{t=1}^{T} \alpha_t} \cdot \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} + \sum_{i=1}^{RC} \beta_i}. \tag{3.25}$$

3.4 Latent Fluorescent Dirichlet Allocation (LFDA)

In fluorescence spectroscopy many variables have complex interactions affecting the measurements in many different ways. For example, higher temperatures cause faster diffusion and thus a larger amount of collisional quenching. On the other hand, higher temperatures also cause disassociation of weakly bound compounds and thus a smaller amount of static quenching [10]. Therefore, an increase in temperature may or may not cause a quenching effect depending on the actual temperature change, the specific quencher and the particular fluorophore. A model that can naturally incorporate these variables would offer a more reliable interpretation and could offer more accurate predictions. Extending the LDA model to incorporate such control variables can improve the flexibility of the model to deal with the factors that may impact the fluorescence of a sample.

3.4.1 Model Definition

The LFDA model is an extension of the model introduced in section 3.1. The new model incorporates a set of M control variables (pH, temperature, etc.). These variables can affect the measured fluorescence of a given sample through a variety of chemical and physical interactions. We assume each control variable is drawn from a multinomial distribution Λ_m with $m \in \overline{1,M}$ with dimension $|\Lambda_m|$ corresponding to the number of discretized values of the m-th control variable. Let us denote these control variable via $\xi_{dm} \in \overline{1,|\Lambda_m|}, d \in \overline{1,D}$. The extended graphical model can be seen in fig. 3.3.

Note that in this model the matrix Φ is no longer an $RC \times T$ matrix but a $RC \times T \times |\Lambda_1| \times |\Lambda_2| \times \cdots \times |\Lambda_M|$ multidimensional array. The distribution ϕ_t for aspect t over the RC matrix elements is dependent of the the observed variables ξ_{dm} with $m \in \overline{1,M}$ and $d \in \overline{1,D}$. This differs from our previous generative model in that the fluorescent block $q_{d\ell}$ is no longer picked from a probability distribution conditioned only on the aspect assignment $p(q_{d\ell}|\Phi, z_{d\ell})$, denoted as $p(q_{d\ell}|\phi_{z_{d\ell}})$, but from a distribution conditioned also on the control

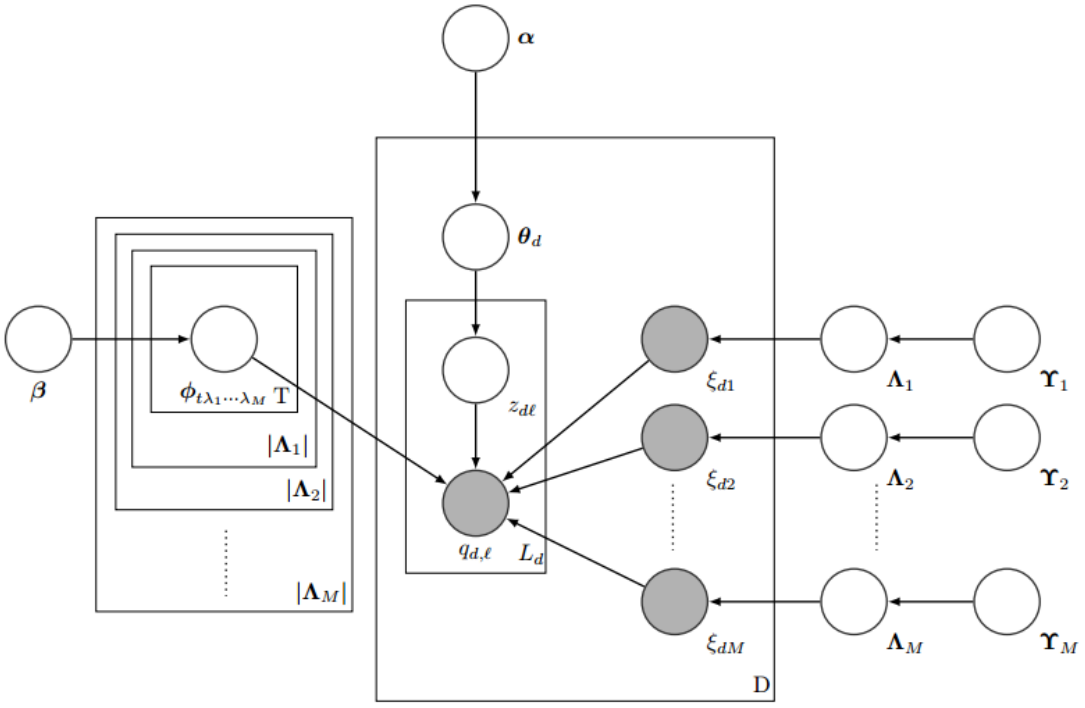


Figure 3.3: The LFDA model as an extension of the LDA model incorporating observed control variables.

variables $p(q_{d\ell}|\Phi, z_{d\ell}, \xi_{d1}, \xi_{d2}, \cdots, \xi_{dM})$, denoted as $p(q_{d\ell}|\phi_{z_{d\ell}\xi_{d1}\cdots\xi_{dM}})$. Thus, we must redefine the multinomial distributions $\phi_t = \{\phi_{ti}\}$ with $t \in \overline{1,T}$ and $i \in \overline{1,RC}$ previously defined in section 3.2.

3.4.2 Model Derivation

For each latent aspect enumerated via $t = \overline{1.T}$ and for each value of the control variables $\lambda_m \in \overline{1,|\Lambda_m|}$, let $\phi_{t\lambda_1\cdots\lambda_M} = \{\phi_{t\lambda_1\cdots\lambda_M i}\}$, $i \in \overline{1,RC}$ denote a multinomial distribution over the RC EEM matrix entries, conditioned on aspect t and control variables $\lambda_1, \cdots, \lambda_M$. As before we assume that $\phi_{t\lambda_1\cdots\lambda_M}$ is drawn from a Dirichlet distribution $Dir_{RC}(\beta = \{\beta_i\})$. Also in order to take a fully Bayesian approach we assume the multinomial distributions Λ_m have Dirichlet priors Υ_m with $m \in \overline{1,M}$.

Under this graphical model, the generative process of creating the d-th discrete EEM matrix takes the following form:

1. ***Latent aspect distributions in samples,*** $\phi_{t\lambda_1\cdots\lambda_M}$***:*** For each $t \in \overline{1,T}$ and for each value of the control variables $\lambda_m \in \overline{1,|\Lambda_m|}$, $m \in \overline{1,M}$, choose a multinomial distribution over the *RC* EEM matrix entries from $\phi_{t\lambda_1\cdots\lambda_M} \sim Dir_{RC}(\beta)$.

2. ***Multinomial control variables,*** Λ_m***:*** For each control variable choose a multinomial distribution $\Lambda_m \sim Dir_{|\Lambda_m|}(\Upsilon_m)$.

3. ***Latent aspect proportions,*** θ_d***:*** Choose a multinomial distribution over the T latent aspects from $\theta_d \sim Dir_T(\alpha)$.

4. ***Control variables*** ξ_{dm}***:*** Choose a value for each control variable by drawing it from its multinomial distribution $\xi_{dm} \sim Multi_{|\Lambda_m|}(\Lambda_m)$.

5. ***Number of fluorescence blocks,*** L_d***:*** Choose the number of fluorescence blocks from $L_d \sim Poisson(\vartheta)$.

6. Consider the fluorescence blocks $\ell = \overline{1,L_d}$. Assign $\ell \leftarrow 1$ and **BEGIN:**

   (a) ***Latent aspect assignment for each block,*** $z_{d\ell}$***:*** Choose a latent aspect from $z_{d\ell} \sim Multi_T(\theta_d)$.

   (b) ***Landscape of fluorescence blocks,*** $\{q_{d\ell}\}$***:*** For this latent aspect $z_{d\ell}$, choose the weight of the associated fluorescence block from $q_{d\ell} \sim Multi_{RC}(\phi_{z_{d\ell}\xi_{d1},\cdots,\xi_{dM}})$.

   (c) ***Fluorescence intensity,*** $\widehat{\mathbf{x}}_d = \{\widehat{x}_{d\ell}\}$***:*** Increase the fluorescence intensity at the location corresponding to $q_{d\ell}$ by incrementing the value of $\widehat{x}_{dq_{d\ell}}$ by 1.

   (d) If $\ell < L_d$, assign $\ell \leftarrow \ell + 1$ and repeat from Step 6a; otherwise, **STOP.**

The joint p.d.f. $p(\mathbf{Z}, \widehat{\mathbf{X}}, \Xi_1, \cdots, \Xi_M, \Lambda_1, \cdots, \Lambda_M \Theta, \Phi \,|\, \alpha, \beta, \Upsilon_1, \cdots, \Upsilon_M)$, where $\Xi_m = \{\xi_{dm}\}$ with $m \in \overline{1,M}$, $d \in \overline{1,D}$ factorizes according to the graph in fig. 3.3 as:

$$
\begin{aligned}
&p(\mathbf{Z},\widehat{\mathbf{X}},\Xi_1,\cdots,\Xi_M,\Lambda_1,\cdots,\Lambda_M\Theta,\Phi\,|\,\alpha,\beta,\Upsilon_1,\cdots,\Upsilon_M)\\
&=\left(\prod_{\lambda_1=1}^{|\Lambda_1|}\cdots\prod_{\lambda_M=1}^{|\Lambda_M|}\prod_{t=1}^{T}p(\phi_{t\lambda_1\cdots\lambda_M}\,|\,\beta)\right)\left(\prod_{d=1}^{D}p(\theta_d\,|\,\alpha)\left(\prod_{m=1}^{M}p(\Lambda_m|\Upsilon_m)p(\xi_{dm}|\Lambda_m)\right)\right.\\
&\qquad\left.\prod_{\ell=1}^{L_d}p(z_{d\ell}\,|\,\theta_d)\,p(q_{d\ell}\,|\,\phi_{z_{d\ell}\xi_{d1}\cdots\xi_{dM}})\right).
\end{aligned}
\tag{3.26}
$$

As in section 3.3.3, the parameters $\Lambda_m, m\in\overline{1,M}$, Θ and Φ can be integrated out and a sampler can be derived to find the value of the latent variables. However, note that knowledge of the variables Λ_m does not offer information about the chemical composition of each sample. Instead, these variables offer information about the distribution of the control parameters in the chemical experiment. In general, the analyst is interested in the way control variables affect fluorescence but not in the distribution of these variables. This fact can be used to simplify the distribution in eq. (3.26).

The objective is finding the joint distribution of $\widehat{\mathbf{X}}$ and $\mathbf{Z}$ from this new graphical model. Thus, given the control variables ξ_{dm}, the variables $q_{d\ell}$ and $z_{d\ell}$ are d-separated from the variables Λ_m and Υ_m. Therefore, the joint distribution of the graphical model can be rewritten as:

$$
\begin{aligned}
&p(\mathbf{Z},\widehat{\mathbf{X}},\Theta,\Phi\,|\,\Xi_1,\cdots,\Xi_M,\alpha,\beta)\\
&=\left(\prod_{\lambda_1=1}^{|\Lambda_1|}\cdots\prod_{\lambda_M=1}^{|\Lambda_M|}\prod_{t=1}^{T}p(\phi_{t\lambda_1\cdots\lambda_M}\,|\,\beta)\right)\left(\prod_{d=1}^{D}p(\theta_d\,|\,\alpha)\prod_{\ell=1}^{L_d}p(z_{d\ell}\,|\,\theta_d)\,p(q_{d\ell}\,|\,\phi_{z_{d\ell}\xi_{d1}\cdots\xi_{dM}})\right).
\end{aligned}
\tag{3.27}
$$

Note that as shown in section 3.3.3 integrating out the parameters Θ and Φ from eq. (3.27) requires separating the equation into two factors labeled P_1 and P_2. Factor P_1 contains the integral over Θ and has the same form and derivation process as shown in section 3.3.3. Factor P_2 has now the form:

$$\underbrace{\int_{\Phi}\left(\prod_{\lambda_1=1}^{|\Lambda_1|}\cdots\prod_{\lambda_M=1}^{|\Lambda_M|}\prod_{t=1}^{T}p(\phi_{t\lambda_1\cdots\lambda_M}\,|\,\beta)\right)\left(\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}p(q_{d\ell}\,|\,\phi_{z_{d\ell}\xi_1\cdots\xi_M})\right)d\Phi}_{P_2}. \tag{3.28}$$

Substituting the Dirichlet priors and the multinomial distributions

$$P_2=\int_{\Phi}\left(\prod_{\lambda_1=1}^{|\Lambda_1|}\cdots\prod_{\lambda_M=1}^{|\Lambda_M|}\prod_{t=1}^{T}\underbrace{\frac{\Gamma\left(\sum_{i=1}^{RC}\beta_i\right)}{\prod_{i=1}^{RC}\Gamma(\beta_i)}\left(\prod_{i=1}^{RC}\phi_{t\lambda_1\cdots\lambda_M i}^{\beta_i-1}\right)}_{p(\phi_{t\lambda_1\cdots\lambda_M}\,|\,\beta)}\right)\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\underbrace{\phi_{z_{d\ell}\xi_1\cdots\xi_M q_{d\ell}}}_{p(q_{d\ell}\,|\,\phi_{z_{d\ell}\xi_1\cdots\xi_M})}\,d\Phi. \tag{3.29}$$

In order to define this sampler, we will redefine the counting variable introduced in eq. (3.30) for factor P_2 as:

$$n_{di}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle=\sum_{\ell=1}^{L_d}\mathbb{1}_t(z_{d\ell})\cdot\mathbb{1}_i(q_{d\ell})\cdot\mathbb{1}_{\lambda_1\cdots\lambda_M}(\xi_{d1}\cdots\xi_{dM}), \tag{3.30}$$

Let us now consider the term $\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\phi_{z_{d\ell}\xi_1\cdots\xi_M q_{d\ell}}$ in P_2. As shown in eq. (3.16) we can observe that by exponentiating the probabilities $\phi_{t\lambda_1\cdots\lambda_M i}, t\in\overline{1,T}, i\in\overline{1,RC}, \lambda_m\in\overline{1,|\Lambda_m|}$ to the count variable that indicates how many times the probability appears in the factorization, the term can be rewritten as

$$\prod_{d=1}^{D}\prod_{\ell=1}^{L_d}\phi_{z_{d\ell}\xi_1\cdots\xi_M q_{d\ell}}=\prod_{\lambda_1=1}^{|\Lambda_1|}\cdots\prod_{\lambda_M=1}^{|\Lambda_M|}\prod_{t=1}^{T}\prod_{i=1}^{RC}\phi_{t\lambda_1\cdots\lambda_M i}^{n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle}. \tag{3.31}$$

Substitute eq. (3.31) in eq. (3.29) and switch the integral and the product operators:

$$P_2 = \int_{\Phi} \left(\prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \prod_{t=1}^{T} \frac{\Gamma\left(\sum_{i=1}^{RC} \beta_i\right)}{\prod_{i=1}^{RC} \Gamma(\beta_i)} \left(\prod_{i=1}^{RC} \phi_{t\lambda_1\cdots\lambda_M i}^{\beta_i - 1} \right) \right) \left(\prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \prod_{t=1}^{T} \prod_{i=1}^{RC} \phi_{t\lambda_1\cdots\lambda_M i}^{n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle} \right) d\Phi$$

$$= \prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \prod_{t=1}^{T} \int_{\phi_{t\lambda_1\cdots\lambda_M}} \frac{\Gamma\left(\sum_{i=1}^{RC} \beta_i\right)}{\prod_{i=1}^{RC} \Gamma(\beta_i)} \prod_{i=1}^{RC} \phi_{t\lambda_1\cdots\lambda_M i}^{n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i - 1} \, d\phi_{t\lambda_1\cdots\lambda_M}. \tag{3.32}$$

Following the steps used in eq. (3.18) the terms containing only β_i can be treated as constants. Noting that the integrals are unormalized Dirichlet distributions, multiply the inside of the integral by a normalization factor and the outside by the factor's reciprocal:

$$P_2 \propto \prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \prod_{t=1}^{T} \frac{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i)}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i\right)}$$

$$\underbrace{\int_{\phi_{t\lambda_1\cdots\lambda_M}} \frac{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i\right)}{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i)} \prod_{i=1}^{RC} \phi_{t\lambda_1\cdots\lambda_M i}^{n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i - 1} \, d\phi_{t\lambda_1\cdots\lambda_M}}_{=1}. \tag{3.33}$$

With the multivariate integral over $\phi_{t\lambda_1\cdots\lambda_M}$ simplifying to yield 1, we have

$$P_2 \propto \prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \prod_{t=1}^{T} \frac{\prod_{i=1}^{RC} \Gamma(n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i)}{\Gamma\left(\sum_{i=1}^{RC} n_{*i}^{(t)}\langle\lambda_1\cdots\lambda_M\rangle + \beta_i\right)} \tag{3.34}$$

Following the steps on eq. (3.22) through eq. (3.25) we can show

$$P_2 \propto \prod_{\lambda_1=1}^{|\Lambda_1|} \cdots \prod_{\lambda_M=1}^{|\Lambda_M|} \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})}\langle\lambda_1\cdots\lambda_M\rangle + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})}\langle\lambda_1\cdots\lambda_M\rangle + \sum_{i=1}^{RC} \beta_i}. \tag{3.35}$$

Isolating the terms that still depend on d':

$$P_2 \propto \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \sum_{i=1}^{RC} \beta_i} \prod_{\lambda_1 \neq \xi_{d'1}} \cdots \prod_{\lambda_M \neq \xi_{d'M}} \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} \langle \lambda_1 \cdots \lambda_M \rangle + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} \langle \lambda_1 \cdots \lambda_M \rangle + \sum_{i=1}^{RC} \beta_i}. \quad (3.36)$$

Incorporating the constant terms in the products into the proportionality constants:

$$P_2 \propto \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \sum_{i=1}^{RC} \beta_i}. \quad (3.37)$$

Finally, conclude that the joint distribution $p(\mathbf{Z}, \widehat{\mathbf{X}}, \Theta, \Phi \,|\, \Xi_1, \cdots, \Xi_M, \alpha, \beta)$ is proportional to:

$$p(z_{d'\ell'} | z_{\neg(d'\ell')}, \widehat{\mathbf{X}}, \Xi_1, \cdots, \Xi_M, \alpha, \beta) \propto \frac{\neg n_{d'*}^{(z_{d'\ell'})} + \alpha_{z_{d'\ell'}}}{\neg n_{d'*}^{(*)} + \sum_{t=1}^{T} \alpha_t} \cdot \frac{\neg n_{*q_{d'\ell'}}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \beta_{q_{d'\ell'}}}{\neg n_{**}^{(z_{d'\ell'})} \langle \xi_{d'1} \cdots \xi_{d'M} \rangle + \sum_{i=1}^{RC} \beta_i}. \quad (3.38)$$

This expression can be intuitively understood. The first term is the probability of observing aspect $z_{d'\ell'}$ in matrix d'; the second term is the probability of placing a fluorescent block at location $q_{d'\ell'}$ given aspect $z_{d'\ell'}$ and control variables $\xi_{d'1 \cdots \xi_{d'M}}$. This sampler would allow the modeling of fluorescent matrices while taking into account the other parameters that affect fluorescence.

Using this sampler we can run an experiment using dummy control valiables that affects the way in wich a fluorophore presnets in an EEM. Assuming an EEM that can be represented in a 6×6 matrix we generate a simple fluorescence representation, consisting of four fluorophores affected by a single control variable that can take 1 out of 9 states. Figure 3.4 shows all possible representations fof these dummy fluorophores under all passible 9 states of the control variable such that ϕ_{ab} represents fluorophore $t = a$ under control variable $\lambda_1 = a$.

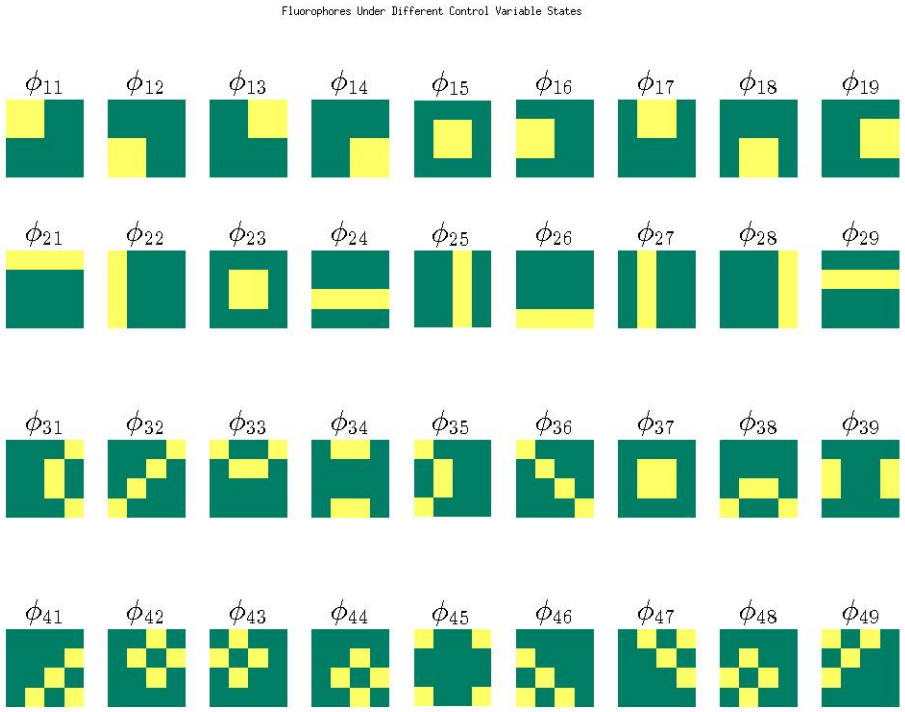


Figure 3.4: Ground truth structure of 4 fluorophores under a single control variable.

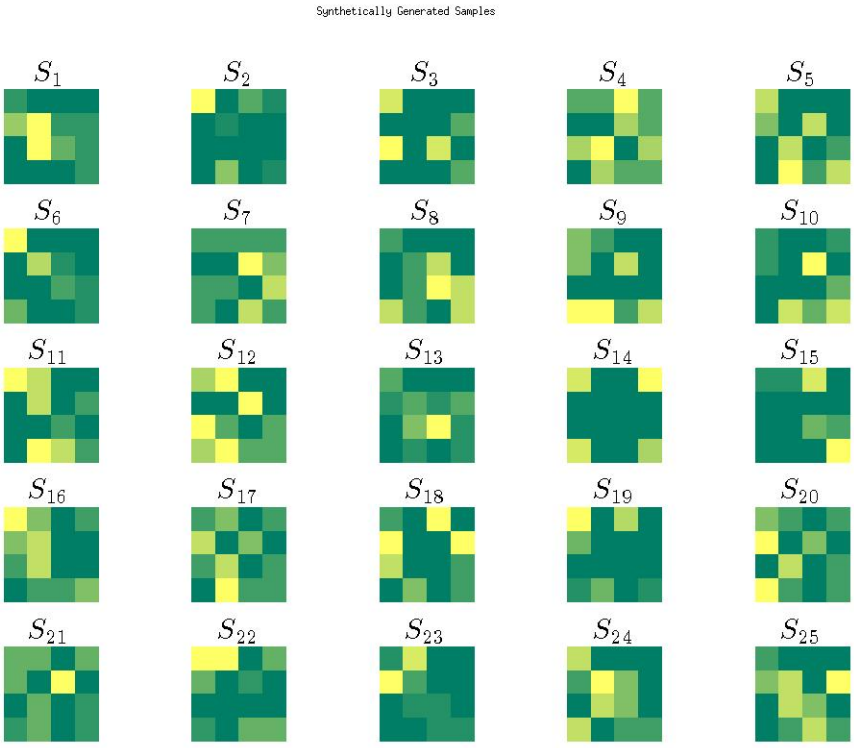


Figure 3.5: Samples generated from the underlying structure shown in 3.4.

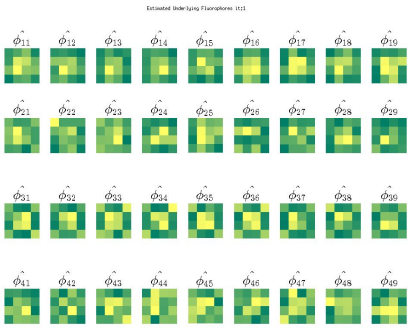
Having knowledge of the ground truth structure, we can generate samples by randomizing diffrent proportions of fluorophores under different values of the control variable λ_1 as shown in 3.5. Once the samples have been generated we run the sampler using 3.38, 3.6 shows the state of the sampler at diffrent stages of the calculation.

3.5 Classification Enhancement Technique (CET)

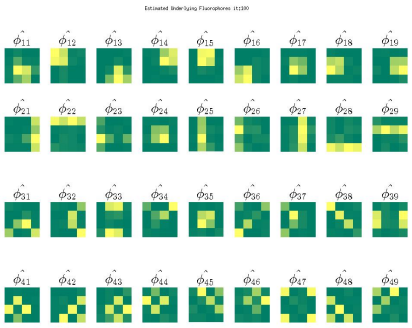
For the task at hand, we introduce a classification enhancement technique using the scatter matrix separability measure to enhance the difference between EEMs that represent different classes. We make use of this technique to highlight EEM features that might be more important for classification purposes and as a means of selecting an appropriate parameter δ for the discretization step in Section 3.1. It is worth mentioning that such a step is useful in classification problems but may introduce unwanted artifacts when the goal of the analysis is quantitative in nature. In a categorical classfication analysis we can have a clear picture of which areas of the spectrtum are the most relevant for the task. However, when the purpuse of the analysis is estimating fluorescent proportions or regresing a variable we simply make sure that the discretization parameter is constant and offers enough resolution to correctly repressent the full spectrum.

Assuming a binary classification problem, let us separate our original data set $\mathbf{X}$ into two three-way arrays $\mathbf{X}^{(1)} = \{x^{(1)}_{drc}\} \in \mathbb{R}^{D^{(1)} \times R \times C}$ and $\mathbf{X}^{(2)} = \{x^{(2)}_{drc}\} \in \mathbb{R}^{D^{(2)} \times R \times C}$ containing the samples for class 1 and 2, respectively. To establish which excitation-emission pairs are the most useful for classification purposes, we utilize the matrix $\Omega = \{\omega_{rc}\} \in \mathbb{R}^{R \times C}$ where

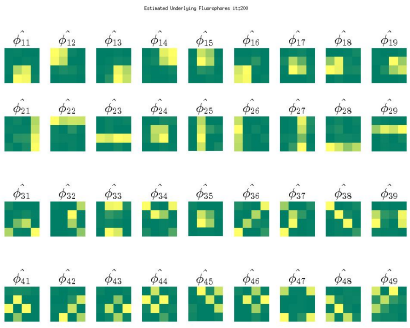
$$\omega_{rc} = \frac{P^{(1)}(\mu^{(1)}_{\cdot rc} - \mu_{\cdot rc})^2 + P^{(2)}(\mu^{(2)}_{\cdot rc} - \mu_{\cdot rc})^2}{P^{(1)}\sigma^{(1)^2}_{\cdot rc} + P^{(2)}\sigma^{(2)^2}_{\cdot rc}}. \tag{3.39}$$



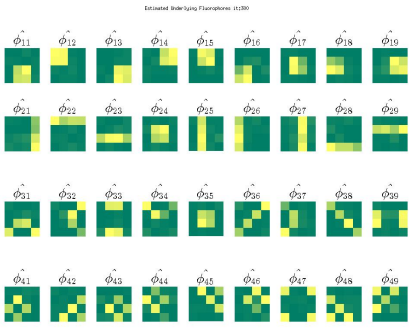
(a) Random initialization.



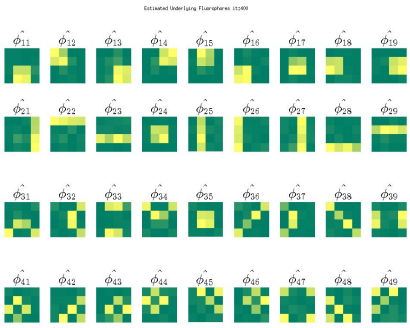
(b) Iteration 100.



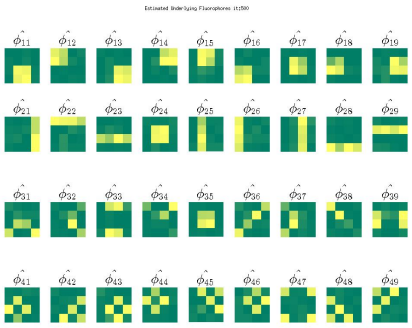
(c) Iteration 200.



(d) Iteration 300.



(e) Iteration 400.



(f) Iteration 500.

Figure 3.6

Here, for $i = 1, 2$,

$$\begin{aligned} \mu_{\cdot rc}^{(i)},\ \sigma_{\cdot rc}^{(i)^2} &= \text{mean and variance of samples } x_{drc}^{(i)} \text{ in } \mathbf{X}^{(i)}; \\ \mu_{\cdot rc} &= \text{mean of all samples } x_{\cdot rc} \text{ in } \mathbf{X}; \\ P^{(i)} &= \text{prior probability for class } i, \\ &\quad \text{i.e., } P^{(i)} = |D^{(i)}|/|D|. \end{aligned}$$

The numerator and denominator of eq. (3.39) are referred to as the *between-class scatter* and the *within-class scatter,* respectively. The value of ω_{rc} is a measure of *intra-class similarity* and *inter-class separability* defined as the ratio between intra-class and inter-class variance. Larger values of ω_{rc} indicate excitation-emission pairs that are highly descriptive and useful for classification purposes; smaller values indicate excitation-emission pairs with information that is not correlated with class membership.

Our objective is to find an optimal value for the quantization parameter δ so that small but informative entries in the matrix can be appropriately quantized with minimum loss of information. To accomplish this, we will first scale the matrix $\Omega = \{\omega_{rc}\}$ to generate $\Omega' = \{\omega'_{rc}\}$ such that its elements ω'_{rc} lie within the interval $(0, 1]$. We then use a location specific fluorescence quantization parameter δ_{rc} which is inversely proportional to ω'_{rc}:

$$\delta_{rc} \propto \frac{1}{\omega'_{rc}}. \tag{3.40}$$

This enables δ_{rc} to be adaptively changed based on the relevance of the location for classification. Locations having a higher ω'_{rc} value correspond to a lower quantization value, thus offering a higher resolution; locations having a lower ω'_{rc} value identifies areas of lower importance for the classification task, thus offering a higher quantization value which corresponds to a lower resolution.

Once all entries δ_{rc} are found, the EEM data set can be quantized to generate discrete EEM data as

$$\widehat{x}_{drc} = \left\lfloor \frac{x_{drc}}{\delta_{rc}} \right\rfloor. \tag{3.41}$$

Chapter 4

Minority Oversampling

Due to the nature of some application domains, it is common for many of the available datasets to be highly biased. These biased or skewed datasets are commonly referred to in literature as *imbalanced* datasets. The imbalanced learning problem is not new to the machine learning and data mining communities and several approaches to address it have been proposed in the past [48], [49], [50]. Nevertheless, some of these approaches change the underlying structure of the data, while others precipitate over-fitting [51], [52]. Learning from imbalanced data is a common and recurring problem that affects a wide range of application domains and causes a variety of difficulties for automated systems [53], [54], [55]. Thus, a method that preserves the underlying structure of the data and avoids over-fitting, while synthetically balancing the dataset, is highly desirable.

In real world applications, imbalanced datasets are common in a variety of fields including fraud detection [53], [56], text classification [54] and medical diagnosis [57]. Given that most classification mechanisms assume balanced training sets, imbalanced datasets can pose a significant challenge, especially when a high precision classification of the minority class is desired. Obtaining a balanced dataset is not always practical and in certain cases can be even impossible. In fields such as fraud detection and network intrusion detection, imbalanced datasets are the norm and approaches to efficiently and soundly address this issue are necessary. While the machine learning and data mining communities have devel-

oped some classifiers to make do with imbalanced data, it is advantageous to synthetically balance a dataset in order to apply existing and proven learning methods, which are usually sensitive to class imbalance.

4.1 Previous Work

Most of the techniques in the machine learning literature that address the class imbalance problem can be classified into two categories: 1) *Data balancing methods,* which change the training data so that a balanced training set can be used by the classifier; and 2) *algorithmic methods,* which modify the classifier so that misclassifications of minority class instances have a higher penalty. The idea behind both approaches is to artificially bias the classifier or the dataset in such a way as to neutralize the original bias present in the training data. In this paper, we will focus on data balancing methods, which can be used in conjunction with any existing standard learning algorithm.

4.1.1 Random Under-Sampling and Oversampling Methods

The simplest data balancing methods make use of random oversampling, random undersampling, or a mixture of both. Suppose a training set $T = (\mathbf{x}_i, y_i)$, $i = 1, ..., n$, where $\mathbf{x}_i = [x_1, x_2, ..., x_k]$ is a k-dimensional feature vector and $y_i \in 1, ..., C$ is a class label associated with feature vector $\mathbf{x}_i$, is given. Let $T_{mi} \subset \{T\}$ and $T_{ma} \subset \{T\}$ be the subsets representing the minority and majority class respectively such that $T_{mi} \cap T_{ma} = \{\emptyset\}$ and $T_{mi} \cup T_{ma} = \{T\}$. Furthermore, the sets generated by sampling randomly from T_{mi} and T_{ma} will be referred to as S_{mi} and S_{ma} respectively.

Random oversampling techniques balance the dataset by replicating randomly selected samples from T_{mi} and thus increase the underrepresented class by $|S_{mi}|$. This method, though simple, can be very effective in cases where data is not noisy and classes are clearly separable. Nevertheless, random oversampling can precipitate over-fitting, greatly limiting a classifier's ability to generalize [58], [59].

Random undersampling techniques balance the dataset by randomly removing samples from the majority class and thus decreasing the number of samples in the overrepresented class to $|T_{ma}| - |S_{ma}|$. As is the case with random oversampling, random undersampling can be effective, but suffers from a very specific shortcoming that can be extremely problematic. Namely, the removal of data from the training set may result in the deletion of crucial information from the dataset. Randomly removed samples may be more important for concept learning and pattern recognition than the samples left in the balanced dataset [60].

The shortcomings of random undersampling and oversampling are the main reasons *informed undersampling* methods were developed. We will not discuss these methods here for space considerations but, for the interested reader, a thorough explanation and study of these and other methods can be found in [52].

4.1.2 Clustering Methods

A class can usually be represented in feature space as a heterogeneous or a homogeneous concept. A heterogeneous concept is a conglomeration of sub-concepts or disjoint clusters in feature space, representing a single class. A homogeneous concept can be seen as a class that can be represented as a single cluster in feature space. Cluster-based oversampling methods were developed to deal with general imbalanced learning while also addressing the problem known as the *within-class imbalance.* The within-class imbalance problem describes a situation in which certain sub-clusters of a heterogeneous concept are under-represented in the training set [52].

Clustering methods such as *cluster-based oversampling (CBO)* [61] attempt to identify these sub-clusters and oversample the ones that are underrepresented in order to restore within-class balance. In this approach, a clustering algorithm such as k-means is used on samples of a single class. Once the clusters have been defined and the number of samples in each cluster are calculated, the oversampling is done so that the smallest clusters are inflated to match the size of the largest one. By using cluster based methods only in the minority

class, the within-class imbalance problem and the between-class imbalance problem can be addressed simultaneously.

4.1.3 SMOTE : Synthetic Minority Oversampling Technique

The SMOTE class balancing method [62] has had great success in several applications, producing improvements to precision and recall that supersede those of other oversampling techniques.

The SMOTE algorithm creates a set of synthetic new samples S_{mi} by using the similarities between exiting minority samples. During the oversampling procedure, the k nearest neighbors for each sample $\mathbf{x}_i \in T_{mi}$ are selected by picking the K samples possessing the smallest Euclidean distance to sample $\mathbf{x}_i$. A new sample is then generated via a convex combination of a randomly selected neighbor from this K-neighborhood of $\mathbf{x}_i$ and $\mathbf{x}_i$ itself:

$$\mathbf{x}_{new} = \alpha \mathbf{x}_i + (1-\alpha)\mathbf{x}_j, \tag{4.1}$$

where $\mathbf{x}_j$ is a random sample selected from the K-neighborhood of $\mathbf{x}_i$, $\mathbf{x}_{new}$ is the newly generated sample, and α is a random number in the interval $[0,1]$. So, the new sample is situated somewhere on the line that joins sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$.

This type of synthetic oversampling helps address the severe over-fitting introduced by regular oversampling, while still balancing the dataset and frequently improving learning. However, the SMOTE algorithm has a propensity to over generalize when data is noisy and/or when the class distributions overlap in feature space [63]. Several variations to the original SMOTE algorithm have been proposed, such as adaptive synthetic sampling (ADA-SYN) [63], Borderline-SMOTE [64], and SMOTEBoost [65]. These techniques attempt to generate sets of samples that lie close to the decision boundary by trying to formulate probability distributions which give a higher weight to borderline samples or simply by heuristically selecting a subset of borderline samples to pass through a SMOTE algorithm. The idea behind these techniques is that the samples that lie closer to the decision boundary carry more information and are more important to concept learning than other

samples. However, it is worth mentioning that these techniques, like the original SMOTE, can hinder the learning algorithm in cases where the data is noisy and/or where there is considerable class overlap.

4.1.4 Data Cleaning Methods

Some data cleaning methods have been used in conjunction with oversampling techniques in order to address the problem of class overlap [66], [67], [68]. *Tomek links,* defined as those samples whose nearest neighbor belongs to the opposite class, are an excellent example of these methods. More formally, let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the Euclidean distance between samples $\mathbf{x}_i$ and $\mathbf{x}_j \in T^{\tilde{i}}$, where $T^{\tilde{i}}$ is the set including all samples except for sample $\mathbf{x}_i$. The pair $(\mathbf{x}_i, \mathbf{x}_j)$ is called a Tomek link if the minimum distance $d(\mathbf{x}_i, \mathbf{x}_j)$ given a sample $\mathbf{x}_i \in T_{mi}$ corresponds to a sample $\mathbf{x}_j \in T_{ma}$.

Once a pair that forms a Tomek link is identified, the pair is removed from the dataset as it is considered to be noisy or an ambiguous borderline sample. This school of thought believes that removing samples from these borderline areas leads to better defined clusters and boundaries that will facilitate the classification task.

Some studies have been performed combining oversampling techniques and other data cleaning methods such ass one-sided selection (OSS) [66], condensed nearest neighbor rule (CNN) [69], and neighborhood cleaning rule (NCL) [68], reporting favorable results.

4.2 The Proposed Approach: MeMO

4.2.1 The Basic Concept

Our membership-based minority oversampling method merges concepts from cluster-based methods and synthetic oversampling techniques. The method we propose addresses both, the within-class imbalance and between-class imbalance problems, while implementing the oversampling techniques used in the SMOTE algorithm in order to avoid the over-fitting problem present in other oversampling techniques.

The first step in our approach uses a k-means clustering algorithm to divide the minority class samples in T_{mi} into V clusters $T_{mi}^{\nu} \subset T_{mi}$, $\nu = 1, ..., V$. Then, we select which cluster to sample by randomly selecting one cluster from a probability distribution that is inversely proportional to the number of samples in each cluster. The probability of selecting the cluster T_{mi}^{ν} is:

$$p(T_{mi}^{\nu}) = 1 - \frac{|T_{mi}^{\nu} \cup S_{mi}^{\nu}|}{|T_{mi} \cup S_{mi}|}, \tag{4.2}$$

where $|T_{mi}^{\nu} \cup S_{mi}^{\nu}|$ is the number of samples in cluster ν and $|T_{mi} \cup S_{mi}|$ is the total number of minority class samples, including original samples and newly created ones. Initially, when additional samples have not yet been created, both S_{mi}^{ν} and S_{mi} are empty sets.

The imbalance present within the clusters of the minority class is adaptively restored by giving smaller clusters a higher probability of being sampled. However, it is worth noting that once within-class balance is restored cluster selection converges to equal probability for every cluster.

Each sample in a cluster is assigned a *membership* value proportional to the distance of the sample to the center of its cluster. Once a cluster has been chosen (with probability $p(T_{mi}^{\nu})$ determined according to eq. (4.2), we select two samples from this cluster. Each sample will be chosen with a probability proportional to its membership. In other words, we perform a membership-based selection. We assume that the samples closer to the cluster center are more representative of the cluster's "concept" and we assign a probability of selection that is proportional to this membership. Let $d(\mathbf{c}_{\nu}, \mathbf{x}_i)$ be the Euclidean distance between the cluster center $\mathbf{c}_{\nu}$ and sample $\mathbf{x}_i$. Then, the probability of picking sample $\mathbf{x}_i$ is

$$p(\mathbf{x}_i) = \frac{d(\mathbf{c}_{\nu}, \mathbf{x}_i)}{\sum_{\mathbf{x}_j \in T_{mi}^{\nu}} d(\mathbf{c}_{\nu}, \mathbf{x_j})}. \tag{4.3}$$

Once two samples have been selected according to this probability distribution, we use eq. (4.1) to generate a new minority class sample. Note that unlike other SMOTE

variants, we favor sampling near the cluster centers as we believe these samples to be more representative of the minority class, especially in cases of considerable class overlap.

Other SMOTE-like oversampling techniques use the k-NN approach to select the samples to use in eq. (4.1). However, if a minority class is composed of several concepts, it is a non trivial problem to find the best number k of nearest neighbors. The selected value of k must be large enough to generalize the concept but not so large as to introduce unwanted artifacts. It must also be a value that will be small enough to minimize noise, but not so small as to increase the risk of over-fitting.

In other words, the best value of k for a specific locality might not be optimal to the totality of the feature space. The cluster-based oversampling strategy that we propose allows the algorithm to tailor its function to specific areas of the feature space. If an optimum number of clusters is identified through a technique such as cross-validation, the above k-NN related difficulties no longer apply.

4.2.2 Example

To demonstrate this point, we created a toy dataset with two attributes for easy visualization. The data was created by generating 12 random clusters, 6 clusters per class, containing a total of 440 samples, such that the majority class (red) has 300 samples and the minority class (blue) has 140 samples.

The toy domain training set can be seen in Fig. fig. 4.1 and the outcomes of performing MeMO and SMOTE on this dataset can be seen in Fig. 4.2a and in Fig. fig. 4.2b, respectively. MeMO keeps the oversampling limited to the convex hull of each cluster, thus maintaining the underlying structure of the data. On the other hand, SMOTE and SMOTE-like approaches can introduce artifacts, thereby increasing the risk of overlap and noise.

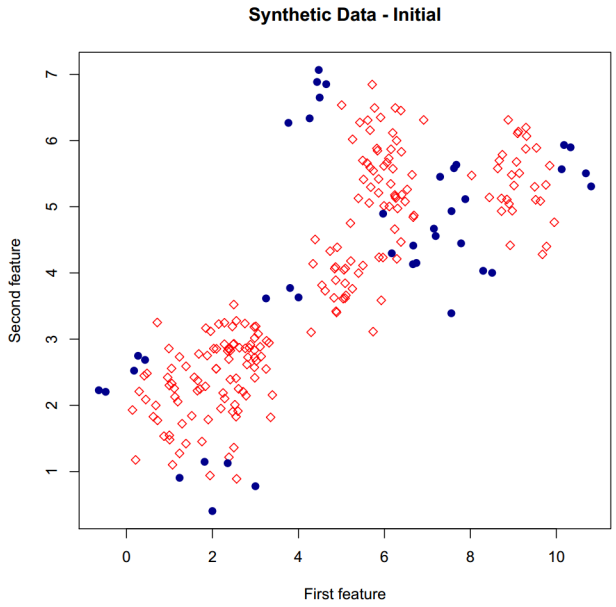
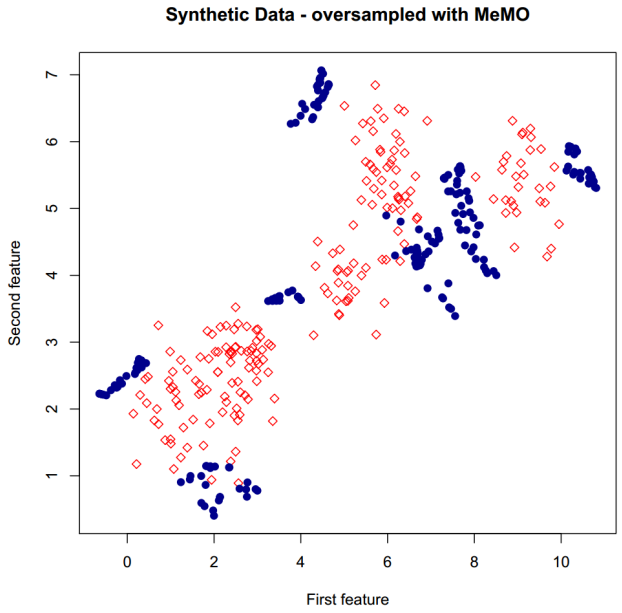
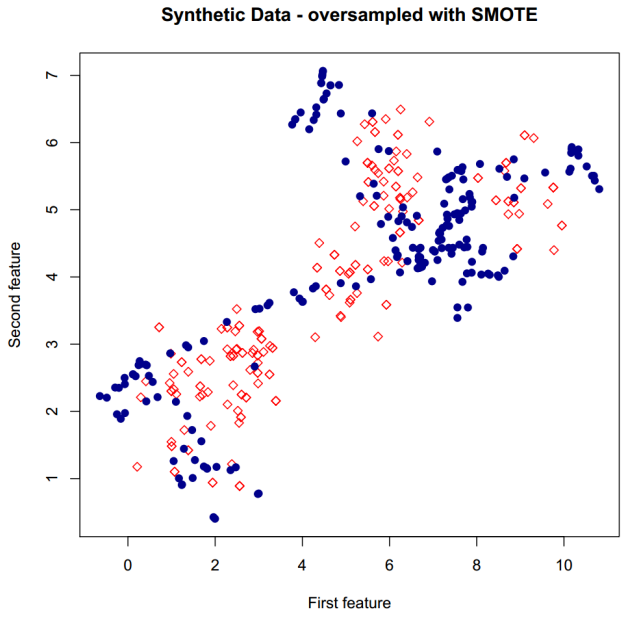


Figure 4.1: Point distribution for initial synthetic dataset.



(a) With MeMO.



(b) With SMOTE.

Figure 4.2: Point distribution after minority class oversampling of synthetic dataset.

Chapter 5

Non-parametric Regression

Nonparametric(NP) estimation techniques allow data to model the relationships among variables, thus, it possesses the ability to detect structures which sometimes remain undetected by traditional parametric estimation techniques[70]. A histogram, the most primitive of non-parametric density estimation techniques, can reveal information that would be hidden under parametric approaches. For example, a histogram is capable of discovering multimodality in a dataset that would be lost with an approach assuming a Gaussian distribution. However, a crude histogram approach has it's limitations, namely, it is highly dependent on the bucket size (or bandwidth) and it is also discontinuous which implies that the gradient along with other important information is lost [71]. These drawbacks lead to the natural conclussion of using a continuous and smooth estimator which consequently lead to the preferred estimators used in literature known as kernel density estimators.

In NP regression, the regression functions are estimated using a local sample for each point, hence the term local estimators. In contrast, parametric estimators are global estimators and in many applied scenarios these regressors enter the conditional mean linearly or with 'judicious' choice of parametric functional form[71]. NP kernel regression estimators use local fits to construct the global function estimator preserving the local unique features inherent to the dataset.

5.1 Overview

In this work we will also explore multivariate density estimators given that univariate density estimation is limited. As an example a univariate density estimation can show bimodality that is not necesarily inherent to the data but rather a dependence on a secondary variable. A well documented example of this drawback was published in [72], where The authors have a visually bimodal univariate density (202 observations) of lean body mass. Subsequent analysis shows that the bimodal nature of this density is linked to the gender of the individual. By splitting the data into 100 men and 102 women, each individual density is strongly unimodal [71].

There are several techniques to find the nonparametric kernel estimator that can compute the smooth function of interest. Here we will present the two most popular methods, local-linear least-square (LLLS) and local-constant least-squres (LCLS) estimator. Theoretically the LLLS estimator has advantages over LCLS providing more accurate measurements of the conditional mean and also an estimator of its first derivative. Using the notation in [71] we can elaborate more on nonparametric estimators. Let a particular maltivariate data point $\mathbf{x}$ be a q-tuple $q \in \mathbb{N}$, such that $\mathbf{x} = (x_1, x_2, \cdots, x_q)$. Then a multivariate kernel $K(\mathbf{x})$ on $\mathbb{R}^q \to \mathbb{R}$ has to satisfy, $K(\mathbf{x}) \geq 0$, $\forall \mathbf{x} \in \mathbb{R}^q$, and $\int_{\mathbb{R}^q} K(\mathbf{x}) d\mathbf{x} = 1$. A kernel which satisfies this requirement is given in (5.1).

$$K_h(\mathbf{x}_i, \mathbf{x}) = \prod_{d=1}^{q} k\left(\frac{x_{id} - x_d}{h_d}\right) \tag{5.1}$$

where, h_d refers to as the bandwidth(which is diferent to the meaning in spectral analysis) of d^{th} dimension and $k(\cdot)$ could be any univariate kernel, such as, uniform, Epanechnikov, biweight, triweight, Gaussian etc ....

Let $f_{\mathbf{x},y}(\cdot)$ denote the joint density of $(\mathbf{x}, y)$ and $f_{\mathbf{x}}(\cdot)$ denote the marginal density of $\mathbf{x}$, whereas the estimators of joint and marginal densities by $\hat{f}_{\mathbf{x},y}(\mathbf{x}, y)$ and $\hat{f}_{\mathbf{x}}(\mathbf{x})$. Assume that we were given a dataset with n instances, $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ with corresponding labels

$\{y_1, \cdots, y_n\}$. The estimators of the marginal and joint densities are given in (5.2) and (5.3) as:

$$\hat{f}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{n|\mathbf{h}|} \sum_{i=1}^{n} K_h(\mathbf{x}_i, \mathbf{x}) \tag{5.2}$$

where $|\mathbf{h}| = h_1 h_2 \cdots h_q$, the product of the q bandwidths and,

$$\hat{f}_{\mathbf{x},y}(\mathbf{x}, y) = \frac{1}{n|\mathbf{h}|h_y} \sum_{i=1}^{n} K_h(\mathbf{x}_i, \mathbf{x}) k\left(\frac{y_i - y}{h_y}\right) \tag{5.3}$$

where h_y is the smoothing parameter associated with y. Applying the statistical definition of conditional densities, $f_{y|\mathbf{x}}(y|\mathbf{x}) = f_{\mathbf{x},y}(\mathbf{x}, y)/f_{\mathbf{x}}(\mathbf{x})$, with $\hat{f}_{\mathbf{x},y}(\mathbf{x}, y)$ and $\hat{f}_{\mathbf{x}}(\mathbf{x})$, we can write an expression for $\hat{f}_{y|\mathbf{x}(y|\mathbf{x})}$, the estimate of conditional density as in (5.4).

$$\hat{f}_{y|\mathbf{x}}(y|\mathbf{x}) = \frac{1}{nh_y} \sum_{i=1}^{n} A_i(\mathbf{x}) k\left(\frac{y_i - y}{h_y}\right), \tag{5.4}$$

where

$$A_i(\mathbf{x}) = \frac{K_h(\mathbf{x}_i, \mathbf{x})}{n^{-1} \sum_{j=1}^{n} K_h(\mathbf{x}_j, \mathbf{x})}. \tag{5.5}$$

Hence, the nonparametric estimator of the conditional expectation $\hat{E}(y|\mathbf{x})$ (or $\hat{m}(\mathbf{x})$) can be given as in (5.6).

$$\begin{aligned} \hat{E}(y|\mathbf{x}) &= \int y \hat{f}(y|\mathbf{x}) dy \\ \hat{m}(\mathbf{x}) &= \sum_{i=1}^{n} A_i(\mathbf{x}) y_i. \end{aligned} \tag{5.6}$$

The selection of the form of $m(\hat{\mathbf{X}})$, referred to hence forth as a, makes the difference between local-constant estimator and local-linear estimator. When a local-constant estimator is used, we minimize a kernel weighted least-square regression of y on a constant as in (5.7).

$$\min_{a} \sum_{i=1}^{n} [y_i - a]^2 K_h(\mathbf{x}_i, \mathbf{x}). \tag{5.7}$$

In a local-linear estimator approach, instead of minimizing a constant, we try to fit a line locally as in (5.8).

$$\min_{a,b} \sum_{i=1}^{n} [y_i - a - (\mathbf{x}_i - \mathbf{x})b]^2 K_h(\mathbf{x}_i, \mathbf{x}). \tag{5.8}$$

5.2 Non-parametric Regression for Microtox Prediction

We use the popular, LLLS estimator [73], over the LCLS estimator (also known as Nadaraya-Watson estimator[74, 75]). In our analysis y represents the percentage drop of Microtox® measurement and $\mathbf{x}$ presensents EEM scores obtained from LDA, LFDA or PARAFAC accompanied with other measurements, namely pH, Conductivity, Dissolved Oxygen and Turbidity. The conditional expectation $\hat{E}(y|\mathbf{x})$ (or $\hat{m}(\mathbf{x})$) is estimated using training data. During the testing phase the non parametric regression will predict the drop in Microtox® measurements using a generalized product kernel [70].

Chapter 6

Fusion-based Classification

Preliminary experiments conducted using PARAFAC and LDA/LFDA, as techniques to analyze EEMs, have shown promising classification results. The combination of both techniques seems to improve overall classification. Thus, we have developed a system that can make predictions using the PARAFAC and LDA/LFDA models independently and in parallel, providing independent class predictions that can subsequently be fused using Dempster-Shaffer theory.

6.1 Dempster-Shafer Theory

Probability theory does not offer a consice and mathematically strict way of representing ignorance or lack of information. Assuming evidence exists for A to occurr with probability $P(A) = 0.6$ the additivity axiom of probability theory imposes that $P(\bar{A}) = 1 - P(A)$. In this case the evidence for A might be satisfactory but the evidence for $\bar{A}$ is being modeled directly from A there is no way to introduce a lack of knowledge or evidence for $P(\bar{A})$. To handle lack of evidence or knowledge a relaxation of the additivity axiom is usually used. One approcah used in literature to fuse informtion using this method is known as Depmster-Shafer theory [76].

Consider a set of mutually exclusive and exhaustive propositions, $\Theta = \{\theta_1, ..., \theta_n\}$, referred to as the *frame of discernment (*FoD*)* representing the "scope of expertise" of some

problem domain. A proposition θ_i, referred to as a *singleton*, represents the lowest level of discernible information and the elements in 2^{Θ}, the power set of Θ, form all propositions of interest. We refer to any proposition that is not a singleton, e.g. (θ_1, θ_2), as a *composite*. For simplicity the cardinality of Θ is assumed to be finite and it is denoted by $|\Theta|$.

DS theory models beliefs by assigning to any set $A \subseteq \Theta$, a numeric value $m(A) \in [0,1]$. The mapping $m : 2^{\Theta} \mapsto [1,0]$ is a *basic belief assignment* (*BBA*) or *mass structure* if $m(\emptyset) = 1$ and $\sum_{A \subseteq \Theta} m(A) = 1$. The *BBA* constitutes the counterpart to the probability measure in probability theory. However, in Dempster-shafer theory the masses can be assigned to non singleton prepositions. The mass is free to move through all the prepositions in Θ creating the notion of ignorace or lack of evidence. In this way evidence commited to a preposition does not inherently imply that the remaining support should be commited to the negation of said preposition. In DS theory, lack of support for a preposition implies support for all others.

The state of complete ignorance is known as a vacuous *BBA* and it can be modeled as:

$$m(A) = 1_{\Theta} = \begin{cases} 1, & \text{for } A = \Theta \\ 0, & \text{for } A \in \Theta \end{cases} \tag{6.1}$$

Propositions possessing a non-zero BBA are referred to as *focal elements* and the set of focal elements is referred to as the *core* and is denoted by $\mathscr{F}$. The triple $\{\Theta, \mathscr{F}, m\}$ is *the body of evidence* (*BoE*) and $|\mathscr{F}|$ is its corresponding number of focal elements.

The set $A \setminus B$ denotes all singletons in $A \subseteq \Theta$ that are not included in $B \subseteq \Theta$ i.e., $A \setminus B = \{\theta_i \in \Theta \mid \theta_i \in A, \theta_i \notin B\}$; $\bar{A} = \Theta \setminus A$. Propositions for which there is no information are not assigned an a priori mass; therefore, as we mentioned previously, committing support for an event $A \subseteq \Theta$ does not imply committing support for its complement $\bar{A}$.

6.1.1 Belief and Plausability

The belief that directly supports a given proposition $A \subseteq \Theta$ is quantified via the *belief function* $Bel : 2^{\Theta} \mapsto [1,0]$, where

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{6.2}$$

and the *plausibility*, which quantifies the upper bound of evidence that can support a given proposition is given by the function $Pl : 2^{\Theta} \mapsto [0,1]$, where

$$Pl(A) = 1 - Bel(\bar{A}) \tag{6.3}$$

Thus, $m(A)$ measures the support assigned to proposition A only; $Bel(A)$ quantifies the support from all proper subsets of A, representing the mass that can move into A without ambiguity; and $Pl(A)$ represents the mass that can move into A and elsewhere, indicating the extent to which one finds A plausible. Note that when $\mathscr{F}$ is composed exclusively of singletons, all the above notions collapse to probability theory

$$m(A) = Bl(A) = Pl(A) = P(A), \forall A \in \mathscr{F} \tag{6.4}$$

Using the above notions of *belief* and *plausability* we can also define the notions of *doubt* and *uncertainty* as:

$$Dou(A) = Bel(\bar{A}) \tag{6.5}$$

$$Un(A) = [Bel(A), Pl(A)] \tag{6.6}$$

respectively, where $[Bel(A), Pl(A)]$ is the interval of the uncertainty associated with the reposition A.

6.1.2 Evidence Combination

Dempster's rule of combination (*DRC*) is the most popular and widely used combination strategy in DS theory. *DRC* makes it possible to fuse the information from several indepen-

dent BoEs that span the same FoD to form a single BoE. Consider two BoEs $\{\Theta, \mathscr{F}_i, m_i\}$, $i = \overline{1,2}$, that span the same FoD Θ. Then,

$$K_{12} = \sum_{B \in \mathscr{F}_1, C \in \mathscr{F}_2 \, C \cap B = \emptyset} m_1(B) m_2(C) \tag{6.7}$$

quantifies the conflict between the evidence present in both BoEs. The amount of conflict can give us information about the compatibility of the BoEs:

1. If $K_{12} = 1$ the two BoEs are incompatible.

2. If $K_{12} < 1$ the two BoEs are compatible.

3. If $K_{12} = 0$ the two BoEs are completely compatible.

The *DRC* approach works only if the combined BoEs are compatible. The fused BBA $m(\cdot) : 2^{\Theta} \mapsto [0,1]$ can be obtained as follows:

$$m(A) = \sum_{B \in \mathscr{F}_1, C \in \mathscr{F}_2 \, C \cap B = A} \frac{m_1(B) m_2(C)}{1 - K_{12}}, \forall A \subseteq \Theta. \tag{6.8}$$

This fusion is denoted as $m(\cdot) = (m_1 \oplus m_2)(\cdot)$ which is the orthogonal sum of m_1 and m_2. The $\oplus$ operator is both associative and commutative allowing the straightforward combination of of multiple BoEs. A variation of the DRC which accounts for evidence reliability makes use of a discounting factor. The BoE would be updated using the reliability of the source as hown bellow

$$\hat{m}_k(A) = \begin{cases} b_k m(A) & \text{for } A \subset \Theta \\ (1 - b_k) + b_k m_k(\Theta) & \text{for } A = \Theta \end{cases} \tag{6.9}$$

Where $b_k \in [0,1]$ is referred to as the *discounting factor* [77]. Note that $b_k = 0$ indicates a fully unreliable source while $b_k = 1$ indicates a fully reliable source.

6.2 DS-Model

In the scope of this work we have experimented with two main methods for EEM analysis. Namely, PARAFAC, LDA and the extension of LDA to EEMs with control variables LFDA. As it will be discussed in 7 these methods can complement each other and can show robustness when used in tandem. Therfore, a method that can offer a reliable and mathematically sound way of fusing the evidence gathered form these different approaches is highly desirable. This section explains the model used to express the results from these methods as a BoE in the context of examining water quality using Microtox measurements.

The scores obtain from PARAFAC-based and LDA-based (LDA/LFDA) models, are used in a non-parametric regression analysis (Section 5) to predict the percentage decrease in Microtox measurements. A Dempster-Shafer mass structure will then be fitted to each prediction and will be referred to as $m_{PA}(\cdot)$, $m_{LDA}(\cdot)$ and $m_{LFDA}(\cdot)$ for the PARAFAC-based LDA-based and LFDA-based models, respectively.

Let us select a threshold value π to discriminate between samples which are acceptable and samples which are not. A frame of discernment can then be defined as $\Theta = \{A, U\}$, where A stands for acceptable and U stands for unacceptable. The regressed variable, henceforth referred to as x, will be compared to the threshold π and the mass structure can then be defined using the following equation:

$$m(i) = \begin{cases} \beta & \text{for } i = A\,, x \geq \pi \\ 0 & \text{for } i = A\,, x < \pi \\ \beta & \text{for } i = U\,, x < \pi \\ 0 & \text{for } i = U\,, x \geq \pi \\ 1-\beta & \text{for } i = \Theta \end{cases} \tag{6.10}$$

where β is given by:

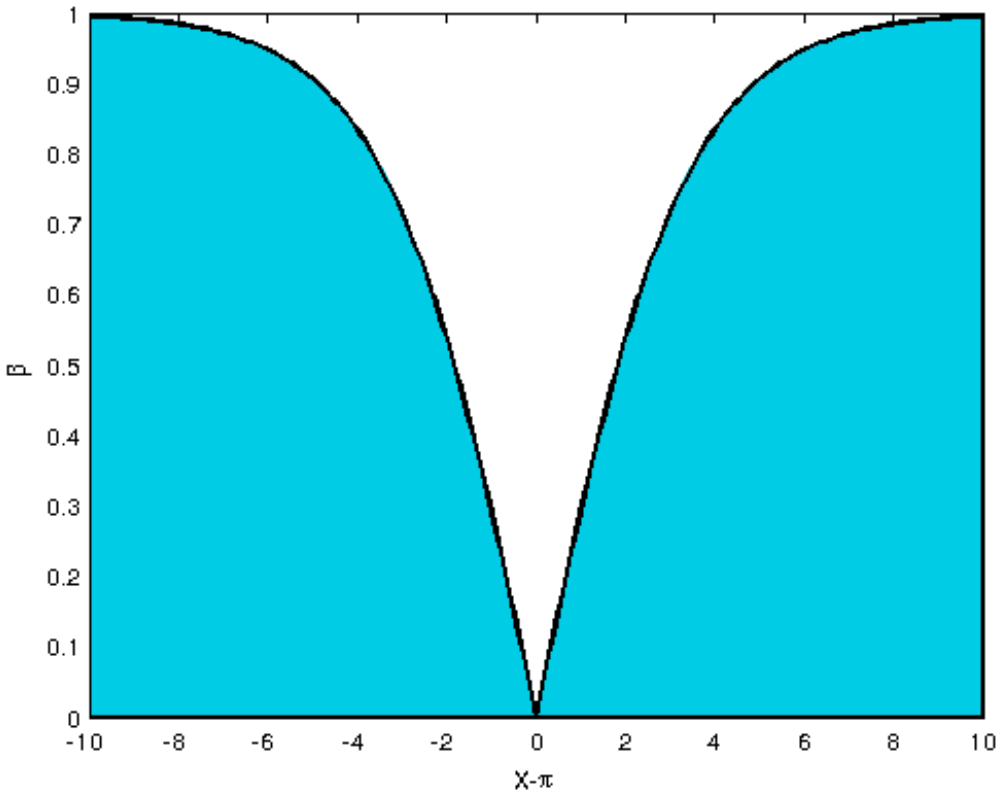


Figure 6.1: Plot of the function β with $\sigma = 5$.

$$\beta = \left| \frac{2}{1+e^{-\sigma x}} - 1 \right| \tag{6.11}$$

where σ controls how fast the uncertainty flows into Θ as x converges to π. Fig. 6.1 shows a graph of how the value of β changes as a function of x.

In order to pool the evidences and test the best performing combination with fused mass structures, we fuse $m_{PA}(\cdot)$ with $m_{LDA}(\cdot)$ and $m_{PA}(\cdot)$ with $m_{LFDA}(\cdot)$ to obtain $(m_{PA} \oplus m_{LDA})(\cdot)$ and $(m_{PA} \oplus m_{LFDA})(\cdot)$ respectively, using (6.8). Results and experiments of using this approach will be reviewd and discussed in 7.4.2

Chapter 7

Results

7.1 MeMO

MeMo creates less noise in cases of class overlap, while allowing an adaptive enlargement of the minority class clusters. The results show that MeMO compares favorably to other techniques in cases of considerable class overlap. However, when class overlap is less, MeMO produces comparable results to SMOTE.

7.1.1 Testbeds

Experiments were performed using the PIMA Indians diabetes dataset, the breast cancer Wisconsin dataset, the LandSat satellite image dataset and a two dimensional toy domain for illustration purposes. In order to quantify the degree of imbalance in each dataset we will use the following *imbalance measure:*

$$d = 1 - \frac{2|T_{mi}|}{|T|}. \tag{7.1}$$

Note that $d = 0$ when the cardinality of the minority set $|T_{mi}|$ is half the cardinality of the entire training set $|T|$. In other words, the imbalance measure d vanishes only when the training set is perfectly balanced. On the other hand, $d = 1$ in the extreme case when the dataset has no minority samples and is completely imbalanced, i.e., $|T_{mi}| = 0$.

The properties of the datasets along with their corresponding imbalance measures are shown in Table 7.1.

Table 7.1: Test beds

		# of instances		
Dataset	**# of attributes**	**Training**	**Testing**	d **(Training)**
Breast cancer (UCI)	30	469	100	0.71
PIMA (UCI)	8	568	200	0.60
LandSat (UCI)	36	4435	2000	0.79
Toy (synthetic)	2	240	200	0.67

7.1.2 Classification Results

The experiments were performed 10 times for each dataset and the samples used in training and testing were randomly chosen during each iteration. The values of k for SMOTE and L for MeMO were set by using cross-validation on the training set. The linear SVM and SMOTE implementations were obtained from the R package "e1071" and "DMwR" respectively. The ROC curves corresponding to the entries in Table 7.1 can be seen in Fig. 7.1, Fig. 7.2, Fig. 7.3 and Fig 7.4.

In Fig. 7.1 we can see the ROC plots and the area under the curve (AUC) for the experiments on the breast cancer Wisconsin dataset. It can be clearly seen that MeMO and SMOTE produce similar results while undersampling actually worsens the prediction. In this particular example the AUC is very high even for the unbalanced dataset. This allows us to conclude that classes are more separable permitting an easier prediction even for the imbalanced case.

Fig. 7.2 shows the results for the PIMA dataset. In this figure we can see that MeMO produces a better peformance than SMOTE and undersampling. It can be seen from the trends of the ROC plots that the classes are not easily separable in feature space. In cases like these, SMOTE can cause the classifier to overgeneralize which makes MeMO a better option.

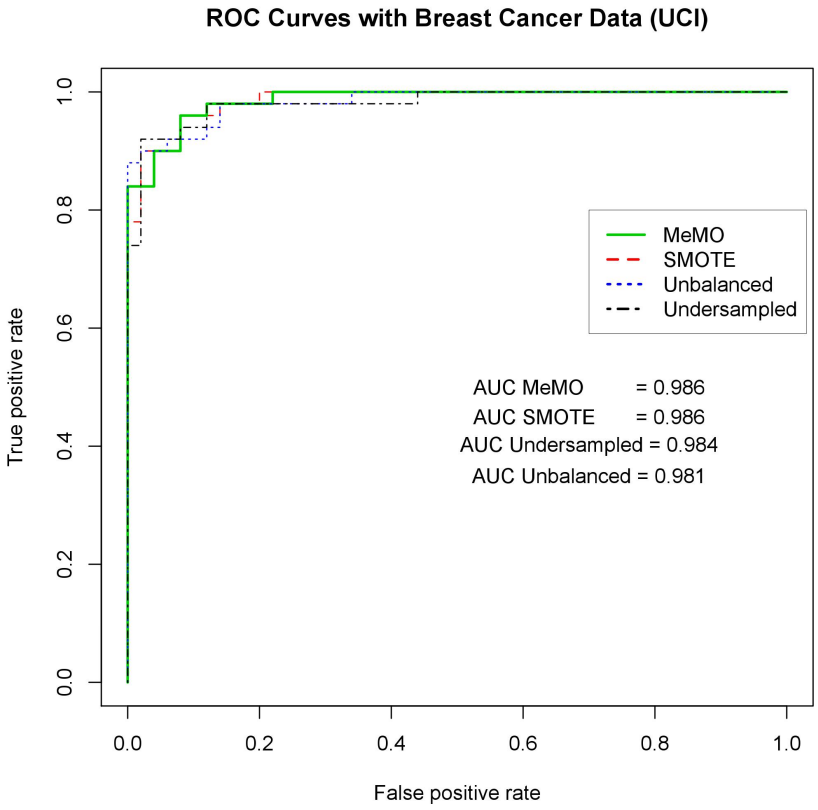


Figure 7.1: ROC curve for Breast Cancer Wisconsin dataset (UCI)

In Fig. 7.3 we can see the ROC and AUC results for the LandSat data. In this figure we can see the clear disadvantages of undersampling. In cases of extreme class imbalance, the undersampling necessary to balance the dataset is harmful for classification. In this case, MeMO shows a slightly better improvement over SMOTE.

Fig. 7.4 shows results obtained for the toy dataset. This domain was synthetically created to have large class overlap and to demonstrate the issues that arise from the *k*-NN approach, as mentioned in Section 4.2. Therefore, it's not a surprise that MeMO compares favorably to SMOTE in this domain.

A t-test was used to evaluate the statistical significance of our results using MeMO and SMOTE. Our results were comparable to SMOTE in the breast cancer dataset where

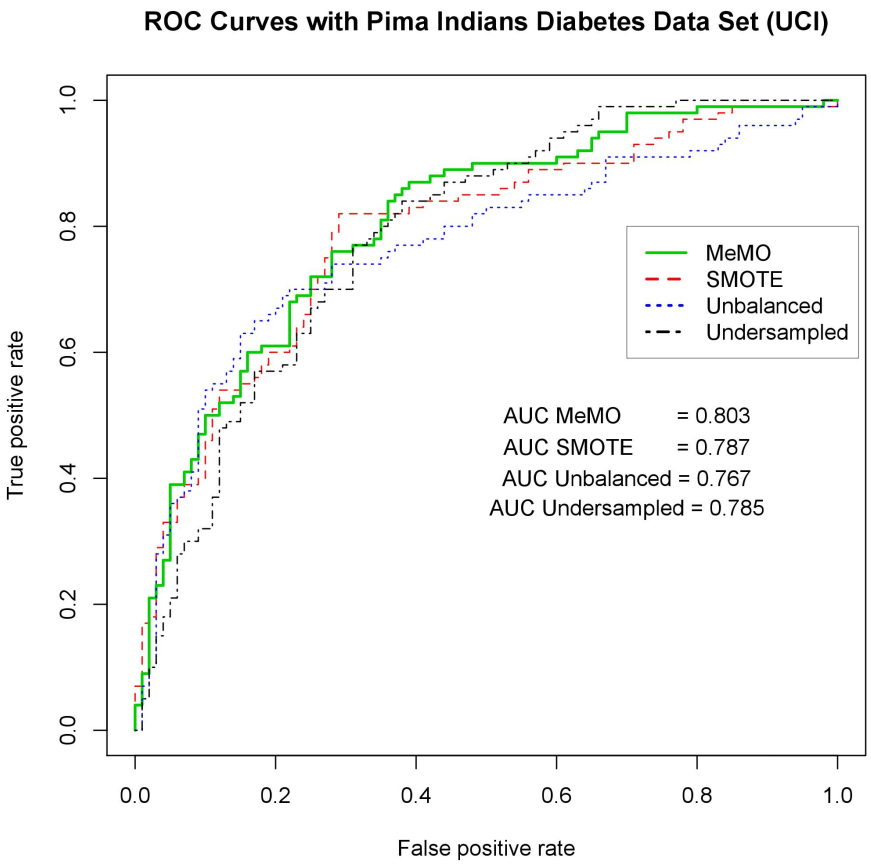


Figure 7.2: ROC curve for Pima Indians Diabetes dataset (UCI)

there seems to be less overlap between classes and the performance of the classifiers is consistently high. In the PIMA dataset we obtained a p-value of 0.0137 and in the Land-Sat dataset we obtained a p-value of 0.00015 with 95% confidence of rejecting the null hypothesis, showing the improvement to be statistically significant in both datasets.

Given the fact that we could control the degree of imbalance present in the toy dataset, we thought it would be interesting to see how SMOTE and MeMO compared to each other as the imbalance was increased systematically. Table 7.2 shows AUC values for SMOTE and MeMO for a toy domain starting at an value of $d = 0.9$ and proceeding to a value of $d = 0.1$.

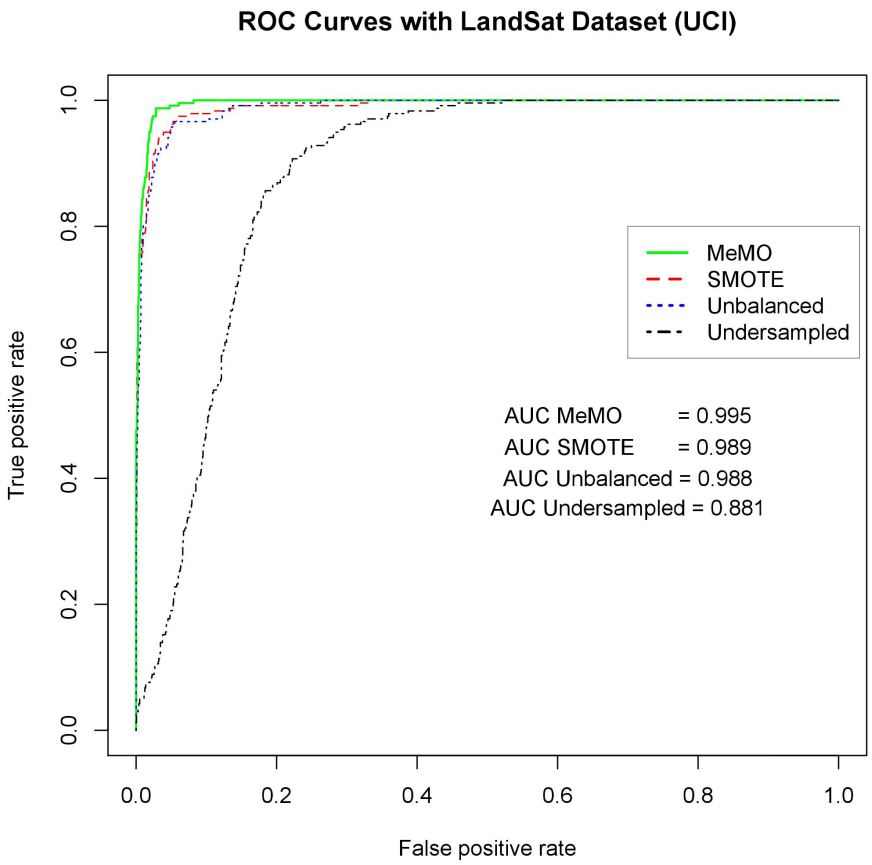


Figure 7.3: ROC curve for LandSat dataset (UCI)

In Table 7.2 we can observe that at higher degrees of imbalance MeMO tends to compare favorably to SMOTE. We believe that this behaviour is due to the fact that at high values of d, SMOTE is prone to overgeneralizing. The lack of neighboring minority samples forces SMOTE to pick minority samples that are too far away, unwillingly causing an increase in noise. On the other hand, MeMO preserves cluster integrity and prevents overgeneralizing under high degrees of imbalance. We can also observe in Table 7.2 that at lower values of d, SMOTE seems to outperform MeMO. This result seems to suggest that different strategies might work better given different degrees of imbalance. Thus, a k-NN approach could be more effective when dealing with a slightly imbalance dataset while a

Table 7.2: Area under the curve of ROC curves for different degree of imbalances with *MeMO* and *SMOTE*.

	Area under the curve	
Degree of imbalance	**MeMO**	**SMOTE**
0.9	0.65	0.59
0.8	0.70	0.67
0.7	0.80	0.77
0.6	0.88	0.84
0.5	0.88	0.85
0.4	0.89	0.86
0.3	0.89	0.88
0.2	0.89	0.90
0.1	0.89	0.91

cluster based approach might be more appropriate when working with a dataset with a high value of d.

7.2 Classification Using PARAFAC and LDA

7.2.1 Performance Criteria

Traditional binary classification tasks use performance measures that are based on the percentage of correctly or incorrectly labeled samples. Such performance measures can be misleading, especially in cases of class imbalance or when the prediction of one class is more important than the other. In statistics, a more complete understanding of the classifier can be obtained by analysing Type I and Type II errors. Type I errors, also known as false positives (FP), occur when a classifier gives a positive label to a negative sample. Type II errors, also known as false negatives (FN), occur when a classifier gives a negative label to a positive sample. On the other hand, positive samples that are correctly classified are called true positives (TP) and negative samples that are correctly classified are called true negatives (TN).

We make use of *precision (Pr)* and *recall (Re)* measures, commonly used in information retrieval literature, because of their ability to offer a more complete characterization of

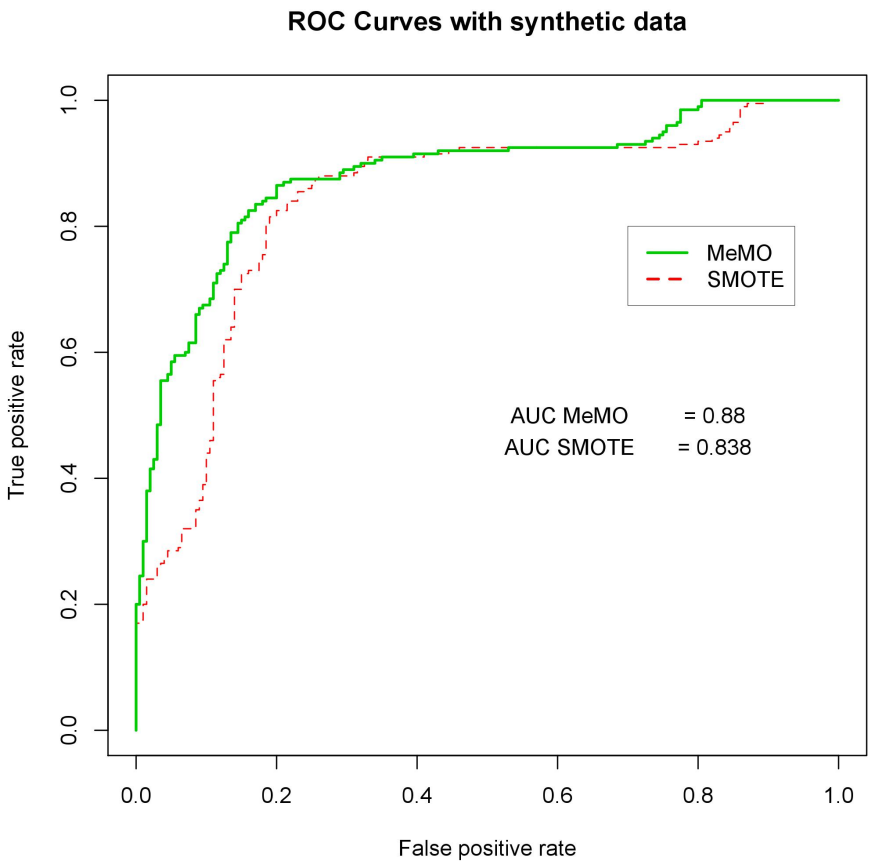


Figure 7.4: ROC curve for synthetic dataset

a predictor in terms of Type I and Type II errors. For simplicity let us assume we are interested only in one particular class. We can define *precision (Pr)* and *recall (Re)* as:

$$Pr = \frac{TP}{TP+FP}, \; Re = \frac{TP}{TP+FN} \tag{7.2}$$

A measure that combines these two into a single metric is the harmonic mean, also called the F_1-measure proposed by Vilar *et al.* [78]:

$$F_1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{7.3}$$

Table 7.3: Micro Averaging and Macro averaging of Precision and Recall

	Macro (M)	Micro (μ)
Precision	$\frac{\sum_{i=1}^{C} Pr_i}{C}$	$\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}$
Recall	$\frac{\sum_{i=1}^{C} Re_i}{C}$	$\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i}$
F_1	$\frac{2 \times Pr_M \times Re_M}{Pr_M + Re_M}$	$\frac{2 \times Pr_\mu \times Re_\mu}{Pr_M + Re_\mu}$

In the multi-class case, each sample can belong to more than one class at the same time. We want to average the performance over all classes in order to have a global performance metric for the predictor. Godbole and Sarawagi [79] proposed two different ways to obtain a global performance measure. The first one, micro-averaging, calculates the above measures by summing over each individual decision. The second one, macro-averaging, calculates the average of the above measures over all classes .

Table 7.3 shows the formulas to calculate these measures for both approaches, where C is the number of classes in the classification problem, Pr_i and Re_i are the precision and recall measured for class i and finally TP_i, FN_i and FP_i denote the values of the basic performance variables for class i.

Given that the proportion of samples that belong to each class is very similar in the cancer and the synthetic datasets, macro-averaging was used to obtain global performance measures. On the other hand, class proportions are varied in the fluorophore dataset; thus, micro averaging was used to obtain a more accurate global performance measure.

7.2.2 Testbeds

For testing and validation of our proposed method, we conducted several experiments using three different datasets. The first two datasets were made available at `http://www.models.life.ku.dk/datasets` by the Department of Food and Science at the Uni-

Table 7.4: Description of Data Sets

Data Set	Excitation Wavelength (nm)			Emission Wavelength (nm)			# of Classes
	Start	End	Increment	Start	End	Increment	
Cancer	250	450	5	300	600	1	4
Fluorophores	230	320	5	230	500	2	6
Synthetic	250	450	1	300	500	1	5

versity of Copenhagen. The third dataset is a toy domain that we have synthetically created in order to introduce different types of noise and artifacts to test the robustness of our approach. Table 7.4 summarizes the datasets we used.

Cancer Data Set: The first dataset contains samples from undiluted human blood plasma obtained from a study conducted at six Danish hospitals. The samples were taken from patients undergoing large bowel endoscopy due to symptoms associated with colorectal cancer. The dataset also contains samples from three control groups, namely, healthy subjects, subjects with other non-malignant diagnosis and subjects with pathologically verified adenomas.

Fluorophores Dataset: The second dataset contains a total of six different fluorophores: catechol, hydroquinone, indole, resorcinol, tryptohpane and tyrosine. The samples contain different combinations of fluorophores at different concentrations. We used this dataset to test the robustness of our approach to fluorescent noise by artificially and systematically reducing the peak-signal-to-noise-ratio (PSNR) of the matrices. Five different versions of the dataset were created, each one with a lower PSNR than the previous dataset. Precision and recall values were recorded for each fluorophore at the different noise levels.

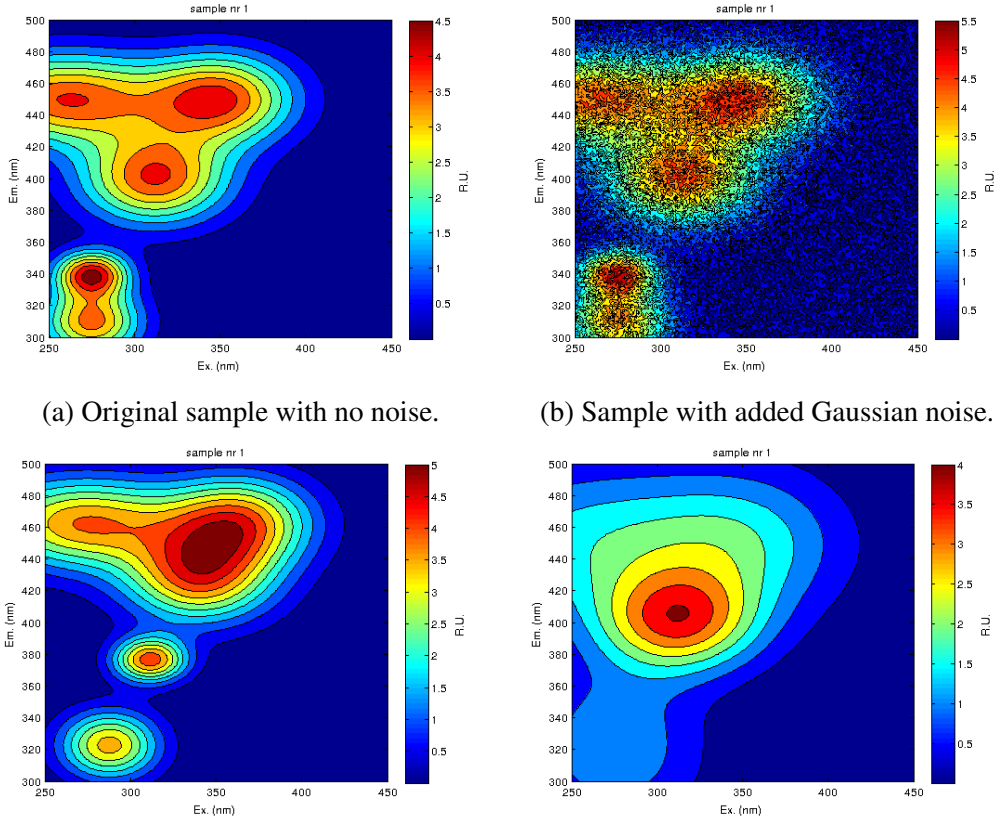
Table 7.5: Major Fluorescent Components With Peaks Designated With Letters as in Coble *et al.* [80]

Peak	Ex_{max} (nm)	Em_{max} (nm)	Remarks
B	275	310	Tyrosine-like, protein-like
T	275	340	Tryptophan-like, protein-like
A	260	380-460	Humic-like
M	312	380-420	Marine humic-like
C	350	420-480	Humic-like

Synthetic Dataset: We created a toy domain using the peak locations for the major fluorescent components reported in [80]. Table 7.5 shows the peak locations and their corresponding compound type. In order to test the robustness of our algorithm, we systematically increased the noise present in the dataset in three different ways. The first noise artifact added was white noise, in order to decrease the PSNR of the spectrum. The second noise artifact added was peak location variability, which will decrease the similarity between samples that belong to the same class while increasing class overlap. Finally, the third noise artifact added was girth and height variability for each peak, in order to try to mimic real world difficulties such as quenching and background fluorescence. Figures 7.5a to 7.5d show a sample under the three different noise conditions.

The samples with white noise were obtained by adding a random number in the interval $[0, n_{WN}]$, where n_{WN} is the white noise parameter. The samples with peak location variability were obtained by shifting the peak location of each sample by a random amount in both the excitation and emission directions. The shift is limited to a square of area $n_{SN} \times n_{SN}$ centered at the location of the original peak, where n_{SN} is the shift noise parameter. Finally, the samples with girth and height variation were obtained by adding variability to the variance of the underlying Gaussian used to create the peak. The variability was obtained by adding a number in the interval $[0, n_{HN}]$, where n_{HN} is the girth/height noise parameter.

NetZero Water treatment system Dataset: The residential urban ambient net-zero water treatment system was presented in [81]. The treatment system proposes a design using a membrane bioreactor, iron-mediated aeration (IMA, reported previously), vacuum ultrafiltration, and peroxone advanced oxidation, with minor rainwater make-up and H_2O_2 disinfection residual. This system neither takes nor releases water off-site, but rather purifies the wastewater by filtration, aeration, oxidation and other methods. The quality of the treatment process is monitored at different stages of the treatment system by taking measurements of pH, turbidity, conductivity, dissolved oxygen, temperature, EEM and



(a) Original sample with no noise.

(b) Sample with added Gaussian noise.

(c) Sample with shifted peak positions.

(d) Sample with girth and height changed.

Figure 7.5: An EEM sample subjected to different types of noise. Figure 7.5a shows a noiseless sample containing all fluorescent components in the toy domain. Figure 7.5b shows the same sample with added Gaussian noise. Figure 7.5c shows the sample after the peak positions have been randomly changed causing some peaks to overlap. Figure 7.5d shows the sample after a random variation of the girth and height of the peaks has been applied.

Microtox® to name a few. The data provided from this treatment facility is used to test the LFDA model using the aditional measurements of pH, turbidity, conductivity, dissolved oxygen and temperature as control variables. The interested reader can find further information about the water treatment mechanism in [81]. Microtox® is a testing system used to examine water, soil and air. The system uses non-pathogenic bioluminescent bacteria (Vibrio fischeri) to detect toxic substances invitro. The bacteria emit light as part of their regular metabolism. When exposed to toxic substances that disrupt their respiration the bacteria undergo a drop in luminescence that can be measured and correlated to the level of toxicity found in a substrate. The residential urban ambient net-zero water treatment process [81] uses an analysis of Microtox® measurements on water samples collected at different stages of the purification process. The treatment also uses a collection of fluorescence spectroscopic data along with other water quality measurements. It is worth mentioning that Microtox® testing protocols are labor intensive and require specially trained technicians to conduct the test, which can be very sensitive to lab conditions. This offers an opportunity to apply machine learning techniques to assess or predict Microtox® reading and the water quality while reducing Microtox® testing time, training and complexity. LFDA has been utilized for modeling and analysis of fluorescent spectroscopic excitation-emission matrices (EEMs) along with parallel factor analysis (PARAFAC).

7.2.3 Results

The data sets used for testing contain samples belonging to more than one class. Therefore, classification was performed using the one-versus-all method. This approach tends to create a very slight class imbalance which as reported in [82] can cause problems in probabilistic graphical models such as LDA. The reported experiments show that rare topics, or in our case rare latent aspects are not accurately modeled when underrepresented in the training dataset. To test if this was indeed the case we tested one class from each dataset used at different degrees of imbalance d, by using MeMO to balance the training samples. The

results can be seen in Table 7.6, Table 7.7 and Table 7.8. We can note that the balancing step does not improve or change the performance of the cancer dataset, however, the balancing step seems to offer some improvement on the synthetic dataset and the Fluorophore dataset.

	Colorectal Cancer Dataset			
d	**Other**	**Cancer**	**Adenoma**	**None**
0	0.43 ± 0.07	0.45 ± 0.06	0.38 ± 0.09	0.43 ± 0.08
0.1	0.44 ± 0.06	0.47 ± 0.05	0.39 ± 0.08	0.43 ± 0.09
0.2	0.42 ± 0.08	0.44 ± 0.07	0.38 ± 0.08	0.44 ± 0.07
0.3	0.41 ± 0.05	0.46 ± 0.06	0.37 ± 0.09	0.43 ± 0.08
0.4	0.45 ± 0.06	0.44 ± 0.04	0.36 ± 0.07	0.42 ± 0.09
0.5	0.43 ± 0.07	0.45 ± 0.06	0.37 ± 0.06	0.42 ± 0.07

Table 7.6: LDA F_1 Prediction performance at different levels of imbalance

	Synthetic Dataset				
d	**Peak M**	**Peak T**	**Peak C**	**Peak B**	**Peak A**
0	0.94 ± 0.15	0.69 ± 0.12	0.98 ± 0.04	0.87 ± 0.11	0.92 ± 0.06
0.1	0.93 ± 0.16	0.68 ± 0.15	0.97 ± 0.06	0.83 ± 0.14	0.95 ± 0.09
0.2	0.91 ± 0.14	0.69 ± 0.10	0.94 ± 0.07	0.80 ± 0.14	0.92 ± 0.06
0.3	0.87 ± 0.19	0.65 ± 0.08	0.96 ± 0.09	0.85 ± 0.21	0.89 ± 0.13
0.4	0.87 ± 0.21	0.66 ± 0.13	0.93 ± 0.07	0.78 ± 0.17	0.83 ± 0.09
0.5	0.85 ± 0.17	0.67 ± 0.12	0.95 ± 0.04	0.78 ± 0.21	0.86 ± 0.11

Table 7.7: LDA F_1 Prediction performance at different levels of imbalance

	Fluorophore Dataset					
d	**Hydroquinone**	**Catechol**	**Indole**	**Resorcinol**	**Tryptophane**	**Tyrosine**
0	0.97 ± 0.03	0.85 ± 0.05	0.95 ± 0.05	0.76 ± 0.10	0.94 ± 0.09	0.62 ± 0.06
0.1	0.97 ± 0.06	0.85 ± 0.05	0.95 ± 0.03	0.76 ± 0.08	0.94 ± 0.11	0.61 ± 0.10
0.2	0.96 ± 0.05	0.85 ± 0.07	0.91 ± 0.07	0.73 ± 0.13	0.91 ± 0.07	0.60 ± 0.11
0.3	0.96 ± 0.06	0.79 ± 0.09	0.91 ± 0.09	0.76 ± 0.14	0.91 ± 0.10	0.61 ± 0.14
0.4	0.92 ± 0.10	0.79 ± 0.11	0.88 ± 0.05	0.76 ± 0.11	0.91 ± 0.09	0.59 ± 0.10
0.5	0.92 ± 0.13	0.79 ± 0.10	0.88 ± 0.07	0.71 ± 0.13	0.88 ± 0.13	0.58 ± 0.13

Table 7.8: LDA F_1 Prediction performance at different levels of imbalance

From the experiments presented in chapter 4, we can conclude that for our level of imbalance and for the multi-class nature of the datasets involved, MeMO is a better choice than other oversampling techniques.

Membership-based oversampling partitions the minority samples into K clusters and then generates new samples by creating convex combinations of sample-pairs from the

same minority cluster. In order to apply this method to the problem at hand, we first generate a similarity matrix, $\mathbf{S} = \{s_{i,j}\}$, where the Euclidean norm is used to generate the similarity score between samples i and j:

$$s_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2}{2\sigma^2}\right). \tag{7.4}$$

Here, σ controls the extent to which the value of the matrix norm $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ is given a high or low similarity score.

We then use a spectral clustering algorithm described in [83] to partition the minority samples into K clusters and consequntly apply MeMO. The training data set is balanced in this manner to ensure that all classes are equally represented for model building in PARAFAC and LDA.

In the following experiments, the feature vectors used to train the classifiers are the PARAFAC model's scores and the LDA model's aspect proportions. The PARAFAC + LDA combination uses a concatenation of the PARAFAC model's scores and the LDA model's aspect proportions. We use the term *raw data* to refer to the direct concatenation of the rows and columns of the original EEM matrices. Each experimental result reported was performed 10 times using 80% of the data for training and 20% for testing.

Table 7.9: Cancer Dataset: Performance Values

		Cancer	**Other**	**Adenoma**	**None**
LDA + SVM	Precision	0.37 ± 0.06	0.31 ± 0.06	0.28 ± 0.06	0.31 ± 0.06
	Recall	0.59 ± 0.12	0.75 ± 0.15	0.65 ± 0.23	0.75 ± 0.15
	F_1	0.45 ± 0.06	$\mathbf{0.43} \pm 0.07$	$\mathbf{0.38} \pm 0.09$	$\mathbf{0.43} \pm 0.08$
PARAFAC + SVM	Precision	0.40 ± 0.09	0.25 ± 0.08	0.27 ± 0.04	0.31 ± 0.05
	Recall	0.64 ± 0.18	0.44 ± 0.17	0.54 ± 0.10	0.64 ± 0.16
	F_1	$\mathbf{0.49} \pm 0.12$	0.32 ± 0.11	0.36 ± 0.06	0.42 ± 0.08
LDA + PARAFAC + SVM	Precision	0.40 ± 0.08	0.25 ± 0.08	0.27 ± 0.05	0.32 ± 0.05
	Recall	0.64 ± 0.18	0.46 ± 0.15	0.57 ± 0.13	0.67 ± 0.12
	F_1	$\mathbf{0.49} \pm 0.12$	0.32 ± 0.10	0.37 ± 0.07	$\mathbf{0.43} \pm 0.06$
Raw + SVM	Precision	0.44 ± 0.10	0.34 ± 0.09	0.35 ± 0.11	0.27 ± 0.07
	Recall	0.46 ± 0.11	0.35 ± 0.15	0.36 ± 0.12	0.34 ± 0.09
	F_1	0.44 ± 0.06	0.33 ± 0.11	0.35 ± 0.10	0.30 ± 0.07

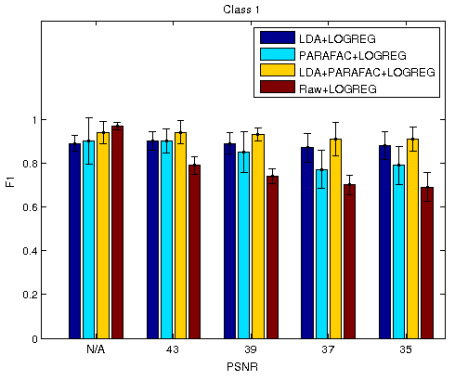
Cancer Dataset: The results for the cancer dataset appear in Table 7.9. For these experiments, we used only the data from the undiluted samples and, unlike the pairwise

Table 7.10: Fluorophore Dataset: Micro Precision, Recall, and F_1 Values Under Different Noise Levels

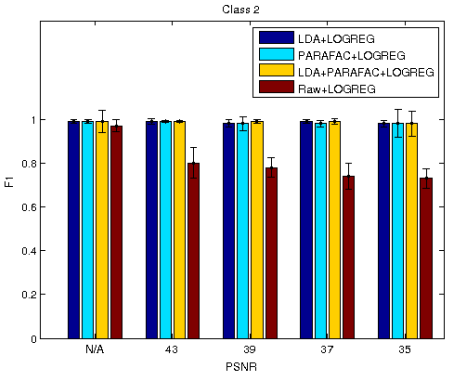
	PSNR		LDA	PARAFAC	PARAFAC + LDA	Raw
Micro	N/A	Precision	0.83 ± 0.18	0.87 ± 0.13	0.91 ± 0.10	$\mathbf{0.98} \pm 0.02$
		Recall	0.88 ± 0.12	0.91 ± 0.08	$\mathbf{0.94} \pm 0.04$	0.93 ± 0.05
		F_1	0.85	0.89	0.92	**0.95**
	43	Precision	0.82 ± 0.18	0.87 ± 0.14	$\mathbf{0.93} \pm 0.08$	0.81 ± 0.03
		Recall	0.87 ± 0.14	0.89 ± 0.09	$\mathbf{0.93} \pm 0.06$	0.71 ± 0.07
		F_1	0.84	0.88	**0.93**	0.76
	39	Precision	0.80 ± 0.19	0.81 ± 0.20	$\mathbf{0.90} \pm 0.10$	0.76 ± 0.07
		Recall	0.85 ± 0.10	0.87 ± 0.10	$\mathbf{0.92} \pm 0.08$	0.70 ± 0.05
		F_1	0.82	0.84	**0.91**	0.73
	37	Precision	0.77 ± 0.21	0.78 ± 0.17	$\mathbf{0.90} \pm 0.10$	0.73 ± 0.06
		Recall	0.83 ± 0.13	0.83 ± 0.12	$\mathbf{0.92} \pm 0.07$	0.68 ± 0.05
		F_1	0.80	0.81	**0.91**	0.70
	35	Precision	0.81 ± 0.17	0.87 ± 0.12	$\mathbf{0.92} \pm 0.07$	0.72 ± 0.03
		Recall	0.82 ± 0.12	0.87 ± 0.12	$\mathbf{0.91} \pm 0.09$	0.67 ± 0.03
		F_1	0.81	0.87	**0.91**	0.69

results presented in [11], our results use the one-versus-all method for classification. Our results show that both, PARAFAC + SVM and the combination PARAFAC + LDA + SVM are the most effective methods to distinguish cancer from all other control groups. On the other hand, we can also observe that LDA + SVM seems to perform favorably when compared to the other approaches at identifying all other classes. It is worth mentioning that a random classifier would produce a precision value of approximately 0.25 and a recall of approximately 0.5, given that the four classes have approximately equal proportions in the dataset. However, overall performance is relatively low for all approaches when tested on this dataset. This result was not unexpected. Lawaetz *et al.* [11] performed an exploratory analysis of this data and concluded that there was no clear separation of cancer and control samples. In fact, the variability in the data was not due to age, gender, or other control variables, but it was most likely due to individual differences.

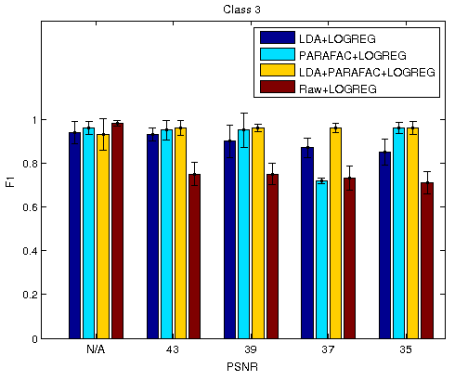
Fluorophores Dataset: Sections 7.2.3 to 7.2.3 show the F_1 values for the fluorophore dataset at different levels of white noise, measured as PSNR. We can see that when the data contains no noise, the raw data vector is enough to obtain a very good prediction. However, a slight increase in the noise level shows how sensitive the raw data classifier is



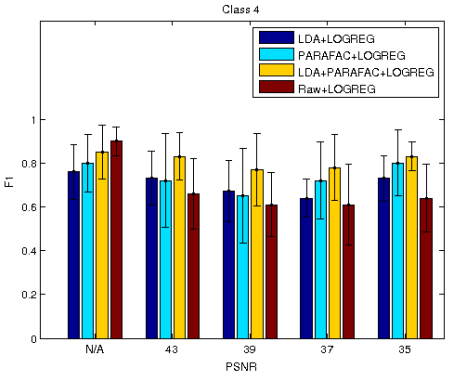
(a) Class 1: Catechol



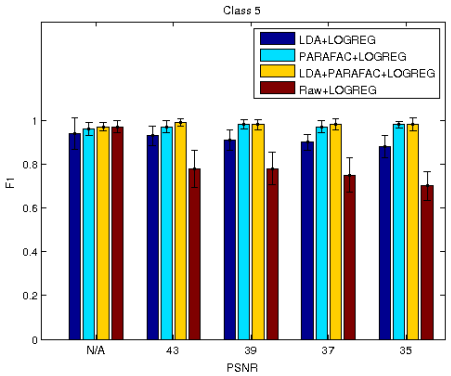
(b) Class 2: Hydroquinone



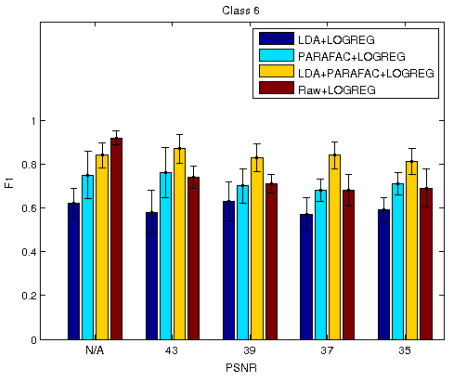
(c) Class 3: Indole



(d) Class 4: Resorcinol



(e) Class 5: Tryptophane



(f) Class 6: Tyrosine

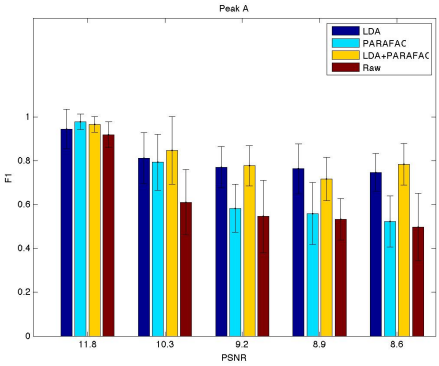
Figure 7.6: F_1 vs. PSNR for fluorophore dataset

to a decrease in the PSNR value. The graphs show how both LDA + Logistic Regression and PARAFAC + Logistic Regression offer robustness to white noise, given that their performance does not decrease as dramatically as the performance of the classifier using raw data. It is also apparent from the results that the combination of both LDA + PARAFAC +Logistic Regression offers additional robustness and the best performance of the tested approaches.

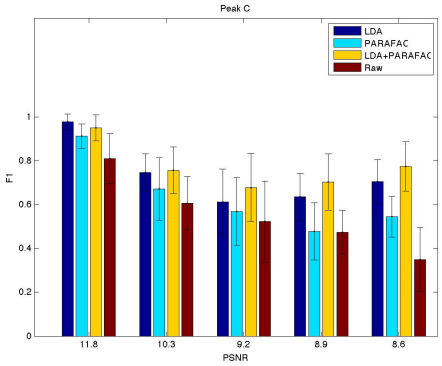
The Fluorophore dataset contains 6 different classses that are not present in the same proportions. Thus, in order to average the classification performance over all classes, we made use of the micro averaging technique in Table 7.3. The overall performance of the approaches in this dataset appear in Table 7.10. We can, once again, clearly see how sensitive the raw data classifier is to a decrease in PSNR values while the PARAFAC + LDA combination offers a consistent and reliable result.

Synthetic Dataset: Sections 7.2.3 to 7.2.3 show the results for the synthetic dataset under different levels of white noise. The same trend of sensitivity to noise seen in the fluorophore dataset results is mirrored here. The most robust method continues to be LDA + PARAFAC and the method that is the most sensitive to noise continues to be the one using raw data.

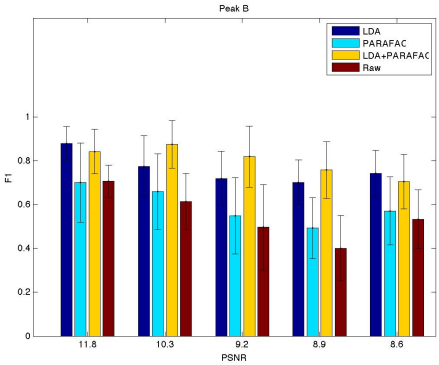
Tables 7.11 and 7.12 show the results for the other two types of noise using macro averaging. These results indicate a higher robustness from LDA when shift noise and height-girth variability are introduced. PARAFAC relies on a trilinear model to describe the data. In such a model, variability comes from the scores as fixed loadings are shared by all samples. Shift noise violates trilinearity assumptions and inhibits the performance of the PARAFAC model. On the other hand, the robustness of the LDA model can be attributed to the fact that it makes no such assumption. The underlying distribution of each latent aspect is fixed, but there is an inherent flexibility in probabilistic models that allows for sample



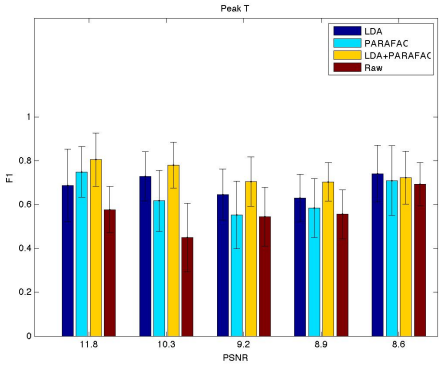
(a) Peak A



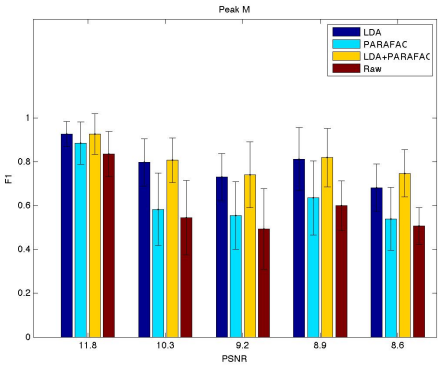
(b) Peak C



(c) Peak B



(d) Peak T



(e) Peak M

Figure 7.7: F1 vs. PSNR for synthetic dataset.

Table 7.11: Synthetic Dataset: Macro Precision, Recall, and F_1 Values Under Noise Created by Shifting Peak Locations

	n_{SN}		LDA	PARAFAC	PARAFAC + LDA	Raw
Macro	10	Precision	**0.90** $\pm$ 0.07	**0.90** $\pm$ 0.11	**0.90** $\pm$ 0.07	0.81 $\pm$ 0.12
		Recall	**0.90** $\pm$ 0.09	0.87 $\pm$ 0.12	**0.90** $\pm$ 0.08	0.81 $\pm$ 0.11
		F_1	**0.90**	0.88	**0.90**	0.81
	20	Precision	0.88 $\pm$ 0.13	0.88 $\pm$ 0.12	**0.90** $\pm$ 0.12	0.77 $\pm$ 0.17
		Recall	0.84 $\pm$ 0.12	0.85 $\pm$ 0.11	**0.88** $\pm$ 0.09	0.71 $\pm$ 0.16
		F_1	0.86	0.86	**0.89**	0.74
	30	Precision	0.89 $\pm$ 0.13	0.85 $\pm$ 0.13	**0.91** $\pm$ 0.08	0.78 $\pm$ 0.16
		Recall	0.83 $\pm$ 0.13	0.79 $\pm$ 0.13	**0.85** $\pm$ 0.12	0.73 $\pm$ 0.17
		F_1	0.86	0.82	**0.87**	0.76
	40	Precision	**0.89** $\pm$ 0.09	0.84 $\pm$ 0.11	0.88 $\pm$ 0.07	0.75 $\pm$ 0.15
		Recall	0.82 $\pm$ 0.11	0.85 $\pm$ 0.12	**0.90** $\pm$ 0.05	0.72 $\pm$ 0.18
		F_1	0.85	0.84	**0.89**	0.73
	50	Precision	0.84 $\pm$ 0.07	0.77 $\pm$ 0.12	**0.86** $\pm$ 0.05	0.72 $\pm$ 0.11
		Recall	0.81 $\pm$ 0.09	0.73 $\pm$ 0.14	**0.82** $\pm$ 0.07	0.67 $\pm$ 0.14
		F_1	0.83	0.75	**0.84**	0.69

variations without a trilinear structure. We can also see that there is an advantage and a measurable enhancement offered by the LDA + PARAFAC combination in both tables.

7.3 Classification Using PARAFAC and LFDA

Microtox® is a testing system used to examine water, soil and air. The system uses non-pathogenic bioluminescent bacteria (Vibrio fischeri) to detect toxic substances invitro. The bacteria emit light as part of their regular metabolism. When exposed to toxic substances that disrupt their respiration the bacteria undergo a drop in luminescence that can be measured and correlated to the level of toxicity found in a substrate. This section will focus on predicting Microtox® measurements using LFDA, PARAFAC, non-parametric regression and DS theory.

7.4 Regression Using PARAFAC and LFDA

A residential urban ambient net-zero water treatment system was presented in [81]. The treatment system proposes a design using a membrane bioreactor, iron-mediated aeration (IMA, reported previously), vacuum ultrafiltration, and peroxone advanced oxidation, with minor rainwater make-up and H_2O_2 disinfection residual. This system neither takes nor

Table 7.12: Synthetic Dataset: Macro Precision, Recall, and F_1 Values Under Type 3 Noise

	n_{HN}		LDA	PARAFAC	PARAFAC + LDA	Raw
Macro	10	Precision	**0.89** $\pm$ 0.04	0.84 $\pm$ 0.10	0.97 $\pm$ 0.06	0.75 $\pm$ 0.11
		Recall	**0.84** $\pm$ 0.14	0.77 $\pm$ 0.11	**0.84** $\pm$ 0.05	0.69 $\pm$ 0.11
		F_1	0.88	0.84	**0.90**	0.76
	20	Precision	**0.89** $\pm$ 0.07	0.84 $\pm$ 0.08	0.87 $\pm$ 0.08	0.75 $\pm$ 0.11
		Recall	**0.84** $\pm$ 0.09	0.77 $\pm$ 0.12	**0.84** $\pm$ 0.07	0.69 $\pm$ 0.08
		F_1	**0.87**	0.80	0.85	0.72
	310	Precision	**0.83** $\pm$ 0.11	0.74 $\pm$ 0.09	0.82 $\pm$ 0.09	0.68 $\pm$ 0.11
		Recall	**0.82** $\pm$ 0.09	0.74 $\pm$ 0.08	**0.82** $\pm$ 0.08	0.64 $\pm$ 0.10
		F_1	**0.82**	0.74	**0.82**	0.66
	40	Precision	**0.88** $\pm$ 0.08	0.80 $\pm$ 0.15	**0.88** $\pm$ 0.06	0.75 $\pm$ 0.09
		Recall	0.83 $\pm$ 0.04	0.68 $\pm$ 0.12	**0.86** $\pm$ 0.03	0.67 $\pm$ 0.12
		F_1	0.86	0.74	0.87	0.71
	50	Precision	0.79 $\pm$ 0.11	0.77 $\pm$ 0.16	**0.85** $\pm$ 0.09	0.70 $\pm$ 0.07
		Recall	0.71 $\pm$ 0.11	0.70 $\pm$ 0.13	**0.78** $\pm$ 0.07	0.66 $\pm$ 0.14
		F_1	0.75	0.74	**0.81**	0.68

releases water off-site, but rather purifies the wastewater by filtration, aeration, oxidation and other methods. The quality of the treatment process is monitored at different stages of the treatment system by taking measurements of pH, turbidity, conductivity, dissolved oxygen, temperature, EEM and Microtox® to name a few. In this work we use the data provided from this treatment facility, the interested reader can find further information about the water treatment mechanism in [81].

The residential urban ambient net-zero water treatment process [81] uses an analysis of Microtox® measurements on water samples collected at different stages of the purification process. The treatment also uses a collection of fluorescence spectroscopic data along with other water quality measurements. It is worth mentioning that Microtox® testing protocols are labor intensive and require specially trained technicians to conduct the test, which can be very sensitive to lab conditions. This offers an opportunity to apply machine learning techniques to assess or predict Microtox® values and water quality while reducing technitian training and testing time and complexity. A novel probabilistic graphical model technique, anmely, Latent Fluorescent Dirichlet allocation (LFDA) has been utilized for modeling and analysis of fluorescent spectroscopic excitation-emission matrices (EEMs) along with parallel factor analysis (PARAFAC).

7.4.1 Microtox® Assessment Process

In this section we focus on Microtox® measurement prediction. The objective is to build a model that can ultimately be used to analyse water quality. In order to start our analysis we dichotomize samples into a positive and a negative set. We refer to the positive instances as the water samples for which the Microtox reading has dropped more than the given threshold in a fifteen minute time period. All other samples are considered negative instances. The following is our hypothesis:

H_0 : The sample is a negative instance,(i.e. Microtox luminescence reading does not drop below the given threshold in fifteen minutes)

H_1 : The sample is a positive instance.

As explained in Section 1.2 we use fluorescence spectroscopic excitation-emission matrices (EEMs) based scores along with pH, turbidity, conductivity and dissolved oxygen. EEM scores are extracted using two main techniques: Parallel factor analysis (PARAFAC), and the novel Latent Fluorescent Dirichlet Allocation (LFDA) based probabilistic graphical approach. These techniques are explained in Sections 1.3 and 3.4 respectively. The extracted scores from PARAFAC and LFDA along with other measurements are given as inputs to a non-parametric (NP) regression system, details are given in section 5. The output values of the regression obtained by using PARAFAC scores and LFDA scores are used to create two Dempster-Shafer (DS) theoretic Bodies of Evidence (explained in section 6) which are then fused to give an evidence based classification result. Given sufficient evidence in favour of sample being positive instance leads to a rejection of null-hypothesis H_0.

If we reject H_0, i.e. when the water sample can not be regarded as a negative sample by the above features, further tests are done in order to look for possible cause via extensive laboratory experiments.

Testing was done on a dataset consisting of data points collected over the period April 2013 to December 2014. The data was filtered to avoid missing records and measurements and a subset of 489 samples were used in the analysis with LDA/LFDA and PARAFAC. Non-parametric(NP), linear models(LM) and support vector machine(SVM) based regression methods were applied on the dataset with three sets of variables to predict the drop in Microtox measurement y. The first set of variables $\mathbf{x}_{LDA}$, consists of variables,

$$\mathbf{x}_{LDA} = \left\{ \begin{array}{l} x_{L1} - \text{First LDA score}, \\ x_{L2} - \text{Second LDA score}, \\ x_{pH} - \text{pH value}, \\ x_{Con} - \text{Conductivity value}, \\ x_{DO} - \text{Dissolved Oxygen value}, \\ x_{Tur} - \text{Turbidity value} \end{array} \right\},$$

where x_{L1} and x_{L2} have been obtained from the LDA-based probabilistic graphical model. Figure 7.8 shows the receiver operating characteristic (ROC) curves for NP, LM and SVM based regressions for y labels obtained from assigning class '1' for Microtox reading dropped more than 50% in fifteen minutes, and assigning '0' otherwise. The area under the curve for NP regression was around 75% while for SVM based regression it was around 65%. Under the optimum threshold the sensitivity of 78% and specificity of 60% was reported for NP regression.

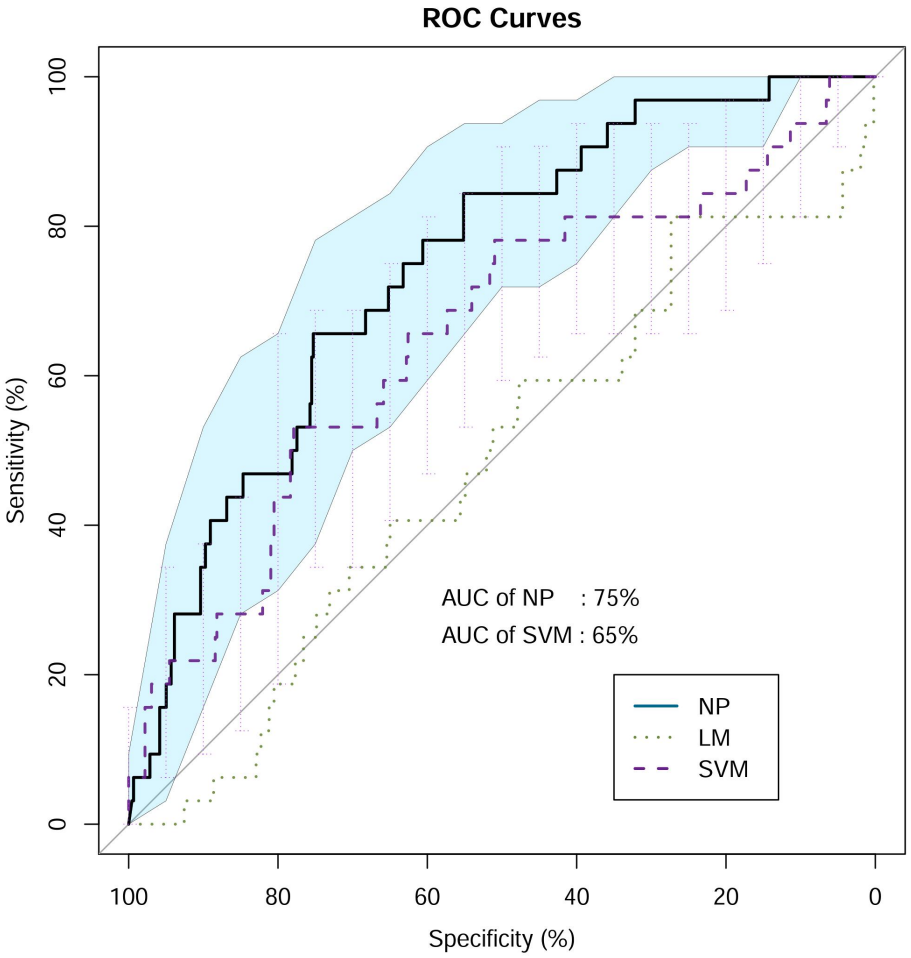


Figure 7.8: The receiver operating characteristic (ROC) curves for non-parametric(NP), linear (LM) and support vector machine(SVM) based regressions are given with variables $\mathbf{x}_{LDA}$. A threshold of 50% reduction in luminescence reading in Microtox has been used to label the positive samples y. The area under the curve (AUC) is given only for NP and SVM based regressions.

The second set of variables $\mathbf{x}_{PA}$, consists of variables,

$$\mathbf{x}_{PA} = \left\{ \begin{array}{l} x_{P1} - \text{First PARAFAC score}, \\ x_{P2} - \text{Second PARAFAC score}, \\ x_{pH} - \text{pH value}, \\ x_{Con} - \text{Conductivity value}, \\ x_{DO} - \text{Dissolved Oxygen value}, \\ x_{Tur} - \text{Turbidity value} \end{array} \right\},$$

where x_{P1} and x_{P2} have been obtained using PARAFAC.

Figure 7.9 shows the ROC curves for NP, LM and SVM based regressions having the same setup that of in figure 7.8, except with variables $\mathbf{x}_{PA}$. The area under the curve for NP regression was around 75% whereas for SVM based regression it was around 66%. For NP regression, under the optimum threshold, the sensitivity and specificity was recorded as 63% and 78%. However, it should be noted that this relatively higher specificity has been given by compromising the sensitivity. For instance, when the threshold of classification was slightly increased sensitivity and specificity values of 73% and 59% can be obtained.

The third set of variables $\mathbf{x}_{LDA}$, consists of variables,

$$\mathbf{x}_{LFDA} = \left\{ \begin{array}{l} x_{LF1} - \text{First LFDA score}, \\ x_{LF2} - \text{Second LFDA score} \end{array} \right\},$$

where x_{LF1}, x_{LF2} and x_{LF3} have been obtained from the LFDA-based probabilistic graphical model. Figure 7.10 shows the receiver operating characteristic (ROC) curves for NP, LM and SVM based regressions for y labels obtained from assigning class '1' for Microtox reading dropped more than 50% in fifteen minutes, and assigning '0' otherwise. The area under the curve for NP regression was around 78% while for SVM based regression it was around 67%. Under the optimum threshold the sensitivity of 83% and specificity of 60% was reported for NP regression.

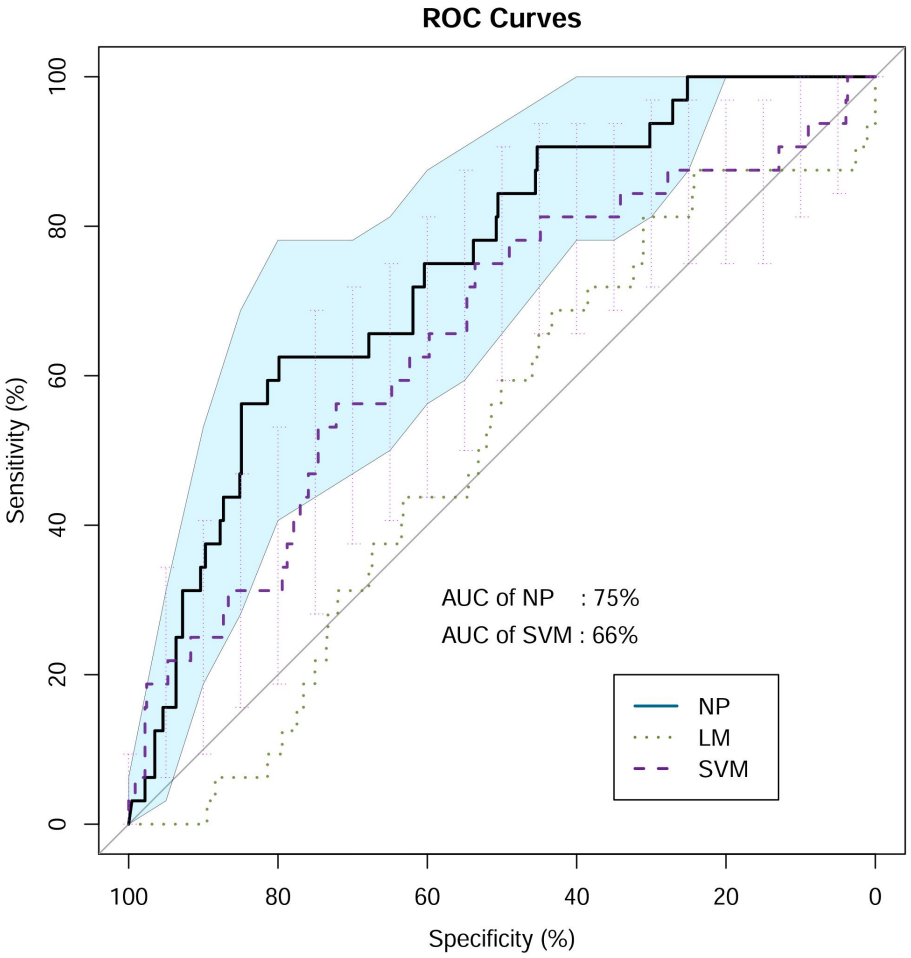


Figure 7.9: The receiver operating characteristic (ROC) curves for non-parametric(NP), linear (LM) and support vector machine(SVM) based regressions are given with variables $\mathbf{x}_{PA}$. A threshold of 50% reduction in luminescence reading in Microtox has been used to label the positive samples y. The area under the curve (AUC) is given only for NP and SVM based regressions.

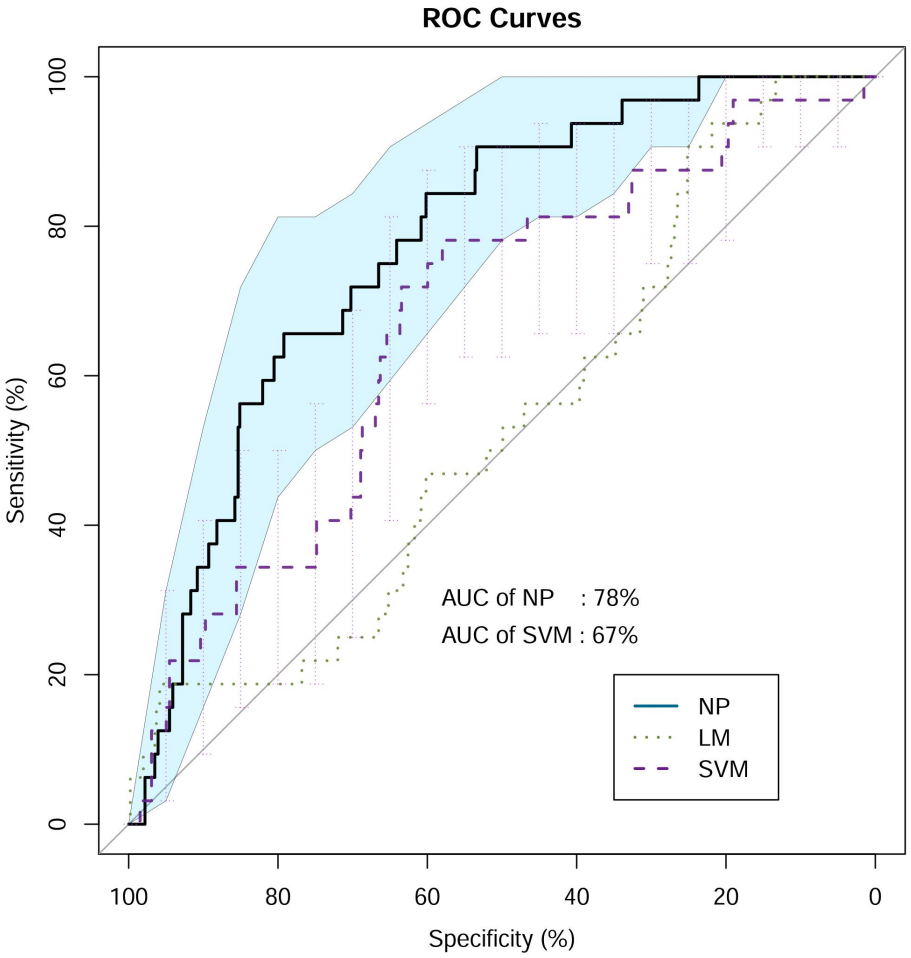


Figure 7.10: The receiver operating characteristic (ROC) curves for non-parametric(NP), linear (LM) and support vector machine(SVM) based regressions are given with variables $\mathbf{x}_{LFDA}$. A threshold of 50% reduction in luminescence reading in Microtox has been used to label the positive samples y. The area under the curve (AUC) is given only for NP and SVM based regressions.

It was seen that non-parametric regression is a better suited approach for our data set. One reason for this improvement maybe related to the fact that non-parametric regression as a local estimator can handle class imbalance in data sets better than global estimators. In contrast, linear regression is a global estimator and support vector machines can be regarded as a 'fusion' of local and global estimators. From Figures 7.8, 7.9 and 7.10 it can be seen the performance gets better when the estimator becomes 'more local'

7.4.2 DS Performance Measures

Since we are getting evidences from two sources, we have used the DS fusion method elaborated in section 6. To report the performance we use DS theoretic specificity and sensitivity based on the definition given for DS theoretic precision and recall in [84]. For that, let us first define the DS framework to use the DS-Model described in section 6.2. The positive samples (with labels '1') are denoted with θ_1 and the negative samples are denoted with θ_2, hence the FoD $\Theta = \{\theta_1, \theta_2\}$. For each testing sample s after fusing the corresponding results of $\widehat{m}_{PA}$ and $\widehat{m}_{LDA}$, the fused BoE can be obtained, say $\{\Theta, \mathbf{F}, m^s\}$. Using pignistic transformations [85] c Under this setting the 'true positives'(TP), 'false positives'(FP), 'false negatives'(FN) and 'true negatives'(TN) can be given as:

$$TP_{DS} \ \sum_{s\in\mathscr{S}_{(+)}} \widehat{BetP}(\theta_1)$$

$$FP_{DS} \ \sum_{s\in\mathscr{S}_{(-)}} \widehat{BetP}(\theta_1)$$

$$FN_{DS} \ \sum_{s\in\mathscr{S}_{(+)}} \widehat{BetP}(\theta_2)$$

$$TN_{DS} \ \sum_{s\in\mathscr{S}_{(-)}} \widehat{BetP}(\theta_2)$$

Now the *DS-Sensitivity*, *DS-Specificity* and the *F-measure* are given in (7.5), (7.6), and (7.7) respectively, following the conventional definitions.

$$\text{DS-Sensitivity} = \frac{TP_{DS}}{TP_{DS} + FN_{DS}}. \tag{7.5}$$

$$\text{DS-Specificity} = \frac{TN_{DS}}{FP_{DS} + TN_{DS}}. \tag{7.6}$$

$$F_{\beta(DS)} = \frac{(1+\beta^2)\cdot TP_{DS}}{(1+\beta^2)\cdot TP_{DS} + \beta^2 \cdot FN_{DS} + FP_{DS}} \tag{7.7}$$

In this application of assessing Microtox® measurements it is desirable to have a higher sensitivity to detect all the sudden changes in Microtox readings. As mentioned in Section 7.4.1, Microtox® readings mimic the impurity of the corresponding water sample. In our analysis we ran five fold cross validation and calculated non-parametric regression values corresponding to $\mathbf{x}_{LDA}$, $\mathbf{x}_{PA}$ and $\mathbf{x}_{LFDA}$ for each training sample. Then for each training sample Dempster-Shafer mass structures $m_{LDA}(\cdot)$, $m_{PA}(\cdot)$ and $m_{LFDA}(\cdot)$ were assigned. The fused mass structures $(m_{LDA} \oplus m_{PA})(\cdot)$ and $(m_{LFDA} \oplus m_{PA})(\cdot)$ were obtained as explained in Section 6.2. Then TP_{DS}, FP_{DS}, FN_{DS} and TN_{DS} we calculated separately under mass structures $m_{LDA}(\cdot)$, $m_{PA}(\cdot)$, $m_{LFDA}(\cdot)$, $(m_{LDA} \oplus m_{PA})(\cdot)$ and $(m_{LFDA} \oplus m_{PA})(\cdot)$.

From DS theory, a large value of $m(\Theta)$ for a particular water sample, essentially means there is a high uncertainty on the prediction for the corresponding sample. Hence when calculating TP_{DS}, FP_{DS}, FN_{DS} and TN_{DS} we only consider samples with low uncertainty. In this analysis 37 samples with uncertainty greater than 90%, in any of the independent evidence sources (i.e., m_{LDA}, $m_{PA}(\cdot)$ or $m_{LFDA}(\cdot)$) were reserved for further analysis. In the remaining 452 samples there were 32 positive samples and 420 negative samples. The DS-sensitivity, DS-specificity and $F_{1(DS)}$ measure obtained for five fold cross validation for the 452 samples are presented in Table 7.13.

Table 7.13: Water Dataset: Performance Values

	DS-Sensitivity	**DS-Specificity**	$\mathbf{F_{1(DS)}}$
$\mathbf{m_{LDA}}$	0.7860	0.5978	0.2218
$\mathbf{m_{PA}}$	0.7831	0.6065	0.2273
$\mathbf{m_{LFDA}}$	0.8201	0.6276	0.2451
$\mathbf{m_{LDA} \oplus m_{PA}}$	0.8486	0.6446	0.2637
$\mathbf{m_{LFDA} \oplus m_{PA}}$	0.8976	0.6858	0.3002

Statistical comparison of $F_{1(DS)}$ values for five fold cross validation on 452 samples have been carried using the post-hoc *Nemenyi test* on significant results of the *Friedman test* (for interesting readers we refer [86]). The Friedman test is a non-parametric equivalent of ANOVA whereas Nemenyi test is similar to Tukey test for ANOVA. The main reason for us to use a non-parametric test is not to assume the distribution, for instance normal distribution, of the compared data. It can be seen that LFDA scores fused with PARAFAC accompanied with pH, Conductivity, DO and Turbidity yields the best results.

Table 7.14 gives the $p-values$ of the pairwise comparisons using Nemenyi post-hoc test with $F_{1(DS)}$ values obtained from five fold cross validation with m_{LDA}, m_{PA}, m_{LFDA}, $m_{LDA} \oplus m_{PA}$ and $m_{LFDA} \oplus m_{PA}$.

Table 7.14: Friedman Nemenyi Post-hoc Test

p-values with $F_{1(DS)}$	$\mathbf{m_{LFDA} \oplus m_{PA}}$	$\mathbf{m_{LDA} \oplus m_{PA}}$	$\mathbf{m_{LFDA}}$	$\mathbf{m_{PA}}$
$\mathbf{m_{LDA} \oplus m_{PA}}$	0.8555	-	-	-
$\mathbf{m_{LFDA}}$	0.2659	0.8555	-	-
$\mathbf{m_{PA}}$	0.0120	0.1796	0.7514	-
$\mathbf{m_{LDA}}$	0.0014	0.0409	0.3735	0.9751

For our discussion let us use the loose notation X^* to denote X accompanies pH, Conductivity, DO and Turbidity in regression step. For instance LDA* denotes that, LDA scores are used along with pH, Conductivity, DO and Turbidity measurements in regression step. In contrast LFDA (without *) denotes, only LFDA scores are used in regression step.

According to the Nemenyi test fused LFDA and PARAFAC* differs significantly ($p < 0.01$) compared to using only LDA*. The former also differs ($p < 0.05$) compared to using only PARAFAC*. The fused LDA* and PARAFAC* differs ($p < 0.05$) from using LDA* only.

In practice each sample's final fused mass structure should be checked. Samples with higher $m(\theta_1)$, i.e., positive with higher evidence, should be considered unsafe as they correlate to severe drops in Microtox® measurements within a 15 minute period. If $m(\theta_2)$ is higher for a particular water sample, then it should not be regarded as unsafe as they

correlate to constant Microtox® measurements within a 15 minute period. However, when $m(\Theta)$ is higher, due to the uncertainty, further testing is necesary to check the quality of water.

7.4.3 LFDA Drawbacks

Like many machine learning algorithms LFDA has certain drawbacks that were made aparent during the testing phase. Given the number of *observed* variables incorporated into the model the need for data is increased. The multidimensional array Φ has a number of dimesions proportional to the number of control variables in the model. This is applicable not only to the number of *observed* variables but also to the cardinality of the discretization of each one of the variables.

The model needs to learn the behavior of the fluorescent latent aspects given several dependencies, the larger the number of dependencies the larger the needed dataset. In our specific use case data was readily available given that water samples are tested at the treatment facility on a weekly basis. Other problem domains might not have such a readily available large dataset and might be ill suited for analysis using LFDA. Several runs of experiments were conducted to measure the sensitivity of LFDA compared to the other approaches presented in this work. Table 7.15 shows the results of this test.

Table 7.15: AUC for NP regression

Percent of full dataset	20%	40%	60%	80%	100%
PARAFAC	70%	73%	74%	75%	75%
LDA	63%	70%	72%	75%	75%
LFDA	59%	68%	70%	74%	78%

These experiments show the over-sensitivity of LFDA to a reduction of the dataset size. PARAFAC and LDA show some reduction of performance but LFDA is specially afected by the curse of dimensionality due to the inter-dependency of the control variables in the model.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

We have presented a novel LDA-based approach to analyze two-way data. Our experiments demonstrate an improved accuracy and robustness to different types of noise on EEM matrices. Our proposed probabilistic graphical model based approach seems to offer advantages specifically for shift and height-girth variability of the peaks corresponding to fluorescent components. Improvements on precision and recall values were observed when classifiers using both PARAFAC and LDA are combined.

We have also presented a technique on assessing Microtox® measurements with EEMs. We have introduced the novel LFDA model which is based on our previous LDA based graphical model in [87]. In our analysis we have used LDA and LFDA along with the three way analysis technique PARAFAC to generate fetures for each EEM. Non-parametric regression techniques were applied to get regression values which were used in a DS framework to generate DS masses for each testing sample. These DS masses were then fed to a DS fusion method to obtain DS theoretic predictions on water quality as indicated by Microtox® measurements. The use of LDA along with PARAFAC for EEM analysis was already justified in [87]. In this work it was evident that the novel LFDA along with

PARAFAC accompanied with pH, Conductivity, DO and Turbidity gave the best performance in predicting Microtox® measurements.

8.2 Future Work

The LFDA model offers insights into changes induced by external variables. In recent literature other extended graphical models have included prediction models such as supervides latent Dirichlet allocation (sLDA) presented in [88]. In these extended models a response variable is included to predict a class or rating for each sample. These extended models not only perform latent aspect modeling but also make predictions by incorporating a class variable. The results of these models illustrate certain benefits of sLDA versus modern regularized regression, as well as versus an unsupervised LDA analysis followed by a separate regression. Extending the LFDA model to include a prediction variable might be able to improve prediction further.

Furthermore, the evidential belief methods used in this work apply Dempster's rule of combination (DRC), which is the most basic fusion method used in DS theory. Other methods such as the conditional update equation (CUE) might offer better prediction results and might offer an interesting area of reserach for future work.

Bibliography

[1] B. Hua, F. Dolan, C. Mcghee, T. E. Clevenger, and B. Deng, "Water-source characterization and classification with fluorescence EEM spectroscopy: PARAFAC analysis," *International Journal of Environmental and Analytical Chemistry*, vol. 87, no. 2, pp. 135–147, 2007.

[2] G. Hall and J. Kenny, "Estuarine water classification using EEM spectroscopy and PARAFAC-SIMCA." *Anal Chim Acta*, vol. 581, no. 1, pp. 118–24, 2007.

[3] J. SádeCká and J. TóThoVá, "Fluorescence spectroscopy and chemometrics in the food classificationa review," *Czech Journal of Food Sciences*, vol. 25, no. 4, pp. 159–173, 2007.

[4] J. Christensen, V. Povlsen, and J. Sørensen, "Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage," *Journal of Dairy Science*, vol. 86, no. 4, pp. 1101–1107, 2003.

[5] C. A. Stedmon, S. Markager, and R. Bro, "Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy," *Marine Chemistry*, vol. 82, no. 3 – 4, pp. 239 – 254, 2003.

[6] N. Hudson, A. Baker, and D. Reynolds, "Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters a review," *River Research and Applications*, vol. 23, no. 6, pp. 631–649, 2007.

[7] T. Ohno and R. Bro, "Dissolved organic matter characterization using multiway spectral decomposition of fluorescence landscapes," *Soil Science Society of America Journal*, vol. 70, no. 6, pp. 2028–2037, 2006.

[8] G. J. Hall, K. E. Clow, and J. E. Kenny, "Estuarial fingerprinting through multidimensional fluorescence and multivariate analysis," *Environmental Science and Technology*, vol. 39, no. 19, pp. 7560–7567, 2005.

[9] B. Hua, K. Veum, A. Koirala, J. Jones, T. Clevenger, and B. Deng, "Fluorescence fingerprints to monitor total trihalomethanes and N-nitrosodimethylamine formation potentials in water," *Environmental Chemistry Letters*, vol. 5, pp. 73–77, 2007.

[10] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*. Springer Verlag, 2009.

[11] A. Lawaetz, R. Bro, M. Kamstrup-Nielsen, I. Christensen, L. Jorgensen, and H. Nielsen, "Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer," *Metabolomics*, vol. 8, pp. 111–121, 2012.

[12] S. Madhuri, P. Aruna, M. I. Summiya Bibi, V. S. Gowri, D. Koteeswaran, and S. Ganesan, "Ultraviolet fluorescence spectroscopy of blood plasma in the discrimination of cancer from normal," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 2982, May 1997, pp. 41–45.

[13] A. Uppal, N. Ghosh, A. Datta, and P. Gupta, "Fluorimetric estimation of the concentration of NADH from human blood samples," *Biotechnology and Applied Biochemistry*, vol. 41, no. 1, pp. 43–47, 2005.

[14] M. Leiner, R. Schaur, G. Desoye, and O. Wolfbeis, "Fluorescence topography in biology. III: Characteristic deviations of tryptophan fluorescence in sera of patients with gynecological tumors." *Clin Chem*, vol. 32, no. 10, pp. 1974–8, 1986.

[15] S. Madhuri, S. Suchitra, P. Aruna, T. Srinivasan, and S. Ganesan, "Native fluorescence characteristics of blood plasma of normal and liver diseased subjects." *Midical Science Research*, vol. 27, no. 2, pp. 635–639, 1999.

[16] S. Madhuri, N. Vengadesan, P. Aruna, D. Koteeswaran, P. Venkatesan, and S. Ganesan, "Native fluorescence spectroscopy of blood plasma in the characterization of oral malignancy," *Photochemistry and Photobiology*, vol. 78, no. 2, pp. 197–204, 2003.

[17] V. Masilamani, K. Al-Zhrani, M. Al-Salhi, A. Al-Diab, and M. Al-Ageily, "Cancer diagnosis by autofluorescence of blood components," *Journal of Luminescence*, vol. 109, pp. 143 – 154, 2004.

[18] R. Bro and H. A. L. Kiers, "A new efficient method for determining the number of components in PARAFAC models," *J. Chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[19] C. M. Andersen and R. Bro, "Practical aspects of PARAFAC modeling of fluorescence excitation-emission data," *Journal of Chemometrics*, vol. 17, no. 4, pp. 200–215, 2003.

[20] R. Bro, "PARAFAC. tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, 1997.

[21] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, 2005.

[22] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 1, pp. 6 –20, jan. 2009.

[23] L. R. Tucker, "The extension of factor analysis to three-dimensional matrices," in *Contributions to Mathematical Psychology.* New York: Holt, Rinehart and Winston, 1964, pp. 110–127.

[24] R. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, 1970.

[25] J. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, September 1970.

[26] R. Cattell, "Parallel proportional profiles and other principles for determining the choice of factors by rotation," *Psychometrika*, vol. 9, pp. 267–283, 1944.

[27] R. Sands and F. W. Young, "Component models for three-way data: An alternating least squares algorithm with optimal scaling features," *Psychometrika*, vol. 45, no. 1, pp. 39–67, 1980.

[28] E. Sanchez and B. R. Kowalski, "Tensorial resolution: a direct trilinear decomposition," *Journal of Chemometrics*, vol. 4, no. 1, pp. 29–45, 1990.

[29] S. Li and P. J. Gemperline, "Eliminating complex eigenvectors and eigenvalues in multiway analyses using the direct trilinear decomposition method," *Journal of Chemometrics*, vol. 7, no. 2, pp. 77–88, 1993.

[30] R. A. Harshman and M. E. Lundy, "The PARAFAC model for three-way factor analysis and multidimensional scaling," *Research Methods for Multimode Data Analysis*, pp. 122–215, 1984.

[31] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[33] L. D. Brown, "Fundamentals of statistical exponential families with applications in statistical decision theory," *Lecture Notes-monograph Series*, pp. i–279, 1986.

[34] N. Metropolis and S. Ulam, "The Monte Carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.

[35] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, p. 1087, 1953.

[36] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.

[37] A. F. Smith and G. O. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 3–23, 1993.

[38] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. CRC Press, 1996, vol. 2.

[39] R. Kannan, "Markov chains and polynomial time algorithms," in *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*. IEEE, 1994, pp. 656–671.

[40] G. O. Roberts and J. S. Rosenthal, "General state space markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.

[41] S. P. Brooks and A. Gelman, "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 434–455, 1998.

[42] M. K. Cowles and B. P. Carlin, "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.

[43] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed., ser. Springer Series in Statistics. Springer Verlag, 1996.

[44] J. Besag, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems," *Statistical Science*, pp. 3–41, 1995.

[45] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[46] T. Griffiths and M. Steyvers, "Finding scientific topics." *Proc Natl Acad Sci U S A*, vol. 101 Suppl 1, 2004.

[47] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, 2008.

[48] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 592–602.

[49] L. Li, T. A. Darden, C. Weingberg, A. Levine, and L. G. Pedersen, "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method," *Combinatorial Chemistry & High Throughput Screening*, vol. 4, no. 8, pp. 727–739, 2001.

[50] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine Learning*, vol. 30, no. 2, pp. 195–215, 1998.

[51] R. Prati, G. Batista, and M. Monard, "Class imbalances versus class overlapping: an analysis of a learning system behavior," *MICAI 2004: Advances in Artificial Intelligence*, pp. 312–321, 2004.

[52] H. He and E. A. Garcia, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[53] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study." *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[54] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[55] T. Fawcett and F. Provost, "Combining data mining and machine learning for effective user profiling," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 8–13.

[56] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[57] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 21, no. 2-3, p. 427, 2008.

[58] D. Mease, A. J. Wyner, and A. Buja, "Boosted classification trees and class probability/quantile estimation," *The Journal of Machine Learning Research*, vol. 8, pp. 409–439, 2007.

[59] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, 2003.

[60] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 2, pp. 539–550, 2009.

[61] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004.

[62] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[63] H. He, Y. Bai, E. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, June, pp. 1322–1328.

[64] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," *Advances in Intelligent Computing*, pp. 878–887, 2005.

[65] N. Chawla, A. Lazarevic, L. Hall, and K. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," *Knowledge Discovery in Databases: PKDD 2003*, pp. 107–119, 2003.

[66] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 179–186.

[67] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[68] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," *Artificial Intelligence in Medicine*, pp. 63–66, 2001.

[69] K. Gowda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (corresp.)," *Information Theory, IEEE Transactions on*, vol. 25, no. 4, pp. 488–490, 1979.

[70] J. Racine and Q. Li, "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, vol. 119, no. 1, pp. 99 – 130, Mar. 2004.

[71] D. J. Henderson and C. F. Parmeter, *Applied Nonparametric Econometrics*. Cambridge University Press, 2015.

[72] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 401–407, 1996.

[73] J. Racine and Q. Li, "Cross-validation local linear nonparametric regression," *Statistica Sinica*, vol. 14, pp. 485–512, 2004.

[74] E. A. Nadaraya, "On non-parametric estimates of density functions and regression curves," *Theory of Probability & Its Applications*, vol. 10, no. 1, pp. 186 – 190, Jul. 1965.

[75] G. S. Watson, "Smooth regression analysis," *Sankhyã: The Indian Journal of Statistics, Series*, vol. 26, no. 4, pp. 359 – 372, Dec. 1964.

[76] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976, vol. 1.

[77] ——, "Belief functions and parametric models," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 322–352, 1982.

[78] D. Vilar, M. J. Castro, and E. Sanchis, "Multi-label text classification using multinomial models," in *Advances in Natural Language Processing*. Springer Verlag, 2004, pp. 220–230.

[79] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*. Springer Verlag, 2004, pp. 22–30.

[80] P. G. Coble, "Characterization of marine and terrestrial dom in seawater using excitation-emission matrix spectroscopy," *Marine Chemistry*, vol. 51, no. 4, pp. 325–346, 1996.

[81] J. D. Englehardt, T. Wu, and G. Tchobanoglous, "Urban net-zero water treatment and mineralization: Experiments, modeling and design," *Water Research*, vol. 47, no. 13, pp. 4680–4691, 2013.

[82] T. M. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying rare and subtle behaviors: A weakly supervised joint topic model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 12, pp. 2451–2464, 2011.

[83] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[84] T. Wickramarathne, K. Premaratne, M. Kubat, and D. Jayaweera, "Cofids: A belief-theoretic approach for automated collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 175–189, Feb. 2011.

[85] P. Smets, "Practical uses of belief functions," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. Morgan Kaufmann Publishers Inc., 1999, pp. 612–621.

[86] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[87] O. Martinez, R. Dabarera, K. Premaratne, and M. Kubat, "LDA-based probabilistic graphical model for excitation-emission matrices," *International Journal of Intelligent Data Analysis*, vol. 19, no. 5, 2015.

[88] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems*, 2008, pp. 121–128.