MULTI-VIEW CANONICAL CORRELATION ANALYSIS

Jan Rupnik

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Dunja Mladenić, Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
Co-Supervisor: Prof. Dr. John Shawe-Taylor, Department of Computer Science, University College London, Gower Street London WC1E 6BT, United Kingdom

Evaluation Board:
Prof. Dr. Nada Lavrač, Chair, Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
Prof. Dr. Bor Plestenjak, Member, Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 21, Ljubljana, Slovenia
Prof. Dr. Nicolò Cesa-Bianchi, Member, Department of Computer Science, University of Milan, via Comelico 39 20135 Milano, Italy

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Jan Rupnik

MULTI-VIEW CANONICAL CORRELATION ANALYSIS

Doctoral Dissertation

KANONIČNA KORELACIJSKA ANALIZA ZA VEČ MNOŽIC SPREMENLJIVK

Doktorska disertacija

Supervisor: Prof. Dr. Dunja Mladenić

Co-Supervisor: Prof. Dr. John Shawe-Taylor

Ljubljana, Slovenia, January 2016

To Zala

Acknowledgments

I would like to express great appreciation to my PhD supervisor Prof. Dunja Mladenić and co-supervisor Prof. John Shawe-Taylor for their guidance and advice throughout my studies. Special thanks go to Primož Škraba for his invaluable advice and contributions to my work. I would also like to thank Marko Grobelnik for providing inspiration and encouragement for my research work.

Assistance provided by my collaborators Andrej Muhič, Blaž Fortuna, Gregor Leban and Sabrina Guettes was greatly appreciated. It was my great pleasure working with you. I would also like to thank my friends from UCL: Tom Diethe, David Roi Hardoon and Zakria Hussain for great company and discussions during my visits to UCL.

My gratitude goes to the members of my doctoral committee, Prof. Nada Lavrač, Prof. Bor Plestenjak and Prof. Nicolò Cesa-Bianchi, for their valuable comments and remarks.

Finally, I wish to thank Zala, my parents Darja and Erik and my sister Nika for their support and encouragement.

The author gratefully acknowledges funding by the projects SMART (IST-5-033917-STP), X-LIKE (ICT-257790-STREP), MultilingualWeb (PSP-2009.5.2 Agr.# 250500), TransLectures (FP7-ICT-2011-7), PlanetData (ICT-257641-NoE), RENDER (ICT-257790-STREP), XLime (FP7-ICT-611346), and META-NET (ICT-249119-NoE).

Abstract

Matrix factorization represents a popular approach in pattern analysis and is used to tackle many problems, such as: collaborative filtering, imputing missing data, denoising data, dimensionality reduction, data visualization and exploratory analysis.

This thesis is focused on factorization based pattern analysis methods for *multiview learning* problems: that is problems where each data instance is represented by multiple *views* of an underlying object, encoded by multiple *feature sets*. As an example of a multiview problem consider a dataset where each instance has two representations: a visual image and a textual description. The patterns of interest are pairs of functions over images and texts that are strongly related over the observed data.

Canonical Correlation Analysis (CCA) is designed to extract patterns from data sets with two views. This thesis focuses on two generalizations of CCA, which were proposed in the literature: *Sum of Correlations* (SUMCOR) and *Sum of Squared Correlations* SSCOR. The SUMCOR problem formulation is interesting from the optimization perspective by its own right, since it emerges in other problems as well.

We study several aspects of the generalizations. We first present a provably convergent novel algorithm for finding non-linear higher order patterns, which is based on an iterative approach for solving multivariate eigenvalue problems. We show that SUMCOR in general is NP-hard and then study its reformulation to a computationally tractable Semidefinite Programming (SDP) problem. Based on the reformulation we derive several computationally feasible bounds on global optimality, which complement the locally optimal solutions. We introduce a new preprocessing step for dealing with large scale SDP problems that arise from an application to cross-lingual text analysis. We investigated how to apply our methods to real datasets with missing data. The particular structure of missing data in the problem considered leads to a simplification of the SSCOR optimization problem, which is reduced to a tractable eigenvalue problem. We show how the algorithms apply to building cross-lingual similarity models and apply the models on the task of cross-lingual cluster linking. The approach to cross-lingual cluster linking is used in a real-time global analysis of news streams in multiple languages.

Povzetek

Metode, ki temeljijo na matrični faktorizaciji, predstavljajo pomemben pristop k analizi vzorcev in podatkovnemu rudarjenju. Naloge, ki jih lahko prevedemo na matrične razcepe, vključujejo izbiranje s sodelovanjem (ang. *collaborative filtering*), vstavljanje manjkajočih podatkov (ang. *missing data imputation*), zmanjševanje dimenzij (ang. *dimensionality reduction*), odstranjevanje šuma (ang. *denoising*), vizualizacijo podatkov (ang. *data visualization*) in raziskovalno analizo podatkov (ang. *exploratory data analysis*).

V disertaciji se ukvarjamo z *večpoglednim učenjem* (ang. *multiview learning*), kjer predpostavljamo, da imamo za podatke dva ali več *pogledov* (ang. *views*), kar konkretneje pomeni, da imamo za vsako podatkovno instanco na voljo dve ali več množic značilk (ang. *feature sets*), ki predstavljajo različne poglede na nek objekt. Primer podatkovne množice, primerne za večpogledno učenje, je množica parov slik in tekstovnih opisov slik. Predpostavljamo, da lahko slike in besedila predstavimo kot objekte v dveh vektorskih prostorih, katerih dimenzije ustrezajo značilkam za analizo slik oziroma besedil. V tem primeru iščemo vzorce (predstavljene kot funkcionale) v prostoru slik in tekstovnem prostoru, ki so paroma močno povezani (na primer visoko korelirani vzdolž učne množice).

Kanonična korelacijska analiza (KKA) predstavlja enega od najpomembnejših pristopov za analizo podatkov, kjer sta na voljo dva pogleda oziroma dve množici spremenljivk. V pričujočem delu preučujemo dve posplošitvi metode KKA za analizo poljubnega števila množic značilk: metodo vsote korelacij (VKOR) (ang. *Sum Of Correlations*) in metodo vsote kvadratov korelacij (VKKOR).

Omenjeni posplošitvi VKOR in VKKOR preučimo z več vidikov. Prvi prispevek k znanosti predstavlja dokazano konvergentni algoritem za iskanje več množic nelinearnih vzorcev, ki temelji na iterativni metodi za reševanje multivariatnih problemov lastnih vrednosti (ang. *multivariate eigenvalue problems*). Dokažemo, da je problem KKRO v splošnem NP-težak, kar nas privede do analize konveksne relaksacije in prevedbe na optimizacijsko nalogo semidefinitnega programiranja (SDP) (ang. *Semidefinite Programming*). Na podlagi SDP formulacije predstavimo številne nove spodnje meje za vrednost globalno optimalne rešitve. Čeprav so meje izračunljive v polinomskem času, je njihov izračun v praksi lahko težaven. Zato predlagamo pristop, ki temelji na zmanjšanju števila spremenljivk s pomočjo naključnih projekcij. Predstavimo tudi aplikacijo posplošitev KKA na problemu učenja jezikovno neodvisne mere podobnosti, kjer naletimo na problem manjkajočih učnih podatkov. Pokažemo, da določena struktura manjkajočih podatkov pripelje do poenostavitve optimizacijskega problema KKRO, ki ga prevedemo na računsko manj zahteven problem lastnih vrednosti. Pokažemo, kako lahko uporabimo jezikovno neodvisno mero podobnosti za medjezično povezovanje gruč (ang. *clusters*) dokumentov. Pristop uporabimo v sistemu za globalno analizo tokov novic v več jezikih.

Contents

List of Figures xv

List of Tables xvii

List of Algorithms xix

Abbreviations xxi

Symbols xxiii

Glossary xxv

1 Introduction 1
- 1.1 Overview and Questions Addressed . . . 2
- 1.2 Scientific Contributions . . . 3
- 1.3 Thesis Structure . . . 4

2 Notation and Definitions 5
- 2.1 Sample Datasets and the Multiview Assumption . . . 6
- 2.2 Kernel Methods . . . 7

3 Background 9
- 3.1 k-means Clustering . . . 9
- 3.2 Singular Value Decomposition . . . 10
- 3.3 Canonical Correlation Analysis . . . 10
- 3.4 Kernel Methods . . . 12
  - 3.4.1 Kernel k-means . . . 13
  - 3.4.2 Kernel PCA . . . 13
  - 3.4.3 Kernel CCA . . . 14
- 3.5 Semidefinite programming . . . 14

4 Nonlinear Multiview Canonical Correlation Analysis 17
- 4.1 Related Work . . . 17
- 4.2 Sum of Correlations . . . 18
- 4.3 Local Solutions . . . 20
- 4.4 Proposed Extensions . . . 20
  - 4.4.1 Dual Representation and Kernels . . . 21
  - 4.4.2 Computing Several Sets of Canonical Vectors . . . 23
  - 4.4.3 Implementation . . . 26

5 Relaxations and Bounds 29
- 5.1 NP-Hardness . . . 29
- 5.2 Semidefinite Programming Relaxation . . . 30

5.3 Upper Bounds on QCQP . . . 33
5.4 Random Projections and Multivariate Regression . . . 37

6 Cross-Lingual Document Similarity 39
6.1 Problem Definition . . . 39
6.2 Cross-Lingual Models . . . 40
6.3 Related Work . . . 41
6.4 Notation . . . 42
6.5 k-means . . . 43
6.6 Cross-Lingual Latent Semantic Indexing . . . 44
6.7 Bi-Lingual Document Analysis CCA . . . 45
6.8 Hub Language Based CCA Extension . . . 46

7 Applications to Cluster Linking 51
7.1 Problem Definition . . . 53
7.2 Algorithm . . . 54

8 Experiments 57
8.1 Synthetic Experiments . . . 57
8.1.1 Generating Synthetic Problem Instances . . . 57
8.1.2 Convergence of Horst's algorithm on Synthetic Data . . . 58
8.1.3 SDP and Horst Solutions on Synthetic Problems . . . 58
8.2 Experiments on EuroParl Corpus . . . 61
8.2.1 Dataset and preprocessing . . . 62
8.2.2 Local Versus Global Approaches . . . 62
8.3 Experiments on the Wikipedia Corpus . . . 63
8.3.1 Wikipedia Comparable Corpus . . . 64
8.3.2 Experiments With Missing Alignment Data . . . 64
8.3.3 Evaluation Of Cross-Lingual Event Linking . . . 66
8.3.4 Remarks on the Scalability of the Implementation . . . 70
8.3.5 Remarks on the Reproducibility of Experiments . . . 71

9 Conclusions 73
9.1 Discussion . . . 73
9.2 Future Work . . . 74

References 75

Bibliography 81

Biography 83

List of Figures

Figure 1.1: The main contributions and related work 3

Figure 4.1: The block structure 18

Figure 6.1: Multilingual corpus matrices 43
Figure 6.2: k-means algorithm and coordinate change. 44
Figure 6.3: LSI multilingual corpus matrix decomposition. 45

Figure 7.1: An example of an event 52
Figure 7.2: Cluster linking 53

Figure 8.1: Convergence rates: varying problem instances 59
Figure 8.2: Convergence rates: varying initial conditions 60
Figure 8.3: Average of mean reciprocal ranks. 68

List of Tables

Table 8.1: Random Gram matrix. 61
Table 8.2: Random spectrum sampling. 61
Table 8.3: Random 1-dim structure sampling. 62
Table 8.4: Train and test sum of correlation. 63
Table 8.5: Training – test sizes 65
Table 8.6: Pairwise retrieval 66
Table 8.7: Dimensionality drift 66
Table 8.8: Accuracy of cluster linking for several cross-lingual similarity models . . 69
Table 8.9: Accuracy of cluster linking: large vs small clusters 69
Table 8.10: Story linking accuracy for several feature sets 70
Table 8.11: Story linking accuracy for several language pairs 70

List of Algorithms

Algorithm 4.1: Horst's algorithm 21
Algorithm 4.2: Horst's algorithm for computing a k-dimensional representation . . 26

Algorithm 5.1: Random projections basis generation 38

Algorithm 7.1: Algorithm for identifying candidate clusters 54

Abbreviations

CCA	...	Canonical Correlation Analysis
MCCA	...	Multiview Canonical Correlation Analysis
KCCA	...	Kernel Canonical Correlation Analysis
SUMCOR	...	Sum of Correlations
SSCOR	...	Sum of Squared Correlations
LSI	...	Latent Semantic Indexing
SVD	...	Singular Value Decomposition
TSVD	...	Truncated Singular Value Decomposition
PCA	...	Principal Component Analysis
CG	...	Conjugate Gradient
SDP	...	Semidefinite Programming
fMRI	...	functional Magnetic Resonance Imaging
i.i.d	...	independently and identically distributed
QCQP	...	Quadratically Constrained Quadratic Program
MEP	...	Multivariate Eigenvalue Problem
BQO	...	Binary Quadratic Optimization
TF	...	Term Frequency
IDF	...	Inverse Document Frequency
TFIDF	...	Term Frequency Inverse Document Frequency
GB	...	Gigabyte
GCCA	...	Generalized Canonical Correlation Analysis
dim	...	dimension
CCAR	...	Canonical Correlation Analysis Regression
CL-VSM	...	Cross-Lingual Vector Space Model
JPLSA	...	Joint Probabilistic Latent Semantic Analysis
CPLSA	...	Coupled Probabilistic Latent Semantic Analysis
PLTM	...	Polylingual Topic Models
CPLSA	...	Coupled Probabilistic LSA
PCLLSA	...	Probabilistic Cross-Lingual LSA
CL-ESA	...	Cross-Lingual Explicit Semantic Analysis
OPCA	...	Oriented Principal Component Analysis
MMR	...	Mean Reciprocal Rank
AMMR	...	Average Mean Reciprocal Rank

Symbols

$\mathbb{R}$	...	real numbers
$\mathbb{R}^{m \times n}$	...	matrices with m rows and n columns
$\mathbb{N}$	...	natural numbers
$\mathbb{S}_{+}^{n}$	...	space of symmetric positive semidefinite n-by-n matrices
$\mathbb{S}_{++}^{n}$	...	space of symmetric positive definite n-by-n matrices
$\vec{1}_k$	...	column vector with k dimensions with all coefficients equal to 1
ρ	...	correlation coefficient
μ_X	...	empirical mean of a column sample matrix X
$Cov(\cdot, \cdot)$	...	covariance function that takes two aligned sample matrices as input
$\kappa(\cdot, \cdot)$	...	kernel function
$\|\cdot\|_F$	...	Frobenius norm
$\|\cdot\|_1$	...	operator norm corresponding to ℓ_1 norm
$\phi(\cdot)$	...	feature map from a set to a Hilbert space

Glossary

Correlation coefficient measures the degree of linear dependence between two univariate random variables.
Canonical Correlation Analysis is a way of measuring the linear relationship between two multidimensional variables.
Principal Component Analysis is a dimensionality reduction technique based on maximization of variance.
Singular Value Decomposition is a factorization of a real or complex matrix.
Vector Space Model is a representation of textual data in a vector space, based on counting the occurrences of words, which correspond to vector space dimensions.
Latent Semantic Indexing is a text analysis technique based on the singular value decomposition of the corpus matrix.
k-means Clustering is a grouping algorithm that groups objects according to their similarity.
Symmetric positive semidefinite matrix is a symmetric matrix with nonnegative eigenvalues.
Semidefinite programming is a subfield of convex optimization concerned with the optimization of a linear objective function over the intersection of the cone of positive semidefinite matrices with an affine space.
Dual Representation expresses vectors as linear combinations over the training set.
Hilbert space is a vector space equipped with an inner product.
Kernel functions provide a way to manipulate data as though it were projected into a higher dimensional space.
Kernel methods are a class of algorithms for pattern analysis, based on embeddings into a Hilbert space.

Chapter 1

Introduction

Pattern analysis is the process of finding structure or regularity in a set of data. For example, if each data instance represents a point in a vector space, we might be interested in the following question: does the dataset lie in a lower dimensional subspace (does it admit a more compact representation)? In this case, the subspace represents a pattern (structure or regularity) discovered in the data. Principal Component Analysis provides a solution to such a question.

This thesis deals with finding patterns in datasets that exhibit a *multi-view* aspect: that is, for each instance of data there are two or more representations (views) available. We refer to such datasets as *aligned* (multi-view) datasets. As an example of a two-view dataset, consider a dataset where each instance is represented by a visual image and a textual description. Another example is a *parallel multi-lingual corpus*, where given n languages, each data instance consists of n documents, one for each language and the documents are related by being translations of each other. The patterns that we are interested in represent regularities within each view that have associated regularities in other views. For example, when dealing with text, a type of pattern that is often of interest is a distribution over words from a fixed vocabulary, referred to as a *topic vector*. Given a collection of documents in a single language, a typical problem is to find relevant topic vectors that summarize the document collection. The multi-view variant of the problem then corresponds to finding sets of multiple representations of topic vectors (one per language). Methods that extract such multi-representation patterns represent the main subject of the thesis.

There are several possible applications of such an analysis. The patterns themselves can be of interest for explorative analysis. For example, given an aligned dataset of fMRI brain scans and visual images that were shown to the subjects as scans were taken, we can investigate how the brain functions by looking at relationships between brain activation regions and patterns in visual images. Another example of application is to use the multi-view patterns as maps into a representation independent space. For example, representing visual images and textual descriptions in the same space can be used for cross-modal information retrieval, where one searches for images relevant to a query text, (or documents relevant to a given image) by using standard information retrieval techniques. In addition, the optimization problem related to one of the generalizations of Canonical Correlation Analysis (CCA) that we study appears in applications that range from control theory, blind source separation to multiple subject fMRI analysis.

1.1 Overview and Questions Addressed

We will now provide a high-level overview of the results presented in the thesis and highlight the related work that motivated or enabled the results, all of which is summarized in Figure 1.1. Canonical Correlation Analysis (CCA) [1], a well established method that looks for patterns in two-view datasets, has been extended by other authors in several ways: a nonlinear extension was proposed in [2], which was later applied to text in [3]. It has been extended to more than two sets of variables in [4], where a formulation called *Sum of Correlations* (SUMCOR) was presented, together with an iterative algorithm to finding local solutions, known as the *Horst's algorithm*. Results on global optimality of a subset of SUMCOR problems was established in [5] and [6]. Several alternative generalizations of CCA were proposed in [7], where the most relevant extension to the thesis is the *Sum of Squared Correlations* (SSCOR).

The thesis starts with two questions:

- How can we extend the Horst's algorithm to handle nonlinear patterns and how to find several sets of canonical vectors? Does the extension provably converge?
- What is the computational complexity of the SUMCOR problem formulation?

We present an extension that is closely related to [2] and show that it does converge to local solutions. We prove that in general the computational complexity of the SUMCOR problem is NP-hard. In light of these results, several questions arose:

- Can we find a convex relaxation of the problem?
- Can we obtain computationally tractable bounds on the SUMCOR objective?

We show how to relax the problem to an instance of a Semidefinite Programming (SDP), whose solutions yield computable bounds on global optimality. The results related to SUMCOR complexity and SDP relaxations are available in [8] and submitted for publication.

Applying the theory to practice opened up the following questions:

1. How to apply the SDP bounds to high dimensional data?
2. How can one use the methods to perform cross-lingual document analysis?
3. How does one handle missing data?

We addressed the questions in the following way:

1. We proposed a preprocessing step that reduces the number of variables in the SDP derived from a SUMCOR problem instance which makes the relaxation computationally tractable.
2. We present the methodology for building cross-lingual similarity functions and apply it to the task of cross-lingual cluster linking. The application is relevant to global analysis of high-volume multi-lingual news streams.
3. We address the problem of missing data in our application to cross-lingual text mining for datasets where data was missing in a structured way and show that under certain assumptions, the SSCOR problem formulation can be reduced to a low-dimensional eigenvalue problem. The results related to SSCOR reduction and cross-lingual applications are available in [9]. An alternative application of the SSCOR reformulation to cluster linking is published in [10].

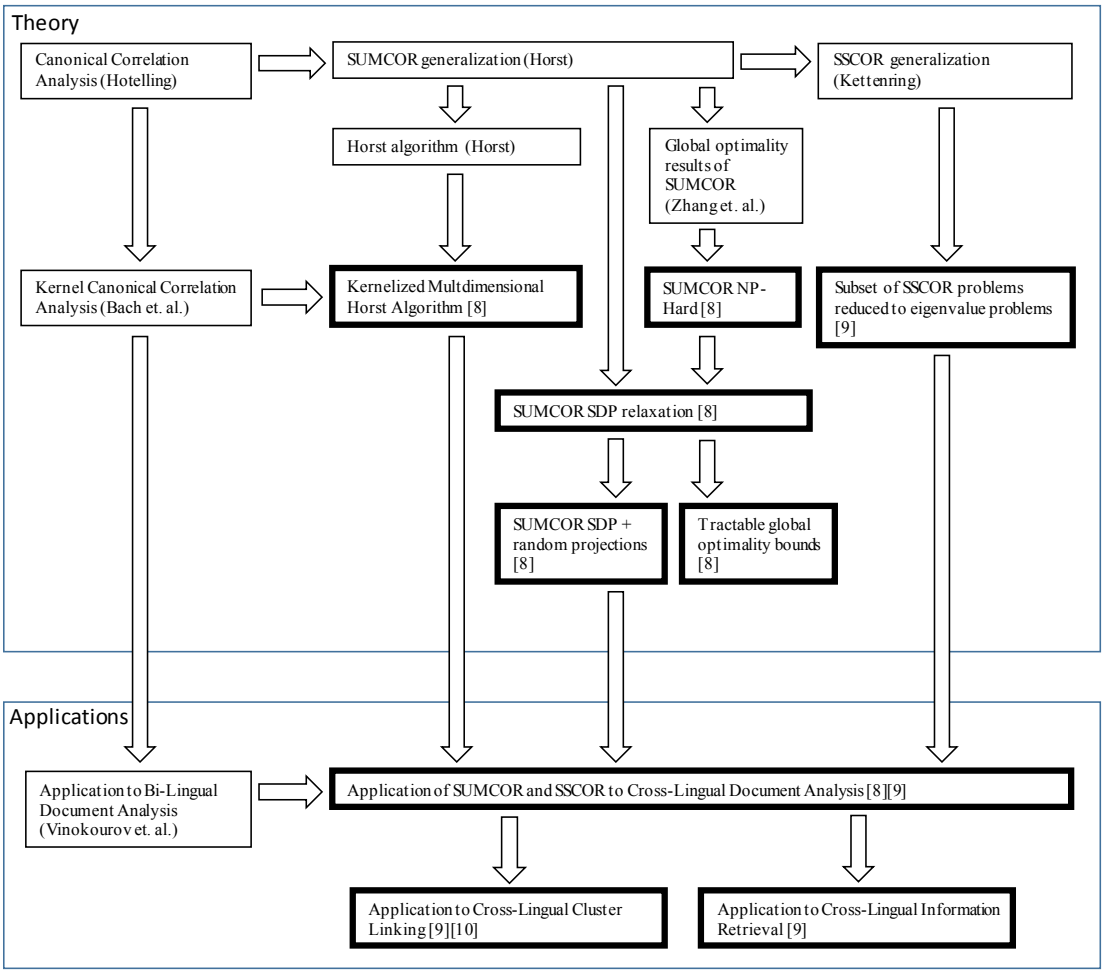


Figure 1.1: The main contributions, represented by text boxes with thick border, are positioned with respect to the related work.

1.2 Scientific Contributions

We now list the main scientific contributions of the thesis and their references:

- A novel algorithm based on the Horst's algorithm that can extract several sets of nonlinear patterns [8]
- A proof that in general the Sum of Correlations problem is NP-hard [8]
- A semidefinite programming relaxation of the SUMCOR problem and several new bounds on global optimization of the SUMCOR problems [8]
- A novel approach to apply the SDP bounds on high-dimensional data [8]
- A novel approach to building cross-lingual similarity functions and its application to cross-lingual information retrieval and cross-lingual cluster linking [9][10]
- Addressing the missing data problem, a novel reduction of a subset of SSCOR problems to eigenvalue problems [9]

1.3 Thesis Structure

The rest of the thesis is structured as follows. Chapter 2 introduces notation and some definitions. For background we describe three pattern analysis methods that are the most relevant for the thesis and explain how they can be adapted for analysis of nonlinear patterns in Chapter 3. Chapter 4 introduces a central problem of the thesis: generalizations of Canonical Correlation Analysis (CCA) and the original contributions that extend the method to nonlinear and higher-dimensional setting. In Chapter 5 we prove the result on the complexity of a particular generalization and study global optimality guarantees based on semidefinite relaxations. Chapter 6 discusses an application of multiview learning to building cross-lingual similarity models. We show how a particular structure of the data can be exploited to express a particular generalization of CCA as an eigenvector problem. Chapter 7 then shows how the cross-lingual similarity measures can be used to perform cross-lingual cluster linking, relevant for large scale monitoring of global news in multiple languages. In Chapter 8 several experiments are presented both on synthetic and real datasets. Finally, Chapter 9 concludes the thesis and discusses possible future directions.

Chapter 2

Notation and Definitions

We first introduce the notation we use throughout the thesis:

- Column vectors are denoted by lowercase letters, e.g. x and matrices are denoted by uppercase letters, e.g. X.
- Subscripts are used to enumerate vectors or matrices, e.g. x_1, x_2, X_1, except in the special case of the identity matrix, I_n and the zero matrix $0_{k,l}$. In these cases, the subscripts denote row and column dimensions.
- We use superscripted symbol T for vector and matrix transpose, e.g. x^T.
- Let $\|v\|$ or $\|v\|_2$ denote the ℓ_2 norm of the vector v and let $\|A\|_F$, $\|A\|_1$ and $\|A\|_2$ denote the Frobenius norm and the operator norms induced by 1-norm and 2-norm respectively.
- MATLAB notation [11]
  - The i-th element of vector x is denoted by $x(i)$ and the matrix entry in the i-th row and j-th column is denoted by $X(i,j)$.
  - The i-th row of matrix X is denoted by $X(i,:)$ and the j-th column by $X(:,j)$. matrix elements, rows and columns (e.g. $X(i,j), X(i,:), X(:,j)$)
  - Matrix concatenation: $[A\ B]$ represents horizontal concatenation and $[A;B]$ represents vertical concatenation.
  - $diag(v)$ denotes a diagonal matrix whose diagonal entries correspond to vector v.
  - 1_k denotes a column vector with k elements all equal to 1.
- Spaces
  - $\mathbb{R}^n$ denotes the n-dimensional real vector space.
  - $\mathbb{R}^{n\times m}$ denotes the $(n\cdot m)$-dimensional vector space used when specifying matrix dimensions.
  - $\mathbb{N}$ denotes the natural numbers.
  - $\mathbb{S}^n_+$ denotes the space of symmetric positive semidefinite n-by-n matrices.
  - $\mathbb{S}^n_{++}$ denotes the space of symmetric positive definite n-by-n matrices.
- Random vectors are denoted by calligraphic letters, e.g. $\mathcal{X}$ and $\mathcal{X} \in \mathbb{R}^n$ denotes their dimension.

2.1 Sample Datasets and the Multiview Assumption

The following definitions will be relevant for our discussion of kernel versions of the methods relevant to this thesis.

Definition 2.1. A *sample dataset* with ℓ samples and n dimensions is a set

$$S := \{x_1, \ldots, x_\ell\},$$

where $x_i \in \mathbb{R}^n$ are generated independently and identically distributed (i.i.d.) according to an underlying distribution.

Definition 2.2. A $n \times \ell$ *sample matrix* based on a dataset S with ℓ samples and n dimensions is obtained by horizontally concatenating the samples:

$$X := [x_1 \cdots x_\ell]\,.$$

Definition 2.3. A *multiview sample dataset* with ℓ samples and m views is a set:

$$S = \left\{ \left(x_1^{(1)}, \ldots, x_1^{(m)}\right), \ldots, \left(x_\ell^{(1)}, \ldots, x_\ell^{(m)}\right) \right\},$$

where $x_i^{(j)} \in \mathbb{R}^{n_j}$ corresponds to the j-th view of the i-th sample. We assume that each sample point was generated independently and identically distributed (i.i.d.) according to an underlying distribution with a specific structure. We assume that the samples represent different views of an underlying object, that is, the observed random vectors are functions of an unobserved random vector:

$$\left(\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(m)}\right) = (f_1(\mathcal{O}), \ldots, f_m(\mathcal{O}))\,.$$

Definition 2.4. Given a multiview dataset S with ℓ samples and m views, we form a matrix for view i by using horizontal concatenation:

$$X_i := \left[x_1^{(i)} \cdots x_\ell^{(i)}\right].$$

We refer to the set $\{X_1, \ldots, X_m\}$ as *multiview aligned matrices.*

In general, we will say that two matrices are *aligned*, if their columns form observation vector pairs, related to a multiview dataset.

Definition 2.5. A *multiview stacked matrix* based on a dataset S with ℓ samples and m views is a matrix obtained by vertically concatenating the multiview aligned matrices:

$$X := [X_1; \cdots; X_m]\,.$$

Definition 2.6. The sample matrix is centered if its rows sum to zero.

Definition 2.7. Given a $n \times \ell$ sample matrix X, the *empirical mean* $\mu_X \in \mathbb{R}^n$ is computed as:

$$\mu_X(i) = \frac{1}{\ell} \sum_j X(i,j).$$

Definition 2.8. Given a sample matrix $X \in \mathbb{R}^{n \times \ell}$ the empirical covariance is defined as:

$$Cov(X) := \frac{1}{n-1}(X - \mu_X \cdot \vec{1}_\ell^T) \cdot (X - \mu_X \cdot \vec{1}_\ell^T)^T.$$

Definition 2.9. *Empirical variance* $Var(X)$ is defined as the empirical covariance for single dimensional sample matrices, that is, when n equals 1.

Definition 2.10. Given two aligned sample matrices $X_1 \in \mathbb{R}^{n_1 \times \ell}$ and $X_2 \in \mathbb{R}^{n_2 \times \ell}$ the empirical cross-covariance is defined as:

$$Cov(X_1, X_2) := \frac{1}{n-1}(X_1 - \mu_{X_1} \cdot \vec{1}_\ell^T) \cdot (X_2 - \mu_{X_2} \cdot \vec{1}_\ell^T)^T.$$

2.2 Kernel Methods

The following definitions will be relevant for our discussion of kernel versions of the methods relevant to this thesis. For definitions of standard concepts from topology we refer the reader to standard texts [12].

Definition 2.11. A *metric space* is an ordered pair (M, d), where M is a set and $d : M \times M \to \mathbb{R}$ is a *metric* on M, i.e., a function which satisfies for all $x, y, z \in M$:

1. $d(x, y) \geq 0$,
2. $d(x, y) = 0 \iff x = y$,
3. $d(x, y) = d(y, x)$,
4. $d(x, z) \leq d(x, y) + d(y, z)$.

Definition 2.12. Let X be a metric space equipped with a metric $d : X \times X \to \mathbb{R}$. A sequence $(x_1, x_2, x_3, \ldots)$ is a *Cauchy* sequence, if for every $\epsilon > 0$ there exists a positive integer N such that for all $m, n > N$:

$$d(x_m, x_n) < \epsilon.$$

Definition 2.13. A metric space X is complete, if every Cauchy sequence of elements in X converges to an element of X.

Definition 2.14. A topological space H is *separable* if it contains a countable dense subset; that is, there exists a sequence $(x_n)_{n=1}^{\infty}$ such that every nonempty open subset of the space contains at least one element of the sequence.

Definition 2.15. A *Hilbert space* H is an inner product space that is both *separable* and *complete*.

Definition 2.16. Let $V \subset \mathbb{R}^n$. A *kernel* is a function $\kappa : V \times V \to \mathbb{R}$ that satisfies:

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle, \quad \forall x, y \in V,$$

where $\phi : V \to H$ is a function from V to a Hilbert space.

Definition 2.17. Given a sample matrix $X \in \mathbb{R}^{n \times \ell}$ and a kernel function κ, we define a *kernel matrix* $K \in \mathbb{R}^{\ell \times \ell}$ as:

$$K(i, j) := \kappa(x_i, x_j).$$

A special case of a kernel matrix is the *Gram* matrix, where the standard inner product is used as the kernel function (also referred to as the *linear kernel*).

Definition 2.18. A matrix A is *positive semidefinite* if:

$$x^T A x \geq 0, \quad \forall x.$$

The space of positive semidefinite matrices is denoted by $\mathbb{S}_+$.

Definition 2.19. A matrix A is *positive definite* if:

$$x^T A x > 0, \quad \forall x.$$

The space of positive definite matrices is denoted by $\mathbb{S}_{++}$.

Chapter 3

Background

The central subjects in the thesis revolve around statistical approaches to finding structure in one, two or more sets of variates. We will introduce two methods that find structure in a single set of variates: k-means clustering and Singular Value Decomposition (SVD) for dimensionality reduction, which is closely related to Principal Component Analysis (PCA). We will then present Canonical Correlation Analysis (CCA), a method for studying two sets of variates. We will also briefly cover kernel method extensions of the methods and present some results on Semidefinite Programming.

3.1 k-means Clustering

The k-means algorithm [13] is perhaps the most well-known and widely-used clustering algorithm. In spirit of analysis on multiview methods that is to be presented, we will formulate k-means as a matrix factorization problem. Given an $n \times \ell$ sample matrix (Definition 2.2) the goal is to find the best rank k approximation under additional constraints:

$$\begin{aligned} \underset{C \in \mathbb{R}^{n \times k}, P \in \mathbb{R}^{\ell \times k}}{\text{minimize}} \quad & \|X - C \cdot P^T\|_F^2, \\ \text{subject to} \quad & P(i,j) \in \{0,1\}, \quad \forall i,j \\ & \sum_j P(i,j) = 1, \quad \forall i. \end{aligned} \tag{3.1}$$

The interpretation of the additional constraints on matrix P is that they force each sample vector (column in X) to select precisely one column of C to approximate it and the objective function corresponds to minimizing a sum of squared errors made by approximating points with centroids.

The matrix C in Equation 3.1 is uniquely defined for a given P, since for any given set of points in $\mathbb{R}^n$, the point that minimizes the sum of squared distances to the set is the mean. Since each column of P selects a subset of columns of X, C can be expressed as:

$$C := X \cdot P \text{diag}(\vec{1}_\ell^T \cdot P)^{-1} P^T, \tag{3.2}$$

where the inverse of the diagonal matrix corresponds to division by the set size when computing the mean ($\vec{1}_\ell^T \cdot P$ counts the number of points assigned to each of the k clusters). In addition, given C, the assignment P that minimizes the sum of squared errors can be found by:

$$P(i, j^*) = 1, \quad \text{where} \quad j^* = \arg\min_j \|X(:,i) - C(:,j)\| \tag{3.3}$$

A popular approach [13] to solving the problem in Equation 3.1 is to start with an initial assignment and alternate between updating C given P and vice versa. The approach is

widely used in practice, even though it is susceptible to finding local minima. In general, the problem is known to be NP-hard [14].

3.2 Singular Value Decomposition

The second factorization based approach that is relevant to our work is based on the Truncated Singular Value Decomposition [11] (TSVD). It is closely related to Principal Component Analysis [15] (PCA), a well established approach to dimensionality reduction.

Given an $n \times \ell$ sample matrix (Definition 2.2) the goal is to find a best approximation with rank at most k under additional constraints:

$$\begin{aligned} \underset{U \in \mathbb{R}^{n \times k}, S \in \mathbb{R}^{k \times k}, V \in \mathbb{R}^{\ell \times k}}{\text{minimize}} \quad & \|X - U \cdot S \cdot V^T\|_F, \\ \text{subject to} \quad & U^T U = I_k \\ & V^T V = I_k \\ & S = \operatorname{diag}(\sigma), \quad \sigma \in \mathbb{R}^k, \quad \sigma(i) \geq 0. \end{aligned} \tag{3.4}$$

The method of PCA is based on a low rank decomposition of the empirical covariance matrix, computed based on the sample matrix. The main idea is to find a subspace that accounts for as much as variability in the data as possible. The first principal component is defined as the one-dimensional subspace that maximizes the variance of the data when projected onto it. Formally, it solves the following problem:

$$\begin{aligned} \underset{u \in \mathbb{R}^n}{\text{maximize}} \quad & \operatorname{Var}(u^T \cdot X), \\ \text{subject to} \quad & \|u\| = 1. \end{aligned} \tag{3.5}$$

The other principal vectors can be obtained by deflation [16], or equivalently solving the eigenvalue problem:

$$\begin{aligned} \underset{U \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad & \|\operatorname{Cov}(X) - U \cdot \Lambda \cdot U^T\|_F, \\ \text{subject to} \quad & U^T U = I_k \\ & \Lambda = \operatorname{diag}(\lambda), \quad \lambda \in \mathbb{R}^k. \end{aligned} \tag{3.6}$$

One of the main applications of PCA is as a dimensionality reduction technique, where the data is projected to the space spanned by the normalized eigenvectors (also called principal vectors). In typical applications a truncated eigenvalue decomposition is used, where one discards the principal vectors with small eigenvalues (similar to truncated SVDs).

If the data matrix is centered, then the solution U of TSVD and the eigenvector basis U of PCA will coincide.

3.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [1] is a general procedure for studying relationships between two sets of random variables. It is based on analyzing the cross-covariance matrix between two random vectors with the aim of identifying linear relationships between them. We will start with intuitions and then give a formal presentation.

Roughly speaking, given two random vectors $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ we are interested in "non-trivial" pairs of functions $(f^{(1)}, f^{(2)})$ such that there is a "dependence" between $f^{(1)}(\mathcal{X}^{(1)})$ and $f^{(2)}(\mathcal{X}^{(2)})$. The "dependence" we consider is linear (possibly in a Hilbert space). The "non-triviality" of the functions is a requirement that guards us against trivial solutions,

such as $f^{(1)}(x) := 0 \cdot x$, $f^{(2)}(y) := 0 \cdot y$ - that is, $f^{(1)}(\mathcal{X}^{(1)})$ and $\mathcal{X}^{(1)}$ should share some information, and analogously for $f^{(2)}(\mathcal{X}^{(2)})$ and $\mathcal{X}^{(2)}$. In other words, $f^{(1)}$ and $f^{(2)}$ should not destroy the original signals. When we are interested in more than one good pair of functions, for instance, a family of pairs $(f_i^{(1)}, f_i^{(2)})$, we typically require additional constraints to prevent non-trivial solutions by enforcing that $f_i^{(1)}\left(\mathcal{X}^{(1)}\right)$ and $f_{j\neq i}^{(1)}\left(\mathcal{X}^{(1)}\right)$ share no information, and similarly for $f_i^{(2)}$. We are interested in essentially different function pairs.

There are several possible applications of such an analysis. For example, a common scenario involves analyzing objects $o \in \mathcal{O}$, where $\mathcal{O}$ is some underlying space, which are not directly observable, but are only observable as images of transformations $F^{(1)} : \mathcal{O} \to \mathbb{R}^p$ and $F^{(2)} : \mathcal{O} \to \mathbb{R}^q$. That is, we do not have access to o but only to $\left(F^{(1)}(o), F^{(2)}(o)\right)$. Then finding function pairs $(f_i^{(1)}, f_i^{(2)})$ so that $f_i^{(1)}(F^{(1)}(o))$ behave similarly as $f_i^{(2)}(F^{(2)}(o))$ can be interpreted as finding coupled parametrizations of image spaces of $F^{(1)}$ and $F^{(2)}$ which agree on $\mathcal{O}$. This enables applications such as cross-modal information retrieval, classification, clustering, etc. If $F^{(1)}$ encodes a visual image and $F^{(2)}$ encodes a textual description of the scene, we can perform text input based search over a collection of images, see [17]. Bi-lingual document analysis is another application, see [3], [18]. The pattern functions $(f_i^{(1)}, f_i^{(2)})$ themselves can be interesting to study for exploratory purposes.

Formally, let

$$S = \{\left(F^{(1)}(o_1), F^{(2)}(o_1)\right), \ldots, \left(F^{(1)}(o_n), F^{(2)}(o_n)\right)\}$$

represent a sample of n pairs drawn independently at random according to the underlying distribution, where $F^{(1)}(x_i) \in \mathbb{R}^p$ and $F^{(2)}(x_i) \in \mathbb{R}^q$ represent feature vectors from p and q-dimensional vector spaces. Let $X^{(1)} = [F^{(1)}(o_1), \ldots, F^{(1)}(o_n)]$ and let $X^{(2)} = [F^{(2)}(o_1), \ldots, F^{(2)}(o_n)]$ be the matrices with observation vectors as columns (using MATLAB notation).

The idea is to find two vectors $w^{(1)} \in \mathbb{R}^p$ and $w^{(2)} \in \mathbb{R}^q$ so that the random variables $w^{(1)T} \cdot \mathcal{X}^{(1)}$ and $w^{(2)T} \cdot \mathcal{X}^{(2)}$ are maximally correlated ($w^{(1)T}$ and $w^{(2)T}$ map the random vectors to random variables, by computing weighted sums of vector components). By using the sample matrix notation $X^{(1)}$ and $X^{(2)}$ this problem can be formulated as the following optimization problem:

$$\underset{w^{(1)}\in\mathbb{R}^p, w^{(2)}\in\mathbb{R}^q}{\text{maximize}} \quad \frac{w^{(1)T} Cov(X^{(1)}, X^{(2)}) w^{(2)}}{\sqrt{w^{(1)T} Cov(X^{(1)}) w^{(1)}} \sqrt{w^{(2)T} Cov(X^{(2)}) w^{(2)}}}, \tag{3.7}$$

where $Cov(X^{(1)})$ and $Cov(X^{(2)})$ are empirical estimates of variances of $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ respectively and $Cov(X^{(1)}, X^{(2)})$ is an estimate for the covariance matrix as defined in Definition 2.8 and Definition 2.10. The optimization problem can be reduced to a generalized eigenvalue problem [17]:

$$\begin{bmatrix} 0 & Cov(X^{(1)}, X^{(2)}) \\ Cov(X^{(2)}, X^{(1)}) & 0 \end{bmatrix} \cdot \begin{bmatrix} w^{(1)} \\ w^{(2)} \end{bmatrix} = \lambda \cdot \begin{bmatrix} Cov(X^{(1)}, X^{(1)}) & 0 \\ 0 & Cov(X^{(2)}, X^{(2)}) \end{bmatrix} \cdot \begin{bmatrix} w^{(1)} \\ w^{(2)}. \end{bmatrix} \tag{3.8}$$

If the matrices $Cov(X^{(1)})$ and $Cov(X^{(2)})$ are not invertible, the problem is ill posed. One can use a regularization technique by replacing $Cov(X^{(1)})$ with $(1-\tau)Cov(X^{(1)})+\tau I$, where $\tau \in [0, 1]$ is the regularization coefficient and I is the identity matrix (and analogously for $Cov(X^{(2)})$), see [16] for details. A single canonical variable is usually inadequate

in representing the original random vector and typically one looks for k projection pairs $(w_1^{(1)}, w_1^{(2)}), \ldots, (w_k^{(1)}, w_k^{(2)})$, so that $w_i^{(1)}$ and $w_i^{(2)}$ are highly correlated and $w_i^{(1)}$ is uncorrelated with $w_j^{(1)}$ for $j \neq i$ and analogously for $w^{(2)}$.

The formulation in Equation 3.8 can be reformulated as a symmetric eigenvalue problem and solved efficiently. In case the dimensions of the problem p and q are large and observation vectors are sparse, one can consider an iterative method, for example Lanczos algorithm [19]). Alternatively, if the number of observation vectors n is not prohibitively large, one can reformulate the problem to its dual representation which can be combined with a "kernel trick" [2] to yield nonlinear version of CCA, which will be discussed in the next section.

3.4 Kernel Methods

The methods discussed so far looked for patterns expressed in the same space as the sample dataset. We now discuss how we can extend the methods to finding nonlinear patterns by using the framework of kernel methods. To look for nonlinear patterns in the original space, one first uses a nonlinear map ϕ to map the input data into a Hilbert space, where linear patterns are then extracted. If the Hilbert space is high dimensional the strategy might be computationally intractable. Let ϕ denote a feature map and κ its kernel function as defined in Definition 2.16, that is:

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle.$$

If evaluating the kernel is feasible, then certain methods can be solved efficiently, even if ϕ is a map to an infinite dimensional space. If in a given method the data and model parameters interact only through inner products, then we can attempt to reformulate the problem in terms of kernel matrices.

The general approach to kernelization of a method is to try to express the solution in a dual basis, that is, the basis spanned by the training instances. The following two theorems provide an alternative characterization of kernels and relate the kernel functions with implicit feature maps. For an extended treatment of the concepts and the proofs we refer the reader to [16].

Theorem 3.1. *A function $\kappa : X \times X \to \mathbb{R}$, which is either continuous or has a finite domain, is a kernel function if and only if its kernel matrix is symmetric and positive semidefinite on any finite set of points.*

Theorem 3.2. *Given a kernel function $\kappa : X \times X \to \mathbb{R}$ we can reconstruct the implicit Hilbert space H and the feature map ϕ as:*

$$H := \{\sum_{i=1}^{k} \alpha_i \kappa(x_i, \cdot) : k \in \mathbb{N}, x_i \in X\},$$

$$\phi(x) := \kappa(x, \cdot),$$

and the inner product is defined as:

$$\langle \phi(x), \phi(y) \rangle := \kappa(x, y).$$

3.4.1 Kernel k-means

Instead of working directly with columns of X we now work with $\phi(X(:,i))$ for a ϕ that corresponds to a choice of kernel κ. Applying the iterative procedure described in Section 3.1 to the input space mapped by ϕ involves computing squared distances between columns $X(:,i)$ and centroids, which can be expressed as:

$$\|\phi(X(:,i) - \frac{1}{|S_k|}\sum_{j\in S_k}\phi(X(:,j))\|^2,$$

where S_k denotes the set of indices of points currently assigned to centroids k. Since $\|x-y\|^2 = \langle x-y, x-y\rangle = \langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle$ the above quantity can be expressed using kernel evaluations as:

$$\kappa(X(:,i),X(:,i)) - 2\frac{1}{|S_k|}\sum_{j\in S_k}\kappa(X(:,i),(X(:,j)) + \frac{1}{|S_k|^2}\sum_{j,\ell\in S_k}\kappa(X(:,j),X(:,\ell)).$$

Using Theorem 3.2, a new point x is mapped to $\kappa(x,\cdot)$ and assigned the cluster that minimizes

$$\arg\min_k \|\kappa(x,\cdot) - \frac{1}{|S_k|}\sum_{i\in S_k}\kappa(X(:,i),\cdot)\|^2.$$

Again, computing the cluster assignment can be fully specified through kernel evaluations.

3.4.2 Kernel PCA

We will assume that the data is centered to simplify presentation. The solution to PCA is expressed as an eigenvector decomposition of the covariance matrix. There is a direct correspondence between the eigen-decompositions of the scaled covariance $(\ell-1)Cov(X)$ and the Gram matrix $K = X^TX$. If (v,λ) is an eigenvector-eigenvalue pair for K, then (Xv,λ) is an eigenvalue pair for $(\ell-1)Cov(X)$:

$$(\ell-1)Cov(X)Xv = XX^TXv = XKv = \lambda Xv.$$

Since $\|Xv\| = \sqrt{v^TX^TXv} = \sqrt{\lambda v^Tv} = \sqrt{\lambda}$, the solutions to the original problem are expressed as linear combinations over the training examples of the form $\sqrt{\lambda}Xv$. Motivated by this correspondence, the kernel methods approach thus analyzes the spectrum of the kernel matrix, given a kernel function.

Since the kernel matrix is symmetric and positive-definite, the eigenvectors form an orthonormal set in the Hilbert space and can thus be used as a projection. The normalized eigenvector v (with an associated λ) is expressed in the Hilbert space:

$$\sqrt{\lambda}\sum_i v(i)\phi(X(:,i))$$

which means that projecting a new point $\phi(x)$ in the kernel PCA coordinates is computed as:

$$P(\phi(x))_i := \sqrt{\lambda}\sum_i v(i)\kappa(x,X(:,i)).$$

The centering assumption is not needed, as centering can be implemented as an operation on the kernel matrix, as we will present in Section 4.4.3.

3.4.3 Kernel CCA

We will now present how to apply kernel methods to CCA and obtain the method known as Kernel CCA (KCCA). The idea is to express the optimization problem in its dual form, where we express the solutions as linear combinations of their corresponding training instances. All the interactions with the data will be expressed through inner products, which will make the problem compatible with nonlinear feature maps based on their respective kernels.

Let us express the vectors $w^{(1)}$ and $w^{(2)}$ in Equation 3.8 in their dual form by using new coordinate vectors $\alpha^{(1)} \in \mathbb{R}^\ell$ and $\alpha^{(2)} \in \mathbb{R}^\ell$ so that:

$$w^{(1)} = X^{(1)}\alpha^{(1)},$$
$$w^{(2)} = X^{(2)}\alpha^{(2)}.$$

Assuming that the data is centered, the original optimization problem in Equation 3.9 is expressed as:

$$\underset{\alpha^{(1)}\in\mathbb{R}^\ell,\ \alpha^{(2)}\in\mathbb{R}^\ell}{\text{maximize}} \quad \frac{\alpha^{(1)T}K^{(1)}K^{(2)}\alpha^{(2)}}{\sqrt{\alpha^{(1)T}K^{(1)}K^{(1)}\alpha^{(1)}}\sqrt{\alpha^{(2)T}K^{(2)}K^{(2)}\alpha^{(2)}}}. \tag{3.9}$$

Applying the Lagrangian multiplier technique one can arrive at the dual form of the generalized eigenvalue problem:

$$\begin{bmatrix} 0 & K^{(1)}K^{(2)} \\ K^{(2)}, K^{(1)} & 0 \end{bmatrix} \cdot \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \end{bmatrix} = \tag{3.10}$$

$$\lambda \cdot \begin{bmatrix} K^{(1)}K^{(1)} & 0 \\ 0 & K^{(2)}K^{(2)} \end{bmatrix} \cdot \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)}. \end{bmatrix} \tag{3.11}$$

To prevent overfitting, one can regularize the variances $(1-\tau)Cov(X^{(1)}) + \tau I_p$ and $(1-\tau)Cov(X^(2)) + \tau I_q$. The corresponding regularized dual variances are expressed as: $(1-\tau)K^{(1)}K^{(1)} + \tau K^{(1)}$ and $(1-\tau)K^{(2)}K^{(2)} + \tau K^{(2)}$ can then replace the diagonal blocks of the right side of Equation 3.10.

3.5 Semidefinite programming

Semidefinite programming (SDP) problems are a subclass of convex optimization problems that involve optimizing a linear function over the intersection of the cone of positive semidefinite matrices with an affine space.

A *primal* SDP is given by:

$$\begin{aligned} \min_{X\in\mathbb{S}_+^n} \quad & \operatorname{Tr}(AX) \\ \text{subject to} \quad & \operatorname{Tr}(B_i X) = b_i, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{3.12}$$

A *dual* SDP problem is closely related to the primal and is given by:

$$\begin{aligned} \max_{y\in\mathbb{R}^m} \quad & \langle b, y\rangle \\ \text{subject to} \quad & A - \sum_{i=1}^{m} y_i B_i \in \mathbb{S}_n. \end{aligned} \tag{3.13}$$

An important property of both problems is that they are convex (see [20]) which means that any locally optimal solution is also globally optimal. The following two theorems relate the optimal objective values of primal and dual formulations. For proofs of the next two theorems please refer to a standard textbook on convex optimization [20].

Theorem 3.3 (Weak duality). *Let X^* denote the optimal solution to the problem 3.12 and y^* the optimal solution to the problem 3.13. Then*

$$\mathrm{Tr}(AX^*) - \langle b, y^* \rangle \geq 0.$$

The difference between the objective values is referred to as the duality gap.

Theorem 3.4 (Strong duality). *If the optimization problem 3.12 is bounded from below (optimal solution X^* to the problem 3.12 exists) and there exists $X_0 \in \mathbb{S}^n_{++}$ such that*

$$\mathrm{Tr}(X_0 B_i) = b_i, \quad \forall i = 0, \ldots, m, \tag{3.14}$$

then the optimal solution y^* to the problem 3.13 exists and:

$$\mathrm{Tr}(AX^*) = \langle b, y^* \rangle.$$

Not all SDP problems result in a zero duality gap, but as it will turn out, the problems that we will be interested in do. This is relevant from a practical perspective, since many modern optimization SDP solvers, such as primal-dual interior point methods, converge only when the duality gap is zero.

We now present an important result that has been established [21] on the rank of SDP solutions (relevant to Chapter 5).

Theorem 3.5. *If there is an optimal solution for SDP, then there is an optimal solution of SDP whose rank r satisfies:*

$$\frac{r(r+1)}{2} < m.$$

There also exists a constructive version of the statement [22]. See [23, Chapter 6.5] for a proof.

Summary. This chapter introduced some well known data analysis techniques that play an important role in the following chapters. More concretely, CCA and KCCA serve as the basis of our proposed extensions in Chapter 4, SVD and k-means are used as benchmarks for cross-lingual document analysis in Chapter 8 and SVD is used as a preprocessing step for an original method that will be introduced in Chapter 6. The chapter also introduced SDP problems, which will be used to obtain global optimality results in Chapter 5.

Chapter 4

Nonlinear Multiview Canonical Correlation Analysis

This chapter will discuss generalizations of CCA to analysis of multiple sets of variables. The problem is then known as the Multi-set Canonical Correlation Analysis (MCCA), or sometimes Multiview Canonical Correlation Analysis. Whereas it can be shown that CCA can be solved using an (generalized) eigenvalue computation, MCCA is a much more difficult problem. We will describe a generalization proposed in [7] and present an iterative locally optimal solution proposed in [4], referred to as the *Horst's algorithm*, which represents a starting point of our work.

For use in practical applications, we propose a novel algorithm based on two original contributions of the Horst's algorithm: we adapt the methods to use kernels and to find multi-dimensional solutions, analogous to finding multiple principal directions in PCA and multiple pairs of canonical vectors in CCA.

4.1 Related Work

CCA, introduced in Section 3.3, was developed to detect linear relations between two sets of variables. Typical uses of CCA include statistical tests of dependence between two random vectors, exploratory analysis on multi-view data, dimensionality reduction and finding a common embedding of two random vectors that share mutual information.

CCA has been generalized in two directions: extending the method to finding nonlinear relations by using kernel methods [2][17] and extending the method to more than two sets of variables [7], which we presented in Section 3.4. Among several proposed generalizations in [7] the most important for our work are the *Sum of Correlations* (SUMCOR) and *Sum of Squared Correlations* (SSCOR) generalizations, where SUMCOR will be the main focus of the current chapter and Chapter 5 and SSCOR will be studied in Chapter 6.

There the goal is to project m sets of random variables to m univariate random variables, which are pair-wise highly correlated on average [1]. An iterative method to solve the SUMCOR generalization was proposed in [4] and the proof of convergence was established in [24]. In [24] it was shown that a generic SUMCOR problem admits exponentially many locally optimal solutions. In [6] the authors identified a subset of SUMCOR problems for which the iterative procedure converges to a global maximizer (Their results apply to non-negative irreducible quadratic forms). In Chapter 5 we will discuss results on the problem complexity and global optimality conditions.

[1] Given m univariate random variables, one can compute $\binom{m}{2}$ correlation coefficients, one for each pair of variables.

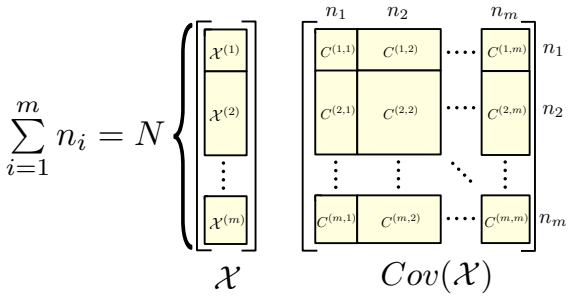


Figure 4.1: The block structure of the random vector $\mathcal{X}$ and the corresponding covariance block structure.

This chapter will focus on the local iterative approach [4] with the aim of extending it so it may be used in applications similar to KCCA. Here we show how the method can be extended to finding non-linear patterns and finding more than one set of canonical variates. Our work is related to [25] where a deflation scheme is used together with the Newton method to find several sets of canonical variates. Our nonlinear generalization is related to [26], where the main difference lies in the fact that we "kernelized" the problem, whereas the authors in [26] worked with explicit nonlinear feature representation.

We now list some applications of the SUMCOR formulation. In [27] an optimization problem for multi-subject functional magnetic resonance imaging (fMRI) alignment is proposed, which can be formulated as a SUMCOR problem (performing whitening on each set of variables). Another application of the SUMCOR formulation can be found in [25], where it is used for group blind source separation on fMRI data from multiple subjects. An optimization problem equivalent to SUMCOR also arises in control theory [28] in the form of linear sensitivity analysis of systems of differential equations.

4.2 Sum of Correlations

In this section we will discuss a generalization of CCA to more than two views, which finds a set of directions (one per view) which maximizes the average correlation (computed for each pair of views).

We assume that we are given a centered random vector $\mathcal{X} \in \mathbb{R}^N$ as defined in Definition 2.3, where $\mathcal{X}$ is composed of m subsets of random variables referred to as views. We assume that the indices of components of each set are contiguous, i.e. $\mathcal{X}$ is a concatenation of blocks that correspond to views.

Additional Notation. Let m denote the number of blocks and N the total number of variables in $\mathcal{X}$. Then

$$b := (n_1, \ldots, n_m)\,, \ \sum_{i=1}^{m} b\,(i) = N,$$

encodes the number of elements in each of the block. We denote the corresponding sub-vectors as $\mathcal{X}^{(i)} \in \mathbb{R}^{n_i}$ (i-th block-row of vector $\mathcal{X}$) and the sub-matrices as $C^{(i,j)} \in \mathbb{R}^{n_i \times n_j}$ (i-th block-row, j-th block column of matrix C); see Figure 4.1. For example, in CCA, there are only two sets, so $m = 2$.

Formally, given $w \in \mathbb{R}^N$ we define m random variables $\mathcal{Z}_i$ (one-dimensional projections of random block components of $\mathcal{X}$) as:

$$\mathcal{Z}_i := \sum_{j=1}^{n_i} \mathcal{X}^{(i)}\,(j)\, w^{(i)}\,(j) = \mathcal{X}^{(i)T} \cdot w^{(i)}.$$

Let $\rho(x, y)$ denote the correlation coefficient between two random variables:

$$\rho(x, y) = \frac{Cov(x, y)}{\sqrt{Cov(x, x)\, Cov(y, y)}}.$$

The correlation coefficient between $\mathcal{Z}_i$ and $\mathcal{Z}_j$ can be expressed as:

$$\rho(\mathcal{Z}_i, \mathcal{Z}_j) = \frac{w^{(i)T} C^{(i,j)} w^{(j)}}{\sqrt{w^{(i)T} C^{(i,i)} w^{(i)}} \sqrt{w^{(j)T} C^{(j,j)} w^{(j)}}}.$$

Initial Problem Formulation.

The problem described above can be stated as finding the set of vectors $w^{(i)}$ which maximize

$$\sum_{i=1}^{m} \sum_{j=i+1}^{m} \rho(\mathcal{Z}_i, \mathcal{Z}_j). \tag{SUMCOR}$$

We refer to this problem as Multi-set Canonical Correlation Analysis (MCCA). Note that it reduces to CCA when $m = 2$. The solution - that is, the set of components $\left(w^{(1)}, \ldots, w^{(m)}\right)$, are referred to as the set of canonical vectors.

Another formulation proposed in [7] is the *Sum of Squared Correlations* (SSCOR)

$$\sum_{i=1}^{m} \sum_{j=i+1}^{m} \rho(\mathcal{Z}_i, \mathcal{Z}_j)^2. \tag{SSCOR}$$

The second formulation is invariant to the signs of the correlation coefficients. It will be of importance in Chapter 6.

Reformulating the Optimization Problem.

Expanding SUMCOR, we get:

$$\max_{w \in \mathbb{R}^N} \quad \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{w^{(i)T} C^{(i,j)} w^{(j)}}{\sqrt{w^{(i)T} C^{(i,i)} w^{(i)}} \sqrt{w^{(j)T} C^{(j,j)} w^{(j)}}}.$$

Observe that the solution is invariant to block scaling (only the direction matters): if $\left(w^{(1)}, \ldots, w^{(m)}\right)$ is a solution, then $\left(\alpha_1 \cdot w^{(1)}, \ldots, \alpha_m \cdot w^{(m)}\right)$ is also a solution for $\alpha_i > 0$. We may therefore impose constraints $w^{(i)T} C^{(i,i)} w^{(i)} = 1$, which only affect the norm. This yields the following equivalent constrained problem:

$$\begin{aligned} \underset{w \in \mathbb{R}^N}{\text{maximize}} \quad & \sum_{i=1}^{m} \sum_{j=i+1}^{m} w^{(i)T} C^{(i,j)} w^{(j)} \\ \text{subject to} \quad & w^{(i)T} C^{(i,i)} w^{(i)} = 1, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.1}$$

We further multiply the objective by 2 and add a constant m. Note that this does not affect the optimal solution. Using the equalities: $w^{(i)T} C^{(i,j)} w^{(j)} = w^{(j)T} C^{(j,i)} w^{(i)}$ and $w^{(i)T} C^{(i,i)} w^{(i)} = 1$, we obtain:

$$\begin{aligned} \underset{w \in \mathbb{R}^N}{\text{maximize}} \quad & \sum_{i=1}^{m} \sum_{j=1}^{m} w^{(i)T} C^{(i,j)} w^{(j)} \\ \text{subject to} \quad & w^{(i)T} C^{(i,i)} w^{(i)} = 1, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.2}$$

This transforms the objective function into a quadratic form $w^T C w$. To simplify the constraints, assume that $C^{(i,i)}$ is strictly positive definite. If $C^{(i,i)}$ is not full rank, then using the eigenvalue decomposition $C^{(i,i)} = V \Lambda V^T$, where $V \in \mathbb{R}^{n_i \times k}$, $\Lambda \in \mathbb{R}^{n_i \times k}$, $\Lambda > 0$,

$k < n_i$, we substitute $\mathcal{X}^{(i)}$ with $V^T\mathcal{X}^{(i)} \in \mathbb{R}^k$, for which the covariance matrix is strictly positive definite.

From the strict positive definiteness it follows that $C^{(i,i)}$ admits a Cholesky decomposition: there exists an invertible triangular matrix D_i such that $C^{(i,i)} = D_i^T D_i$.

Finally, using the block structure b, we substitute $w^{(i)}$ with $D_i^{-1}x^{(i)}$ and define $A \in \mathbb{R}^N$ as:

$$A^{(i,j)} := {D_i}^{-T} C^{(i,j)} {D_j}^{-1},$$

leading to the simplified problem:

$$\begin{aligned} \max_{x \in \mathbb{R}^N} \quad & x^T A x \\ \text{subject to} \quad & x^{(i)T} x^{(i)} = 1, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{QCQP}$$

It turns out that (QCQP) is simpler to manipulate than (SUMCOR), so we use this form from this point on.

Using the technique of Lagrange multipliers (see [24]) we can relate the stationary points of the problem in Equation (QCQP) to solutions of a nonlinear system of equations:

$$\sum_{j=1}^{m} A^{(i,j)} x^{(i)} = \lambda_i x^{(i)}, \quad \forall i = 1, \ldots, m, \tag{4.3}$$

where $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$. This formulation is referred to as a Multivariate Eigenvalue Problem (MEP). The following theorem has been established [24].

Theorem 4.1. *Consider the problem in Equation (4.3). If the matrix A is symmetric and generic, then the number of solutions to the system is exactly*

$$\Pi_{i=1}^{m} 2n_i.$$

This means that in general the number of solutions grows exponentially with the number of views. Notice however that A in Equation (QCQP) is not generic as each of the blocks on its diagonal is an identify matrix.

4.3 Local Solutions

In this section, we give an algorithm that converges to a locally optimal solution of the problem (QCQP), when the matrix A is symmetric, positive-definite and generic (the proof was established in [24]).

The algorithm can be interpreted as a generalization of the power iteration method (also known as the Von Mises iteration), a classical approach to finding the largest solutions to the eigenvalue problem $Ax = \lambda x$. The general iterative procedure is given as Algorithm 4.1.

In the case of $m = 1$, Algorithm 4.1 corresponds exactly to the power iteration. While the algorithm's convergence to a local optimum is guaranteed, its convergence rate is not known. In Chapter 8 we will examine the convergence rate on synthetic data.

4.4 Proposed Extensions

Here we present two extensions of MCCA: how to use kernel methods with MCCA to find nonlinear dependencies in the data; and an algorithm to finding more than one set of correlation vectors.

Algorithm 4.1: Horst's algorithm

Input: matrix $A \in \mathbb{S}_N^+$, block structure $b = (n_1, \ldots, n_m)$, initial vector $x_0 \in \mathbb{R}^N$ with $\|x^{(i)}\| > 0$,

$x \leftarrow x_0$;
for $iter = 1$ to $maxiter$ **do**
  $x \leftarrow Ax$;
  for $i = 1$ to m **do**
    $x^{(i)} \leftarrow \frac{x^{(i)}}{\|x^{(i)}\|}$
  end for
end for

Output: x

4.4.1 Dual Representation and Kernels

We return to formulation (4.1):

$$\begin{aligned} \max_{w \in \mathbb{R}^N} \quad & \sum_{i=1}^{m} \sum_{j=i+1}^{m} w^{(i)T} C^{(i,j)} w^{(j)} \\ \text{subject to} \quad & w^{(i)T} C^{(i,i)} w^{(i)} = 1, \quad \forall i = 1, \ldots, m, \end{aligned}$$

where $b = (n_1, \ldots, n_m)$ denotes the block structure and $\sum_i n_i = N$. In the previous sections, we focused on manipulating covariance matrices only and omitted details on their estimation based on finite samples. In this section, we will use a formulation that explicitly presents the empirical estimates of covariances, which will enable us to apply kernel methods. Let $\mathcal{X}$ be a random vector distributed over $\mathbb{R}^N$ with $E(\mathcal{X}) = 0$. Let $X \in \mathbb{R}^{N \times s}$ represent a sample of s observations of $\mathcal{X}$, where each observation corresponds to a column vector. The empirical covariance of $\mathcal{X}$ based on the sample matrix X is expressed as:

$$\overline{Cov(\mathcal{X})} = \frac{1}{s-1} X X^T.$$

If $s < N$, then $\overline{Cov(\mathcal{X})}$ is singular which makes the optimization problem ill-posed and may lead to overfitting (discovering spurious patterns in the data). These issues are addressed by using regularization techniques, typically a shrinkage estimator $\overline{Cov(\mathcal{X})_\kappa}$ is defined as:

$$\overline{Cov(\mathcal{X})_\kappa} = (1-\kappa) \frac{1}{s-1} X X^T + \kappa I_N,$$

where $\kappa \in [0, 1]$. Higher values of the regularization parameter κ lead to better numerical stability, at a cost of solving a different problem to the one originally posed, with solutions that may be far from optimal. In practice the parameter has to be tuned using validation techniques, for example using cross-validation.

Using the block structure b, Equation (4.1) becomes:

$$\begin{aligned} \max_{w \in \mathbb{R}^N} \quad & \frac{1}{s-1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} w^{(i)T} X^{(i)} X^{(j)T} w^{(j)} \\ \text{subject to} \quad & w^{(i)T} \left(\frac{1-\kappa}{s-1} X^{(i)} X^{(i)T} + \kappa I_N \right) w^{(i)} = 1, \\ & \forall i = 1, \ldots, m. \end{aligned} \tag{4.4}$$

To express each component $w^{(i)}$ in terms of the columns of $X^{(i)}$, let w have block structure $b_w = (n_1, \ldots, n_m)$ where $\sum_i n_i = N$, and let $y \in \mathbb{R}^{m \cdot s}$ have block structure $b_y(i) = s, \forall i = 1, \ldots, m$. The component $w^{(i)}$ can be expressed as:

$$w^{(i)} = \sum_{j=1}^{s} y^{(i)}(j) X^{(i)}(:, j) = X^{(i)} y^{(i)}. \tag{4.5}$$

We refer to y as dual variables. Let $K_i = X^{(i)T} X^{(i)} \in \mathbb{R}^{s \times s}$ denote the Gram matrix. We can now express the problem (4.4) in terms of the dual variables:

$$\begin{aligned} \max_{y \in \mathbb{R}^{m \cdot s}} \quad & \frac{1}{s-1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} y^{(i)T} K_i K_j^T y^{(j)} \\ \text{subject to} \quad & y^{(i)T} \left(\frac{1-\kappa}{s-1} K_i K_i^T + \kappa K_i \right) y^{(i)} = 1, \\ & \forall i = 1, \ldots, m. \end{aligned} \tag{4.6}$$

Expressing the problem in terms of Gram matrices makes it amenable to using kernel methods (see [16]). It remains to check that the formulations (4.4) and (4.6) are equivalent.

Lemma 4.2. There exists a solution to (4.4) which can be expressed as (4.5).

Proof. We prove the lemma by contradiction. Assume that no optimal solution can be expressed as (4.5) and let u be an optimal solution to the problem (4.4). Without loss of generality, assume that $u^{(1)}$ does not lie in the column space of $X^{(1)}$:

$$u^{(1)} = z_\perp + X^{(1)} y^{(1)},$$

where

$$z_\perp \neq 0_{n_1} \quad \text{and} \quad X^{(1)T} z_\perp = 0_s.$$

We show that $\bar{u}$, defined as $\bar{u}^{(i)} := u^{(i)}, \forall i > 1$ and $\bar{u}^{(1)} := \frac{1}{\gamma} X^{(1)} y^{(1)}$, where

$$\gamma := \sqrt{y^{(1)T} X^{(1)T} \left(\frac{1-\kappa}{s-1} X^{(1)} X^{(1)T} + \kappa I_N \right) X^{(1)} y^{(1)}},$$

strictly increases the objective function, which contradicts the assumption that u is optimal. Clearly, $\bar{u}$ is a feasible solution. Positive definiteness of $\frac{1-\kappa}{s-1} X^{(1)} X^{(1)T} + \kappa I_N$, coupled with the fact that $z_\perp^T z_\perp > 0$, implies that $0 < \gamma < 1$. Let $E := \sum_{j=2}^{m} \left(X^{(1)} y^{(1)} \right)^T X^{(1)} X^{(j)T} u^{(j)}$.

If $E < 0$, then the vector $[-u^{(1)T} \; u^{(2)T} \; \cdots \; u^{(m)T}]^T$ strictly increases the objective function, which is a contradiction. We also obtain a contradiction if $E = 0$, since any nonzero $v \in \mathbb{R}^s$ for which $X^{(1)} v \neq 0_{n_1}$ can be used to obtain a solution to the problem (4.4) expressed as (4.5) (after re-scaling such that $\overline{Cov\left(\mathcal{X}^{(1)} v\right)}_\kappa = 1$ and if necessary multiplying it by -1 so that $\sum_{j=2}^{m} \left(X^{(1)} v \right)^T X^{(1)} X^{(j)T} u^{(j)} \geq 0$). Thus, we may assume that $E > 0$.

The following inequality completes the proof, since it shows that $\bar{u}$ increases the objective function:

$$
\begin{aligned}
&\frac{1}{s-1}\sum_{j=2}^{m} u^{(1)T}X^{(1)}X^{(j)T}u^{(j)} \\
&= \frac{1}{s-1}\sum_{j=2}^{m} \left(z_\perp + X^{(1)}y^{(1)}\right)^T X^{(1)}X^{(j)T}u^{(j)} \\
&= \frac{1}{s-1}\sum_{j=2}^{m} \left(X^{(1)}y^{(1)}\right)^T X^{(1)}X^{(j)T}u^{(j)} \\
&< \frac{1}{s-1}\sum_{j=2}^{m} \frac{1}{\gamma}\left(X^{(1)}y^{(1)}\right)^T X^{(1)}X^{(j)T}u^{(j)}.
\end{aligned}
$$

□

Typically the matrices K_i are ill conditioned (or even singular when the data is centered) and it is advantageous to constrain the magnitude of dual coefficients as well as the variance in the original problem. We address this by introducing a first order approximation to the dual regularized variance. Let

$$\widetilde{K_i} := \left(\sqrt{\frac{1-\kappa}{s-1}}K_i + \frac{\kappa}{2}\sqrt{\frac{s-1}{1-\kappa}}I_s\right).$$

The covariance becomes:

$$\overline{Cov\left(\mathcal{X}^{(i)}\right)}_\kappa = \frac{1-\kappa}{s-1}K_iK_i^T + \kappa K_i \approx \widetilde{K_i}\widetilde{K_i}^T.$$

This approximation has two advantages: it is invertible and is in a factorized form. We exploit the latter when obtaining a convergent local method. The final optimization is then expressed as:

$$
\begin{aligned}
\max_{y\in\mathbb{R}^{m\cdot s}} \quad & \frac{1}{s-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} y^{(i)T}K_iK_j^Ty^{(j)} \\
\text{subject to} \quad & y^{(i)T}\widetilde{K_i}\widetilde{K_i}^Ty^{(i)} = 1, \quad \forall i = 1,\ldots,m.
\end{aligned}
\tag{4.7}
$$

The problem can be interpreted as maximizing covariance while constraining variance and magnitude of dual coefficients.

4.4.2 Computing Several Sets of Canonical Vectors

Usually a one-dimensional representation does not sufficiently capture all the information in the data and higher dimensional subspaces are needed. After computing the first set of primal canonical vectors we proceed to computing the next set. The next set should be almost as highly correlated as the first one, but essentially "different" from the first one. We achieve this by imposing additional constraints for every view. Namely, all projection vectors in view i are uncorrelated with respect to $\widetilde{K}_i^2$ (this is similar to the approach in two view regularized kernel CCA[2]).

Let $Y = [y_1,\ldots,y_k] \in \mathbb{R}^{m\cdot s\times k}$ represent k sets of canonical vectors, where

$$Y^{(\ell)T}\widetilde{K_\ell^2}Y^{(\ell)} = I_k \quad \forall \ell = 1,\ldots,m.$$

The equation above states that each canonical vector has unit regularized variance and that different canonical vectors corresponding to the same view are uncorrelated (orthogonal with respect to $\widetilde{K_i^2}$).

We will now extend the set of constraints in the optimization (4.7) to enforce the orthogonality by introducing a modified optimization problem. The problem will involve the matrix Y, whose columns represent k sets of canonical vectors and we will show how to derive the $(k+1)$-th set of canonical vectors y whose block components $y^{(i)}$ will be uncorrelated to corresponding block components of $Y(:,j)^{(i)}$.

$$\begin{aligned} \max_{y\in\mathbb{R}^{m\cdot s}} \quad & \frac{1}{s-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} y^{(i)T}K_iK_j^Ty^{(j)} \\ \text{subject to} \quad & y^{(i)T}\widetilde{K_i}\widetilde{K_i}^Ty^{(i)} = 1, \quad \forall i = 1,\dots,m, \\ & Y^{(i)T}\widetilde{K_i}\widetilde{K_i}^Ty^{(i)} = 0_k, \quad \forall i = 1,\dots,m. \end{aligned} \tag{4.8}$$

To use the Horst's algorithm, we first use the substitutions:

$$Z^{(i)} = \widetilde{K_i}Y^{(i)}, \quad z^{(i)} = \widetilde{K_i}y^{(i)}$$

and define the operators

$$P_i = I_s - \widetilde{K}_iY^{(i)}Y^{(i)T}\widetilde{K}_i = I_s - Z^{(i)}Z^{(i)T},$$

which map to the space orthogonal to the columns of $\widetilde{K}_iY^{(i)}$. Each P_i is a projection operator: $P_i^2 = P_i$, which follows directly from the identities above. The optimization problem in the new variables is:

$$\begin{aligned} \max_{z\in\mathbb{R}^{m\cdot s}} \quad & \frac{1}{s-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} z^{(i)T}\widetilde{K_i}^{-T}K_iK_j^T\widetilde{K_j}^{-1}z^{(j)} \\ \text{subject to} \quad & z^{(i)T}z^{(i)} = 1, \quad \forall i = 1,\dots,m, \\ & Z^{(i)T}z^{(i)} = 0_k, \quad \forall i = 1,\dots,m. \end{aligned} \tag{4.9}$$

Using the projection operators, this is equivalent to:

$$\begin{aligned} \max_{z\in\mathbb{R}^{m\cdot s}} \quad & \frac{1}{s-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} z^{(i)T}P_i^T\widetilde{K_i}^{-T}K_iK_j^T\widetilde{K_j}^{-1}P_jz^{(j)} \\ \text{subject to} \quad & z^{(i)T}z^{(i)} = 1, \quad \forall i = 1,\dots,m. \end{aligned}$$

By multiplying the objective by 2 (due to the symmetries of P_i, K_i and $\widetilde{K_i}$) and shifting the objective function by $\frac{m}{1-\kappa}$, the problem is equivalent to:

$$\begin{aligned} \max_{z\in\mathbb{R}^{m\cdot s}} \quad & \frac{1}{s-1}\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m} z^{(i)T}P_i^T\widetilde{K_i}^{-T}K_iK_j^T\widetilde{K_j}^{-1}P_jz^{(j)} + \frac{1}{1-\kappa}\sum_{i=1}^{m} z^{(i)T}z^{(i)} \\ \text{subject to} \quad & z^{(i)T}z^{(i)} = 1, \quad \forall i = 1,\dots,m. \end{aligned} \tag{4.10}$$

This optimization can be reformulated as:

$$\begin{aligned} \max_{z\in\mathbb{R}^{m\cdot s}} \quad & z^TAz \\ \text{subject to} \quad & z^{(i)T}z^{(i)} = 1, \quad \forall i = 1,\dots,m, \end{aligned} \tag{4.11}$$

where $A \in \mathbb{R}^{m \cdot s}$ with block structure $b(i) = s, \forall i = 1, \ldots, m$, is defined by:

$$A^{(i,j)} = \left\{ \begin{array}{ll} \frac{1}{s-1} P_i^T \widetilde{K_i}^{-T} K_i K_j^T \widetilde{K_j}^{-1} P_j & \text{for: } i \neq j \\ \frac{1}{1-\kappa} I_s & \text{for: } i = j \end{array} \right\}$$

Lemma 4.3. The block matrix A defined above is positive definite (i.e. $A \in \mathbb{S}_{++}^{m \cdot s}$).

Proof. A is symmetric, which follows from $P_i = P_i^T$ and $K_i = K_i^T$. Let $z \in \mathbb{R}^{m \cdot s}$. The goal is to show that $z^T A z > 0$. Let us define an auxiliary matrix W as:

$$W = \frac{1}{1-\kappa} \sum_{i=1}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} \cdot \left(\kappa K_i + \frac{\kappa^2 (s-1)}{4(1-\kappa)} I_s \right) \widetilde{K_i}^{-1} P_i z^{(i)}$$

Each summand is positive-semidefinite, i.e. $W \geq 0$ and $W > 0$ if $\exists i : P_i z^{(i)} = z^{(i)}$. What follows is a sequence of inequalities, some of which must be strict, as will be established:

$$\begin{aligned} z^T A z &= \frac{1}{s-1} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} K_i K_j^T \widetilde{K_j}^{-1} P_j z^{(j)} \\ &\quad + \frac{1}{1-\kappa} \sum_{i=1}^{m} z^{(i)T} z^{(i)} \qquad (4.12) \\ &\geq \frac{1}{s-1} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} K_i K_j^T \widetilde{K_j}^{-1} P_j z^{(j)} \\ &\quad + \frac{1}{1-\kappa} \sum_{i=1}^{m} z^{(i)T} P_i^T P_i z^{(i)} \qquad (4.13) \\ &= \frac{1}{s-1} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} K_i K_j^T \widetilde{K_j}^{-1} P_j z^{(j)} \\ &\quad + \frac{1}{1-\kappa} \sum_{i=1}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} \widetilde{K_i}^T \widetilde{K_i} \widetilde{K_i}^{-1} P_i z^{(i)} \qquad (4.14) \\ &= \frac{1}{s-1} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} K_i K_j^T \widetilde{K_j}^{-1} P_j z^{(j)} \\ &\quad + \frac{1}{s-1} \sum_{i=1}^{m} z^{(i)T} P_i^T \widetilde{K_i}^{-T} K_i K_i^T \widetilde{K_i}^{-1} P_i z^{(i)} + W \qquad (4.15) \\ &= z^T B B^T z + W \geq 0, \end{aligned}$$

where $B \in \mathbb{R}^{m \cdot s \times s}$, defined by $B^{(i)} = \frac{1}{\sqrt{s-1}} (K_i \widetilde{K_i}^{-1} P_i)^T$, with corresponding row block structure $b(i) = s$. The inequality after (4.12) holds since projection operators cannot increase norms. (4.14) is equal to (4.13) using $\widetilde{K_i}^{-T} \widetilde{K_i}^T = I$. Regrouping the terms and applying the definition of W, we obtain (4.15). The final equality follows, since the first two sums form a perfect square.

Now we will show that at least one of the two inequalities is strict. If $P_i z^{(i)} \neq z^{(i)}$ for some i, then the first inequality is strict ($\|P_i z^{(i)}\| < \|z^{(i)}\|$). Conversely, if $P_i z^{(i)} = z^{(i)}$ for all i, then $W > 0$, hence the last inequality is strict. □

Matrix A has all the required properties for convergence, so we apply Algorithm 4.1. Solutions to (4.8) are obtained by back-substituting into $y^{(i)} = \widetilde{K_i}^{-1} z^{(i)}$.

The solution to the above problem can be found by solving the following problem:

$$\sum_{j \neq i} P_i \tilde{K}_i^{-1} K_i K_j \tilde{K}_j^{-1} P_j \alpha_j + \frac{1}{(1-\kappa)^2} \alpha_i + \lambda_i \alpha_i = 0, \forall i,$$

followed by multiplying the solutions α_i by $\tilde{K}_i^{-1}$.

Eigenvalue shifting techniques can be applied to enforce positive-definiteness. The algorithm is shown in Algorithm 4.2.

Algorithm 4.2: Horst's algorithm for computing a k-dimensional representation

Input: $K_1, \ldots, K_m$, κ, $maxiter$, k,
Output: $B_1^k, \ldots, B_m^k$
  $\tilde{K}_i = (1-\kappa) K_i + \kappa I, \forall i$
  for $d = 1$ to k **do**
    Choose random vectors $\alpha_1^0, \ldots, \alpha_m^0$
    if $d > 1$ **then**
      $P_i^d = I - \tilde{K}_i B_i^{d-1} B_i^{d-1\prime} \tilde{K}_i$
      Set $\alpha_i^0 \leftarrow P_i^d \alpha_i^0, \quad \forall i$
    else
      $P_i^d = I \quad \forall i$
    end if
    $u_i^0 = K_i \tilde{K}_i^{-1} \alpha_i^0, \forall i$
    for $i = 1$ to $maxiter$ **do**
      for $j = 1$ to m **do**
        $\alpha_j^i \leftarrow P_j^d \tilde{K}_j^{-1} K_j \sum_{k \neq j} u_k^{i-1} + \left(\frac{1}{(1-\kappa)^2} \right) \alpha_j^{i-1}$
        $\alpha_j^i \leftarrow \frac{\alpha_j^i}{\sqrt{\alpha_j^{i\prime} \alpha_j^i}}$
        $u_j^i \leftarrow K_k \tilde{K}_k^{-1} \alpha_k^i$
      end for
    end for
    for $l = 1$ to m **do**
      $\beta_l^d = \tilde{K}_l^{-1} \alpha_l^{maxiter}$
      $B_l^d = [B_l^d, \beta_l^d]$ if $d > 1$
      $B_l^d = [\beta_l^d]$ if $d = 1$
    end for
  end for

4.4.3 Implementation

The algorithm requires matrix vector multiplications and inverted matrix vector multiplications. If the kernel matrices are products of sparse matrices: $K_i = X^{(i)T} X^{(i)}$ with each $X^{(i)}$ having $s \cdot n$ elements where $s \ll n$, then the kernel matrix vector multiplications cost is $2ns$ rather than n^2. Rather than computing the full inverses, we solve the system $K_i x = y$ for x, every time $K_i^{-1} y$ is needed. Since regularized kernels are symmetric and multiplying them with vectors is fast (roughly four times slower than multiplication with the original sparse matrices $X^{(i)}$), an iterative method like Conjugate Gradient (CG) [11]

is suitable. Higher regularization parameters decrease the condition number of each $\tilde{K}_i$ which speeds up CG convergence.

If we fix the number of iterations, *maxiter*, and number of CG steps, C, the computational cost of computing a k-dimensional representation is upper bounded by: $O\big(C \cdot maxiter \cdot k^2 \cdot m \cdot n \cdot s\big)$, where m is the number of views, n the number of observations and s average number of nonzero features of each observation. Since the majority of computations are sparse matrix-vector multiplications, the algorithm can be parallelized (the sparse matrices are fixed and can be split into multiple blocks).

So far, we have assumed that the data is centered. Centering can be implemented as a preprocessing step on the kernel matrix, or incorporated in the kernel matrix-vector multiplication in order to exploit sparsity in the data. Let $K := X^T X$ denote the kernel and $\mu := \frac{1}{\ell} X \vec{1}_\ell$ denote the empirical mean of the column sample matrix X.

Kernel Matrix Centering. Computing the kernel on centered data is done as follows:

$$\begin{aligned}(X - \mu\vec{1}_\ell^T)^T(X - \mu\vec{1}_\ell^T) &= X^T X - \vec{1}_\ell \mu^T X - X^T \mu \vec{1}_\ell^T + \vec{1}_\ell \mu^T \mu \vec{1}_\ell^T \\ &= X^T X - \frac{1}{\ell}\vec{1}_\ell \vec{1}_\ell^T X^T X - \frac{1}{\ell} X^T X \vec{1}_\ell \vec{1}_\ell^T + \frac{1}{\ell}\frac{1}{\ell}\vec{1}_\ell \vec{1}_\ell^T (\vec{1}_\ell^T X^T X \vec{1}_\ell) \\ &= K - \frac{1}{\ell}\vec{1}_\ell(\vec{1}_\ell^T K) - \frac{1}{\ell}(K\vec{1}_\ell)\vec{1}_\ell^T + \frac{\vec{1}_\ell^T K \vec{1}_\ell}{\ell^2}\vec{1}_\ell \vec{1}_\ell^T.\end{aligned}$$

We see that if computing K is feasible, then the centering only involves a small set of steps with quadratic complexity, which include kernel-vector multiplication and adding rank 1 matrices to the kernel.

Centering on the fly. If $X \in \mathbb{R}^{N \times \ell}$ is sparse, we cannot explicitly compute the kernel matrix K, but we can perform fast matrix-vector multiplication as $Kx = X^T(Xx)$ is a sequence of two fast multiplications. We will now incorporate centering. Then X is centered by subtracting the mean from each column, which in matrix notation corresponds to the matrix $X - \mu\vec{1}_\ell^T$. Then, centering on the fly corresponds to computing:

$$\begin{aligned}(X - \mu\vec{1}_\ell^T)^T(X - \mu\vec{1}_\ell^T)x &= (X - \mu\vec{1}_\ell^T)^T(Xx - (\vec{1}_\ell^T x)\mu) \\ &= X^T(Xx) - ((\mu^T X)x)\vec{1}_\ell - (\vec{1}_\ell^T x)(X^T\mu) + (\vec{1}_\ell^T x)(\mu^T\mu)\vec{1}_\ell.\end{aligned}$$

If the computation is dominated by $X^T(Xx)$, then centering on the fly is suitable for black box matrix-vector based methods. Also note that μ, $X^T\mu$ and $(\mu^T\mu)\vec{1}_\ell$ can be pre-computed.

Summary. This chapter presented two extensions of the SUMCOR problem and showed how the Horst's algorithm can be modified to find solutions of the extended formulations. The first extension is based on a dual representation of the optimization problem and enables looking for non-linear pattern analysis using kernel methods. The second extension enables looking for more than one set of canonical vectors which enables finding higher-dimensional projections of data into a common vector space. The next chapter will investigate the global optimality properties of the problem. Chapter 8 will provide some experimental results.

Chapter 5

Relaxations and Bounds

The current chapter focuses on the optimization aspect of the SUMCOR problem. We present our results on the computational complexity of the SUMCOR problem formulation, as well as several global optimality bounds.

5.1 NP-Hardness

In this section, we prove that the optimization problem QCQP is not only non-convex but NP-hard in general. We present a reduction from a general binary quadratic optimization problem.

Let $A \in \mathbb{R}^{m\times m}$. The binary quadratic optimization problem (BQO) is stated as:

$$\begin{array}{ll} \max\limits_{x\in\mathbb{R}^m} & x^T A x \\ \text{subject to} & x\left(i\right)^2 = 1, \quad \forall i = 1,\ldots,m. \end{array} \tag{BQO}$$

Many hard combinatorial optimization problems (e.g. the maximum cut problem and the maximum clique problem [29]) can be reduced to BQO [30], which is known to be NP-hard.

We reduce a general instance of a BQO problem to an instance of the problem (QCQP). That means that despite the special structure of the problem (QCQP) (maximizing a positive-definite quadratic form over a product of spheres), it still falls into the class of problems that are hard (under the assumption that $P \neq NP$). We begin with a general instance of BQO and through a set of simple transformations, obtain a specific instance of (QCQP), with a block structure $b = (1,\ldots,1)$. More concretely, we will show that any BQO is equivalent to a BQO problem whose assoicated matrix is a correlation matrix (positive definite with ones on the diagonal), which is an instance of a the problem (QCQP.

Consider a BQO with a corresponding generic matrix $A \in \mathbb{R}^{m\times m}$. Since $x^T A x = x^T \frac{\left(A+A^T\right)}{2} x$, we can assume that the matrix A is symmetric. The binary constraints imply that for any diagonal matrix D the quantity $x^T D x = \sum_i D\left(i,i\right)$ is constant. This means that for $c > 0$ large enough, we can replace the objective with an equivalent objective $x^T\left(A + c\cdot I\right)x$ which is a positive-definite quadratic form. Setting $c := \|A\|_1 + 1$ guarantees that $A + c\cdot I$ is positive definite, since it is strictly diagonally dominant. From now on, we assume that the matrix A in the BQO is symmetric and positive-definite. Let $g = \max_i A\left(i,i\right)$ and let $D \in \mathbb{R}^{m\times m}$ be the diagonal matrix with elements $D\left(i,i\right) = g - A\left(i,i\right)$. The BQO is then equivalent to

$$\begin{array}{ll} \max\limits_{x\in\mathbb{R}^m} & x^T \dfrac{\left(A+D\right)}{g} x \\ \text{subject to} & x\left(i\right)^2 = 1, \quad \forall i = 1,\ldots,m. \end{array} \tag{5.1}$$

The matrix $\frac{(A+D)}{g}$ is a correlation matrix since it is a symmetric positive-definite with all diagonal entries equal to 1. The optimization problem corresponds to a problem of maximizing a sum of pairwise correlations between univariate random variables (using block structure notation: $b(i) = 1, \forall i = 1, \ldots, m$). This shows that even the simple case of maximizing the sum of correlations where the optimal axes are known and only directions need to be determined, is NP-hard.

With a fixed number of views, the complexity is polynomial, which follows from recent results on the computational complexity of quadratic maps [31]. The degree of the polynomial is asymptotically equal to the product[1] $\prod_{i=1}^{m} n_i$, where n_i is the dimensionality of the i-th view. In many applications (text mining, fMRI analysis) where the n_i's are large, the computational cost becomes prohibitive even with $m = 3$.

5.2 Semidefinite Programming Relaxation

The Horst's algorithm is scalable and often works well in practice (see Chapter 8). However, it may not converge to a globally optimal solution. We show how to use a relaxation of the problem to obtain candidate solutions for the original problem. The relaxation transforms the problem into a semidefinite program and we prove lower bounds which relates the extracted SDP solution quality and the optimal QCQP objective value. We also present a set of upper bounds on the optimal QCQP objective value in Section 5.3. These can serve as certificates of optimality (or closeness to optimality) for the local solutions (obtained by the local iterative approach for example).

In this section, we relax the formulation QCQP to an SDP problem. Let $A \in \mathbb{S}_{++}^{N \times N}$ and $B_1, \ldots, B_m \in \S_{+}^{N \times N}$ be matrices which share the block structure $b := (n_1, \ldots, n_m), \sum_i b(i) = N$. The blocks $B_i^{(k,l)} \in \mathbb{R}^{n_k \times n_l}$ for $i, k, l = 1, \ldots, m$ are defined as:

$$B_i^{(k,l)} := \begin{cases} I_{n_i} & : k = i, l = i \\ 0_{k,l} & : \text{otherwise} \end{cases},$$

where $I_{n_i} \in \mathbb{R}^{n_i \times n_i}$ is an identity matrix and $0_{k,l} \in \mathbb{R}^{k \times l}$ is a matrix with all entries equal to zero. Since $x^T A x = \operatorname{Tr}\left(A x x^T\right)$, (QCQP) can be rewritten as:

$$\begin{array}{ll} \max\limits_{x \in \mathbb{R}^n} & \operatorname{Tr}\left(A x x^T\right) \\ \text{subject to} & \operatorname{Tr}\left(B_i x x^T\right) = 1, \quad \forall i = 1, \ldots, m. \end{array} \tag{QCQP2}$$

Substituting the matrix xx^T with a positive semidefinite matrix $X \in \mathbb{S}_{+}^{n}$ constrained to have rank one:

$$\begin{array}{ll} \underset{X \in \mathbb{S}_{+}^{n}}{\text{maximize}} & \operatorname{Tr}(AX) \\ \text{subject to} & \operatorname{Tr}(B_i X) = 1, \quad \forall i = 1, \ldots, m \\ & \operatorname{rank}(X) = 1. \end{array}$$

Omitting the rank-one constraint, we obtain a semi-definite program in standard form:

$$\begin{array}{ll} \max\limits_{X \in \mathbb{S}_{+}^{n}} & \operatorname{Tr}(AX) \\ \text{subject to} & \operatorname{Tr}(B_i X) = 1, \quad \forall i = 1, \ldots, m. \end{array} \tag{SDP}$$

Remark. If the solution of the problem (SDP) is rank-one, i.e. X can be expressed as $X = y \cdot y^T$, then y is the optimal solution for (QCQP).

[1] The number of local solutions to our problem of interest is established in [24].

Low Rank Solutions We use the solutions of the SDP relaxation to extract solutions to the original QCQP problem. In the process, we obtain a bound which relates the global SDP bound and the optimal value of QCQP, giving a measure of the quality of the extracted solution. If the solution is rank-one, then the relaxation is exact. Here we consider the case where the solutions are low rank (close to rank-one).

Let X^* be a solution to (SDP) and x^* be the solution to (QCQP). A straightforward way to extract a feasible solution to the problem (QCQP) from X^* is to project its leading eigenvector to the set of constraints. The following inequality always holds:

$$\operatorname{Tr}(AX^*) \geq \operatorname{Tr}\left(A \cdot x^* \cdot x^{*T}\right).$$

The quality of the solution depends on how loose this inequality is, or rather how close the matrix X^* is to rank-one (Proposition 5.3). These depend on the spectral properties of matrix X and matrix A.

The projection of a vector $y \in \mathbb{R}^N, \|y^{(i)}\| \neq 0$ to the feasible set of (QCQP) is given by map $\pi(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$:

$$\pi(y) := \left(\frac{y^{(1)}}{\|y^{(1)}\|}, \ldots, \frac{y^{(m)}}{\|y^{(m)}\|}\right)^T.$$

We require the following technical assumption:

Assumption 5.1. Let $b = (n_1, \ldots, n_m)$ denote the block structure and $\sum_i n_i = N$. Let X^* be the solution to the problem (SDP). Let x_k denote the k-th eigenvector of X^*. The assumption is the following:

$$\|x_1^{(i)}\| > 0, \forall i = 1, \ldots, m.$$

Remark. The assumption ensures that the projection to the feasible set, $\pi(\cdot)$, is well defined. In our experiments, this was always the case, but does not hold in general as one can find counterexamples, for example, let $m = 2$ and define:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

All symmetric positive definite matrices that satisfy the constraints are of the form:

$$X = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix},$$

where $\epsilon < 0$, and attain the same value $\operatorname{Tr}(AX) = 2$, which is thus optimal. Since this also holds for $X = I$ where the leading eigenvector can be defined as $x_1 = [01]^T$, which is in contrast with the assumption. This counterexample has a specific structure: the zeros on the off-diagonal of A make the views completely independent and the problem is in some sense a trivial multi-view problem. It is possible that the assumption holds under additional conditions, for example, that the matrix is not block diagonalizable under the problem's block structure.

The following lemma depends on the projection operator and thus relies on the assumption.

Lemma 5.2. Let $b = (n_1, \ldots, n_m)$ denote the block structure and $\sum_i n_i = N$. Let X^* be the solution to (SDP) for which Assumption 5.1 holds. Let $\alpha_i := \frac{1}{\|x_1^{(i)}\|}$. If X^* can be expressed as:

$$X^* = \lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T,$$

where $\|x_1\| = 1$, $\|x_2\| = 1$, $\langle x_1, x_2 \rangle$ and $\lambda_1 > 1 > \lambda_2$, then

$$\lambda_1 \leq \alpha_i \alpha_j \leq \frac{\lambda_1}{1 - \lambda_2}.$$

Proof. The constraints in problem (SDP) are equivalent to:

$$\lambda_1 \|x_1^{(i)}\|^2 + \lambda_2 \|x_2^{(i)}\|^2 = 1, \forall i = 1, \ldots, m.$$

Since $\lambda_2 < 1$ and $\|x_2^{(i)}\|^2 \leq 1$, we see that $0 \leq \lambda_2 \|x_2^{(i)}\|^2 < 1$. Consequently:

$$0 < \frac{1 - \lambda_2}{\lambda_1} < \|x_2^{(i)}\|^2 \leq \frac{1}{\lambda_1}.$$

From $\alpha_i = \frac{1}{x_1^{(i)}}$ it follows that

$$\sqrt{\lambda_1} \leq \alpha_i \leq \sqrt{\frac{\lambda_1}{1 - \lambda_2}},$$

and finally:

$$\lambda_1 \leq \alpha_i \alpha_j \leq \frac{\lambda_1}{1 - \lambda_2}, \forall i, j = 1, \ldots, m.$$

□

Proposition 5.3. Let $b = (n_1, \ldots, n_m)$ denote the block structure and $\sum_i n_i = N$. Let X^* be the solution to (SDP) and x^* be the solution to (QCQP). Let x_k denote the k-th eigenvector of X^*, $\alpha_i := \frac{1}{\|x_1^{(i)}\|}$, $\psi := \operatorname{Tr}(AX^*)$, and $\phi := \operatorname{Tr}\left(A \cdot x^* \cdot x^{*T}\right)$. If X^* can be expressed as

$$X^* = \lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T,$$

where x_1 and x_2 have unit length and $\lambda_1 > 1 > \lambda_2$, then:

$$\psi - \pi(x_1)^T A \pi(x_1) \leq \left(\frac{1}{1 - \lambda_2} - 1\right) m^2 + \lambda_2 \|A\|_2.$$

Proof. First we note that:

$$\psi \geq \phi \geq \pi(x_1)^T A \pi(x_1).$$

We can expand the left hand side in terms of the two eigenvectors. Grouping the elements by the vector terms and using basic manipulations we arrive at the result.

$$\begin{aligned}
\psi - \pi(x_1)^T A \pi(x_1) &= \lambda_1 \textstyle\sum_{i,j} x_1^{(i)T} A^{(i,j)} x_1^{(j)T} \\
&\quad + \lambda_2 \textstyle\sum_{i,j} x_2^{(i)T} A^{(i,j)} x_2^{(j)T} - \sum_{i,j} \alpha_i \alpha_j x_1^{(i)T} A^{(i,j)} x_1^{(j)T} \\
&\leq \textstyle\sum_{i,j} (\lambda_1 - \alpha_i \alpha_j) x_1^{(i)T} A^{(i,j)} x_1^{(j)T} + \lambda_2 \|A\|_2 \\
&\leq -\textstyle\sum_{i,j} (\lambda_1 - \alpha_i \alpha_j) \cdot |x_1^{(i)T} A^{(i,j)} x_1^{(j)T}| + \lambda_2 \|A\|_2 \\
&\leq \left(\tfrac{\lambda_1}{1-\lambda_2} - \lambda_1\right) \cdot \textstyle\sum_{i,j} \frac{1}{\alpha_i \alpha_j} + \lambda_2 \|A\|_2 \\
&\leq \left(\tfrac{\lambda_1}{1-\lambda_2} - \lambda_1\right) \cdot \tfrac{m^2}{\lambda_1} + \lambda_2 \|A\|_2 \\
&= \left(\tfrac{1}{1-\lambda_2} - 1\right) \cdot m^2 + \lambda_2 \|A\|_2.
\end{aligned}$$

□

Remark. Proposition 5.3 is based on the assumption that the solution is of rank 2. Although the assumption is very specific, it holds in general for $m = 3$ which is a result of Theorem 3.5.

A similar bound can be derived for the general case, provided that the solution to problem (SDP) is close to rank-one.

Proposition 5.4. Let $b = (n_1, \ldots, n_m)$ denote the block structure and $\sum_i n_i = N$. Let X^* be the solution to the problem (SDP) and x^* the solution to the problem (QCQP). Let x_k denote the k-th eigenvector of X^*, $\alpha_i := \frac{1}{\|x_1^{(i)}\|}$ and $\psi := \operatorname{Tr}(AX^*)$. If X^* can be expressed as

$$X^* = \lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T + \cdots + \lambda_n x_n x_n^T,$$

where each x_i has unit length and $\lambda_1 > 1 > \sum_{i=2,\ldots,n} \lambda_i$, then:

$$\begin{aligned} &\psi - \pi(x_1)^T A \pi(x_1) \\ &\leq \left(\frac{1}{1 - \sum_{i=2,\ldots,n} \lambda_i} - 1 \right) m^2 + \left(\sum_{i=2,\ldots,n} \lambda_i \right) \|A\|_2. \end{aligned}$$

5.3 Upper Bounds on QCQP

We now present several upper bounds on the optimal QCQP objective value based on the spectral properties of the QCQP matrix A. We then bound the values of the SDP solutions and present two constant relative accuracy guarantees.

L_2 Norm Bound We show an upper bound on the objective of (QCQP) based on the largest eigenvalue of the problem matrix A.

Proposition 5.5. The objective value of (QCQP) is upper bounded by $m \cdot \|A\|_2$.

Proof. The problem (QCQP) remains the same if we add a redundant constraint $x^T x = m$ obtained by summing the constraints $\sum_{i=1}^m \left(x^{(i)T} x^{(i)} - 1\right) = 0$. We then relax the problem by dropping the original constraints to get:

$$\begin{array}{ll} \max\limits_{x \in \mathbb{R}^N} & x^T A x \\ \text{subject to} & x^T x = m. \end{array} \tag{5.2}$$

Since $\|A\|_2 = \max_{\|x\|_2 = 1} x^T A x$, it follows that the optimal objective value of the problem (5.2) equals $m \cdot \|A\|_2$. □

Bound on possible SDP objective values The next lemma establishes two simple bounds on the optimal value of the SDP problem.

Lemma 5.6. Let X^* be the solution to the problem (SDP) and let $\psi := \operatorname{Tr}(AX^*)$. Then

$$m \leq \psi \leq m^2.$$

Proof. Express X^* as:

$$X^* = \sum_{i=1,\ldots,N} \lambda_i x_i x_i^T,$$

where each x_i has unit length and $\lambda_1 \geq \ldots \geq \lambda_N \geq 0$. The lower bound follows from the fact that ψ upper bounds the optimal objective value of problem (QCQP) which is lower bounded by m. The lower bound corresponds to the case of zero sum of correlations.

To prove the upper bound, first observe that the constraints in (SDP) imply that $\sum_{i=1,\ldots,N} \lambda_i = m$. Let $y \in \mathbb{R}^N$ and $\|y\| = 1$. Let $z := \left(\|y^{(1)}\|, \ldots, \|y^{(m)}\|\right)^T$. Observe that

$\|z\| = 1$ and that $\|zz^T\|_2 = 1$. Define $e \in \mathbb{R}^m, e(i) = 1, \forall i = 1, \ldots, m$. We now bound $\|A\|_2$:

$$\begin{aligned} y^T A y &= \sum_{i,j=1,\ldots,m} y^{(i)T} A^{(i,j)} y^{(j)} = \\ &= \sum_{i,j=1,\ldots,m} \|y^{(i)}\| \|y^{(j)}\| \frac{y^{(i)T}}{\|y^{(i)}\|} A^{(i,j)} \frac{y^{(j)}}{\|y^{(j)}\|} \leq \\ &\leq \sum_{i,j=1,\ldots,m} \|y^{(i)}\| \|y^{(j)}\| = \\ &= e^T (zz^T) e \leq \|e\| \cdot \|zz^T\|_2 \cdot \|e\| = m. \end{aligned}$$

We used the fact that $\frac{y^{(i)T}}{\|y^{(i)}\|} A^{(i,j)} \frac{y^{(j)}}{\|y^{(j)}\|}$ is a correlation coefficient and thus bounded by 1. The upper bound follows:

$$\begin{aligned} \mathrm{Tr}\,(AX^*) &= \mathrm{Tr}\left(A \sum_{i=1,\ldots,N} \lambda_i x^{(i)} x^{(i)T} \right) = \\ &= \sum_{i=1,\ldots,N} \lambda_i x^{(i)T} A x^{(i)} \leq \sum_{i=1,\ldots,N} \lambda_i \cdot m = m^2. \end{aligned}$$

□

Corollary 5.7. The difference (duality gap) between the optimal objective value of the SDP and its the optimal objective value of its dual formulation (see Definition 3.12 and Definition 3.13) is zero.

Proof. This is a consequence of the boundedness of the optimization problem SDP and the fact that a strictly feasible point exists, for example:

$$X_0 := I \in S^N_{++}.$$

The statement follows from Lemma 5.6 and Theorem 3.4. □

Constant Relative Accuracy Guarantee There exists a lower bound on the ratio between the objective values of the original and the relaxed problem that is independent on the problem dimension. The bound is based on the following result from [32] (Theorem 2.1). The theorem refers to convex sets in Euclidean spaces [20] and concepts from general topology ([12]): closed sets, bounded sets and set interior.

Notation Let Square$(\cdot)$ denote componentwise squaring: if $y = \mathrm{Square}(x)$ then $y(i) = x(i)^2$ and let $\mathrm{diag}(X)$ denote the vector corresponding to the diagonal of X.

Definition 5.1. $K \subset \mathbb{R}^n$ is a *convex cone* if it is closed under conic combinations: $\alpha x + \beta y \in K$ for all $x \in K, y \in K, \alpha > 0, \beta > 0$. A cone is *pointed* if it contains the null vector (origin) 0_n.

Theorem 5.8 ([32, page 5]). *Let $A \in \mathbb{R}^{N \times N}$ be symmetric and let $\mathcal{F}$ be a set with the following properties:*

- $\mathcal{F} = \{v \in K : Bv = c\}$, *where K is a convex closed pointed cone in $\mathbb{R}^N$ with non-empty interior, $B \in \mathbb{R}^{k \times N}$, $c \neq 0_k$, and $\{v \in \mathrm{int}K : Bv = c\} \neq \emptyset$.*
- *$\mathcal{F}$ is closed, convex and bounded.*

- *There exists a componentwise strictly positive* $v \in \mathcal{F}$.

Denote:

$$\begin{aligned}
\phi^* &= \max\left\{x^T A x : \text{square}\,(x) \in \mathcal{F}\right\},\\
\phi_* &= \min\left\{x^T A x : \text{square}\,(x) \in \mathcal{F}\right\},\\
\psi^* &= \max\left\{\text{Tr}\,(AX) : \text{diag}\,(X) \in \mathcal{F}, X \in \mathbb{S}_+^N\right\},\\
\psi_* &= \min\left\{\text{Tr}\,(AX) : \text{diag}\,(X) \in \mathcal{F}, X \in \mathbb{S}_+^N\right\},\\
\psi\,(\alpha) &= \alpha\psi^* + (1-\alpha)\,\psi_*.
\end{aligned}$$

Then

$$\psi_* \leq \phi_* \leq \psi\left(1 - \frac{2}{\pi}\right) \leq \psi\left(\frac{2}{\pi}\right) \leq \phi^* \leq \psi^*. \tag{5.3}$$

Theorem 5.9. *Let* x^* *be the solution to the problem (QCQP2) and* X^* *be the solution to the problem (SDP). Let* $b = (n_1, \ldots, n_m)$ *denote the block structure where* $\sum_i n_i = N$. *Let* $\phi^* := \text{Tr}\left(A \cdot x^* \cdot x^{*T}\right)$ *and* $\psi^* := \text{Tr}\,(AX^*)$. *Then*

$$\frac{2}{\pi}\psi^* \leq \phi^* \leq \psi^*.$$

Proof. First note that it follows from $A \in \mathbb{S}_+^n$ that $\psi_* \geq 0$. This is a consequence of the fact that $\text{Tr}\,(AX) \geq 0$, for any $X \in \mathbb{S}_+^n$ (and therefore for the minimizer X_*). The positiveness of the trace can be deduced from: $\text{Tr}\,(AX) = \text{trace}\left(C_A C_A^T C_X C_X^T\right) = \text{trace}\left(C_X^T C_A C_A^T C_X\right) = \|C_A^T C_X\|_F^2 \geq 0$, where $A = C_A C_A^T$ and $X = C_X C_X^T$ are decompositions of A and X, which exist since symmetric matrices are diagonalizable and the eigenvalues are nonnegative.

We now show that the problems (QCQP2) and (SDP) can be reformulated so that the Theorem 5.8 applies.

Note that the feasible sets in (QCQP2) and (SDP) are defined in terms of equalities. Since the corresponding objective functions are convex and the feasible sets are closed and bounded in both cases, the corresponding optima must lie on their borders. For this reason, we can replace the equality constraints with inequality constraints without loss of generality: $x^{(i)T} x^{(i)} \leq 1$ in (QCQP2) and $\text{Tr}\,(B_i X) \leq 1$ in (SDP).

Next, we add redundant constraints to the two problems respectively: $\text{Square}\left(x^{(i)}\right) \geq 0, \forall i = 1, \ldots, m$ and $X\,(j,j) \geq 0, \forall j = 1, \ldots, N$.

Define $\widetilde{\mathcal{F}} = \left\{x \in \mathbb{R}^N | x^{(i)} \in \Delta^{n_i - 1}\right\}$, where

$$\Delta^k = \left\{x \in \mathbb{R}^{k+1} | x\,(i) \geq 0, \forall i \text{ and } \sum_i x\,(i) = 1\right\}.$$

$\widetilde{\mathcal{F}}$ is a product of standard simplices: $\widetilde{\mathcal{F}} = \prod_{i=1}^m \Delta^{n_i - 1}$. It follows that the set is closed, bounded and convex.

$\widetilde{\mathcal{F}}$ can be embedded in $\mathbb{R}^{N+1}$ in order to obtain the desired conic formulation. We now define the matrices referred to in the theorem:

$$\begin{aligned}
K &:= \{t \cdot \begin{bmatrix}1 & x^T\end{bmatrix}^T | t \geq 0, x \in \widetilde{\mathcal{F}}\},\\
B &:= \begin{bmatrix}1 & 0_N^T\end{bmatrix}^T, \quad c = 1, \quad \mathcal{F} := K \cap \{x | Bx = c\}.
\end{aligned}$$

Define $v = \begin{bmatrix}v_1^T \ldots v_m^T\end{bmatrix}^T$, where $v_i\,(j) = \frac{1}{n_i}$. The vector $\begin{bmatrix}1 & v^T\end{bmatrix}^T$ is strictly positive and lies in $\text{int}\,(K) \cap \left\{x \in \mathbb{R}^{N+1} | Bx = c\right\}$. Let $\widetilde{A} \in \mathbb{R}^{N+1}$ be defined as $\widetilde{A}\,(1,i) = 0$, $\widetilde{A}\,(i,1) = 0, \forall i$ and $\widetilde{A}\,(i,j) = A\,(i-1, j-1)\,, \forall i,j > 1$.

The optimization problem (QCQP2) is equivalent (with the same optimal objective value) to:

$$\begin{array}{ll} \max\limits_{x\in\mathbb{R}^{N+1}} & \mathrm{Tr}\left(\widetilde{A}xx^T\right) \\ \text{subject to} & \mathrm{Square}\,(x)\in\mathcal{F}. \end{array}$$

The optimization problem (SDP) is likewise equivalent to the problem:

$$\begin{array}{ll} \max\limits_{X\in\mathbb{S}_+^{N+1}} & \mathrm{Tr}\left(\widetilde{A}X\right) \\ \text{subject to} & \mathrm{diag}\,(X)\in\mathcal{F}. \end{array}$$

Using the definition of $\psi(\alpha)$ and the fact that $\psi_*\geq 0$ it follows that $\psi(\alpha)\geq\alpha\psi^*, \forall\alpha\geq 0$. Substituting $\alpha=\frac{2}{\pi}$ in (5.3) we get the desired result:

$$\frac{2}{\pi}\psi^*\leq\phi^*\leq\psi^*.$$

□

Observe that the bound above relates the optimization problems (QCQP) and (SDP) and not (4.1) with its SDP relaxation. Let $\widetilde{\phi}$ denote the optimum value of the objective function in (4.1) and let $\widetilde{\psi}$ denote the optimum value of the objective function of the corresponding SDP relaxation. It is easy to see that $2\cdot\widetilde{\phi}+m=\phi$ and $2\cdot\widetilde{\psi}+m=\psi$, which is a consequence of transformations of the original problems to their equivalent symmetric positive-definite problems. The $\frac{2}{\pi}$ constant relative accuracy bound becomes a bit weaker in terms of the original problem and its relaxation. This fact is stated in the following corollary.

Corollary 5.10. The optimum values of the objective function in (4.1) and its corresponding relaxation, denoted $\widetilde{\phi}$ and $\widetilde{\psi}$ respectively, are related by:

$$\widetilde{\phi}\geq\frac{2}{\pi}\widetilde{\psi}-\frac{(1-\frac{2}{\pi})m}{2}.$$

Improved Bound on the Relative Accuracy We can exploit the additional structure of the problem to obtain a slightly better bound. The result is based on applying Theorem 3.1 from [32]. Define

$$\omega(\beta):=\beta\arcsin(\beta)+\sqrt{1-\beta^2}.$$

The function $\omega(\beta)$ is increasing and convex with $\omega(0)=1$ and $\omega(1)=\frac{\pi}{2}$.

Theorem 5.11 ([32, page 7]). *Denote*

$$\begin{aligned} \tau^* &= \max\{\langle\mathrm{diag}(A),v\rangle\colon v\geq, v\in\mathcal{F}\}, \\ \tau_* &= \min\{\langle\mathrm{diag}(A),v\rangle\colon v\geq, v\in\mathcal{F}\}, \\ \beta^* &= \frac{\psi^*-\tau^*}{\psi^*-\psi_*}\in[0,1], \\ \beta_* &= \frac{\tau_*-\psi_*}{\psi^*-\psi_*}\in[0,1], \\ \alpha^* &= \max\{\frac{2}{\pi}\omega(\beta_*),1-\beta^*\}, \\ \alpha_* &= \min\{1-\frac{2}{\pi}\omega(\beta^*),\beta_*\}. \end{aligned}$$

The optimal values of the problems in Theorem 5.8 satisfy the following relations:

$$\psi^* \geq \phi^* \geq \psi(\alpha^*),$$
$$\psi_* \leq \phi_* \leq \psi(\alpha^*).$$

Applying the Theorem 5.11 we obtain the result:

$$\max\left\{\frac{2}{\pi}\omega\left(\frac{m}{\psi^*}\right), \frac{m}{\psi^*}\right\}\psi^* \leq \phi^* \leq \psi^*.$$

This results in a minor improvement of the default bound. For example, when $m = 3$ and the fact that $\frac{m}{\psi^*} \geq \frac{1}{3}$ we obtain the following:

$$\frac{2}{\pi}\psi^* \leq \frac{105}{100} \cdot \frac{2}{\pi}\psi^* \leq \phi^* \leq \psi^*.$$

5.4 Random Projections and Multivariate Regression

Although SDP problems are convex and admit polynomial time solutions, their applicability is difficult when the total number of features and the number of instances are large. For example, the applications of multiview learning to text analysis typically involve hundreds of thousands variables and training instances, while typical SDP solvers can find solutions to relaxed forms of QCQPs with up to a few thousand original variables.

To address this issue, we reduce the dimensionality of the feature vectors, resulting in tractable SDP problem dimensions.

One way to analyze a general dataset (single view) is to perform a singular value decomposition of the data matrix as we introduced in Section 3.2. A set of singular vectors corresponding to the largest singular values can be used to project the data into a lower-dimensional subspace. If computing the basis is too expensive, one can generate a random larger set of basis vectors that achieve similar reconstruction errors. However, this random projection basis is not informative in the same sense as the SVD basis is (directions extracted by SVD reflect which directions are prevalent in the data, as opposed to random directions).

When dealing with multiple views, a natural approach to dimensionality reduction is to compute the TSVD of the multiview stacked matrix (introduced in Chapter 2). Using random projections to approximate the computation then seems like a good strategy.

We experimentally observed that the number of random projections needed to approximate the TSVD decomposition thus construct a multi-view aligned basis is prohibitively large - while a relatively small number of random projections is needed to capture the variance in view, a large number of random projections is needed to approximate all the cross-covariance matrices simultaneously. Generating random subspaces with a fixed k sequentially over each view is problematic, since the probability of generating a subspace for the m-th view that is well correlated to the preceding views decreases as m increases.

Our approach is based on the following idea. Generate a set of random vectors for one view and use Canonical Correlation Analysis Regression (CCAR)[33] (a method similar to ridge regression) to find their representatives in the other views. Repeat the procedure for each of the remaining views to prevent bias to a single view. We hypothesize that restricting our search in the spaces spanned by the constructed bases still leads to good solutions, which we will demonstrate in Chapter 8. The procedure is detailed in Algorithm 5.1.

Let m be the number of vector spaces corresponding to different views and n_i the dimensionality of the i-th vector space. Let $X^{(i)} \in \mathbb{R}^{n_i \times N}$ represent the aligned data matrix for the i-th view.

Algorithm 5.1: Random projections basis generation

Input: $X^{(1)}, \dots X^{(m)}$, γ - the regularization coefficient, k - # of projections/block

for $i = 1$ to m **do**
  $P_{(i,i)} :=$ random $n_i \times k$ matrix with elements sampled *i.i.d.* from standard normal distribution.
  Re-scale each column of $P_{(i,i)}$ so that its norm is equal to $\sqrt{\frac{n_i}{k}}$.
  for $j = 1$ to m **do**
    if $j = i$ **then**
      continue
    end if
    $\alpha_{(i,j)} := \left((1-\gamma)\, X^{(j)} X^{(j)T} + \gamma I_j\right)^{-1}$
    $P_{(i,j)} := \alpha_{(i,j)} X^{(j)} X^{(i)T} P_{(i,i)}$, where I_j is the $n_j \times n_j$ identity matrix.
  end for
end for

Output: matrices $P_{(i,j)}$ for $i, j = 1, \dots, m$

The matrices $P_{(i,1)}, \dots, P_{(i,m)}$ form the bases of vector spaces corresponding to $X^{(1)}, \dots, X^{(m)}$. By using horizontal matrix concatenation we define the full basis for the i-th view:

$$P_i := \left[P_{(1,i)}, \dots, P_{(m,i)}\right]. \tag{5.4}$$

Remark. The Algorithm 5.1 is a heuristic dimensionality reduction step that tries to reduce the dimensionality of the space without destroying the pairwise correlation structure. In practice we observed that a purely random projection requires a prohibitively high dimension in order to capture the highly correlated subspaces. For this reason, starting with a random subspace of a particular view, we find its highly correlated ''counterparts" in the other views by performing independent pairwise analyses (similar to linear regression).

Remark on SDP optimization. Efficiently solving large SDP problems remains an active area of research. In this work we relied on an SDP solver implementation of a primal-dual interior point algorithm, see [34][35] for on overview and modern results. Many other approaches exist in the literature: a primal-dual combinatorial approach based on the idea of multiplicative updates [36], first-order methods with low memory requirements[37], spectral bundle methods [38] which can exploit specific problem structure, just to name a few.

Summary. This chapter discussed the optimization aspects of the SUMCOR generalization. We first presented that the problem is NP-hard in general, which motivated our results on global optimality bounds. We presented an SDP relaxation of the problem that yields bounds on the global solution as well as candidate solutions to the original problem under certain assumptions. Although the bound can be computed in polynomial time, applying it on a large and high dimensional dataset is a challenge. To address this we presented a heuristic dimensionality reduction step based on random projections and regression analysis, which significantly reduces the computational cost. This fact will be demonstrated in Chapter 8 on a high-dimensional dataset.

Chapter 6

Cross-Lingual Document Similarity

Document similarity is an important component in techniques from text mining and natural language processing. Many techniques use the similarity as a black box, e.g., a kernel in Support Vector Machines. Comparison of documents (or other types of text snippets) in a monolingual setting is a well-studied problem in the field of information retrieval [39]. We first formally introduce the problem followed by a description of our novel approach.

6.1 Problem Definition

We will first describe how documents are represented as vectors and how to compare documents in a mono-lingual setting. We then define a way to measure cross-lingual similarity which is natural for the models we consider.

Document Representation. The standard vector space model [39] represents documents as vectors, where each term corresponds to a word or a phrase in a fixed vocabulary. Formally, document d is represented by a vector $x \in \mathbb{R}^n$, where n corresponds to the size of the vocabulary, and vector elements x_k correspond to the number of times term k occurred in the document, also called *term frequency* or $TF_k(d)$.

For our document representation we applied a term re-weighting scheme that adjusts for the fact that some words occur more frequently in general. A term weight should correspond to the importance of the term for the given corpus. The common weighting scheme is called *Term Frequency Inverse Document Frequency* ($TFIDF$) weighting. An *Inverse Document Frequency* (IDF) weight for the dictionary term k is defined as $\log\left(\frac{N}{DF_k}\right)$, where DF_k is the number of documents in the corpus which contain term k and N is the total number of documents in the corpus. In the other part of our system, we computed TFIDF vectors on streams of news articles in multiple languages. There the IDF scores for each language changed dynamically - for each new document we computed the IDF of all news articles within a ten day window.

The $TFIDF$ weighted vector space model document representation corresponds to a map $\phi : \text{text} \to \mathbb{R}^n$ defined by:

$$\phi(d)_k = TF_k(d) \log\left(\frac{N}{DF_k}\right).$$

Monolingual similarity. A common way of computing similarity between documents is *cosine similarity*,

$$\text{sim}(d_1, d_2) = \frac{\langle \phi(d_1), \phi(d_2) \rangle}{\|\phi(d_1)\| \|\phi(d_2)\|},$$

where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are standard inner product and Euclidean norm respectively. When dealing with two or more languages, one could ignore the language information and build

a vector space using the union of tokens over the languages. A cosine similarity function in such a space can be useful to some extent, for example "Internet" or "Bowie" may appear both in Spanish and English texts and the presence of such terms in both an English and a Spanish document would contribute to their similarity. In general however, large parts of vocabularies may not intersect. This means that given a language pair, many words in both languages cannot contribute to the similarity score. Such cases can make the similarity function very insensitive to the data.

Cross-Lingual Similarity. Processing a multilingual dataset results in several vector spaces with varying dimensionality, one for each language. The dimensionality of the vector space corresponding to the i-th language is denoted by n_i and the vector space model mapping is denoted by $\phi_i : \text{text} \to \mathbb{R}^{n_i}$. The similarity between documents in language i and language j is defined as a bilinear operator represented as a matrix $S_{i,j} \in \mathbb{R}^{n_i \times n_j}$:

$$\text{sim}_{i,j}(d_1, d_2) = \frac{\langle \phi_i(d_1), S_{i,j}\phi_j(d_2) \rangle}{\|\phi_i(d_1)\| \|\phi_j(d_2)\|},$$

where d_1 and d_2 are documents written in the i-th and j-th language respectively. If the maximal singular value of $S_{i,j}$ is bounded by 1, then the similarity scores will lie on the interval $[-1, 1]$. We will provide an overview of the models in Section 6.2, present related work in Section 6.3 and then introduce additional notation in Section 6.4. Starting with Section 6.5 and ending with Section 6.8 we will describe some approaches to compute $S_{i,j}$ given training data.

6.2 Cross-Lingual Models

In this chapter we will describe several approaches to the problem of computing the multilingual similarities introduced in Section 6.1. We present four approaches: a simple approach based on k-means clustering in Section 6.5, a standard approach based on singular value decomposition in Section 6.6, a related approach called Canonical Correlation Analysis (CCA) in Section 6.7 and finally a new method, which is an extension of CCA to more than two languages in Section 6.8.

CCA can be used to find correlated patterns for a pair of languages, whereas the extended method optimizes a Sum of Squared Correlations (SSCOR) between several language pairs, which was introduced in [7]. The SSCOR problem is difficult to solve in our setting (hundreds of thousands of features, hundreds of thousands of examples). To tackle this, we propose a method which consists of two ingredients. The first one is based on an observation that certain datasets (such as Wikipedia) are biased towards one language (English for Wikipedia), which can be exploited to reformulate a difficult optimization problem as an eigenvector problem. The second ingredient is dimensionality reduction using CL-LSI, which makes the eigenvector problem computationally and numerically tractable.

We concentrate on approaches that are based on linear maps rather than alternatives, such as machine translation and probabilistic models, as discussed in the section on related work. We will start by introducing some notation.

The thesis so far focused on the SUMCOR problem formulation, where two extensions were proposed in Chapter 4 while Chapter 5 presented results on the problem complexity and global optimality bounds. In contrast, this chapter focuses on the SSCOR formulation. This is due to the fact that under an additional assumption (data is biased towards one view), the problem simplifies significantly (we could not derive a similar result for SUMCOR). At the end of the chapter we will comment further on the SUMCOR formulation in light of the additional assumption which made SSCOR computationally tractable.

6.3 Related Work

In this section, we describe previous work described in the literature. We have grouped the approaches to cross-lingual similarity computation as those that are based on: translation and dictionaries, probabilistic topic models, matrix factorization, monolingual models and neural network word embeddings. We also present some related work that also uses the Wikipedia as a language resource.

Translation and dictionary based. The most obvious way to compare documents written in different languages is to use machine translation and perform monolingual similarity, see [40] and [41] for several variations of translation based approaches. One can use free tools such as Moses [42] or translation services, such as Google Translate[1]. There are two issues with such approaches: they solve a harder problem than needs to be solved and they are less robust to training resource quality - large sets of translated sentences are typically needed. Training Moses for languages with scarce linguistic resources is thus problematic. The issue with using online services such as Google Translate is that the APIs are limited and not free. The operation efficiency and cost requirements make translation-based approaches less suited for our system. Closely related are works Cross-Lingual Vector Space Model (CL-VSM) [41] and the approach presented in [43] which both compare documents by using dictionaries, which in both cases are EuroVoc dictionaries [44]. The generality of such approaches is limited by the quality of available linguistic resources, which may be scarce or non-existent for certain language pairs.

Probabilistic topic models. There exist many variants to modelling documents in a language independent way by using probabilistic graphical models. The models include: Joint Probabilistic Latent Semantic Analysis (JPLSA) [45], Coupled Probabilistic LSA (CPLSA) [45], Probabilistic Cross-Lingual LSA (PCLLSA) [46] and Polylingual Topic Models (PLTM) [47] which is a Bayesian version of PCLLSA. The methods (except for CPLSA) describe the multilingual document collections as samples from generative probabilistic models, with variations on the assumptions on the model structure. The topics represent latent variables that are used to generate observed variables (words), a process specific to each language. The parameter estimation is posed as an inference problem which is typically intractable and one usually solves it using approximate techniques. Most variants of solutions are based on Gibbs sampling or Variational Inference, which are nontrivial to implement and may require an experienced practitioner to be applied. Furthermore, representing a new document as a mixture of topics is another potentially hard inference problem which must be solved.

Matrix factorization. Several matrix factorization based approaches exist in the literature. The models include: Non-negative matrix factorization based [48], Cross-Lingual Latent Semantic Indexing CL-LSI [49] and [40], Canonical Correlation Analysis (CCA) [1], Oriented Principal Component Analysis (OPCA) [45]. The quadratic time and space dependency of the OPCA method makes it impractical for large scale purposes. In addition, OPCA forces the vocabulary sizes for all languages to be the same, which is less intuitive. For our setting, the method in [48] has a prohibitively high computational cost when building models (it uses dense matrices whose dimensions are a product of the training set size and the vocabulary size). Our proposed approach combines CCA and CL-LSI. Another closely related method is Cross-Lingual Explicit Semantic Analysis (CL-ESA) [50], which uses Wikipedia (as do we in the current work) to compare documents. It can be interpreted as using the sample covariance matrix between features of two languages to define the dot product which is used to compute similarities. The authors of CL-ESA compare it to CL-LSI and find that CL-LSI can outperform CL-ESA in an information

[1] https://translate.google.com/

retrieval, but is costlier to optimize over a large corpus (CL-ESA requires no training). We find that the scalability argument does not apply in our case: based on advances in numerical linear algebra we can solve large CL-LSI problems (millions of documents as opposed to the 10,000 document limit reported in [50][51]). In addition, CL-ESA is less suited for computing similarities between two large monolingual streams. For example, each day we have to compute similarities between 500,000 English and 500,000 German news articles. Comparing each German news article with 500,000 English news articles is either prohibitively slow (involves projecting all English articles on Wikipedia) or consumes too much memory (involves storing the projected English articles, which for a Wikipedia of size 1,000,000 is a 500,000 by 1,000,000 non-sparse matrix). An interesting combination of dictionaries, CCA and LSI is presented in [52], where the authors show how to incorporate multilingual evidence into independently built (monolingual) vector spaces. In [53] the authors investigated how to optimally select an LSI training dataset for specific classification tasks and thus built domain specific common document representations, which is interesting but less relevant to our problem setting.

Monolingual. Related work also includes monolingual approaches that treat document written in different languages in a monolingual fashion. The intuition is that named entities (for example, "Bowie") and cognate words (for example, "tsunami") are written in the same or similar fashion in many languages. For example, the Cross-Language Character n-Gram Model (CL-CNG) [41] represents documents as bags of character n-grams. Another approach is to use language dependent keyword lists based on cognate words [43]. These approaches may be suitable for comparing documents written in languages that share a writing system, which does not apply to the case of global news tracking.

Word embeddings and neural networks. Many approaches in the recent literature focus on neural network models which construct hierarchical (distributed) representations of data. See [54],[55],[56] and [57] for an empirical comparison of several architectures over a variety of tasks. The approaches typically involve many tuning parameters (learning rates, batch sizes, neural network architectures) and the training tends to be computationally intensive.

Wikipedia based. Finally, there are some works that use Wikipedia to build models, but are not necessarily focused on cross-lingual similarity learning. For example [58] use Wikipedia for building cross-lingual dictionary that maps phrases from other languages to English wikipedia concepts. Wikipedia was also used to build an entity matching model [59], with the aim of aligning Wikipedia infoboxes. In [60] the authors investigated several cross-lingual measures between web documents (based on link analysis, as well as the distribution of cognate words, character n-grams etc.) and found that it especially valuable for under-resourced languages. Another paper [61] used Wikipedia concepts as an inverted index to build cross-lingual similarities. The approach represented words as distributions over Wikipedia concepts, which is closely related to [50].

6.4 Notation

The cross-lingual similarity models presented in this paper are based on comparable corpora. A *comparable corpus* is a collection of documents in multiple languages, with alignment between documents that are of the same topic, or even a rough translation of each other. Wikipedia is an example of a comparable corpus, where a specific entry can be described in multiple languages (e.g., "Berlin" is currently described in 222 languages). News articles represent another example, where the same event can be described by newspapers in several languages.

More formally, a *multilingual document* $d = (u_1, \ldots u_m)$ is a tuple of m documents

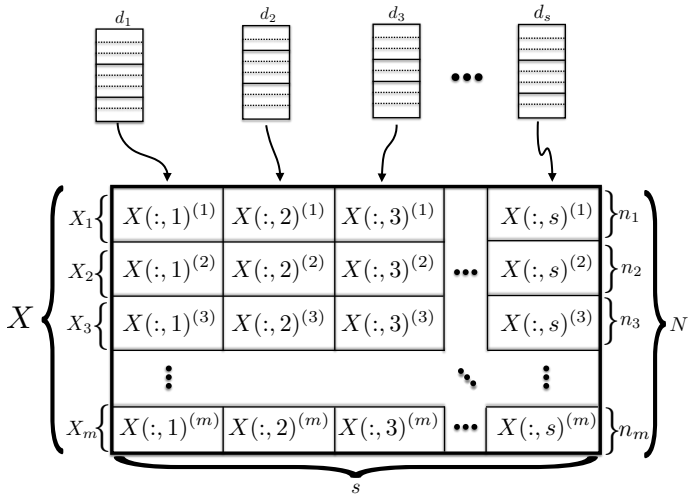


Figure 6.1: Multilingual corpora and their matrix representations using the vector space model.

on the same topic (comparable), where u_i is the document written in language i. Note that an individual document u_i can be an empty document (missing resource) and each d must contain at least **two nonempty documents**. This means that in our analysis we discard strictly monolingual documents for which no cross-lingual information is available. A comparable corpus $D = d_1, \ldots, d_s$ is a collection of s multilingual documents. By using the vector space model, we can represent D as a set of m multiview aligned matrices: $X_1, \ldots, X_m$, where $X_i \in \mathbb{R}^{n_i \times s}$ is the matrix corresponding to the language i and n_i is the vocabulary size of language i. Furthermore, let $X_i(:, \ell)$ denote the ℓ-th column of matrix X_i and the matrices respect the document alignment - the vector $X_i(:, \ell)$ corresponds to the TFIDF vector of the i-th component of multilingual document d_ℓ. We use N to denote the total row dimension of X, i.e., $N := \sum_{i=1}^{m} n_i$. See Figure 6.1 for an illustration of the introduced notation. The dimensions of each view form a block structure $b := (b_1, \ldots, b_m)$ and we will use the parenthesis notation $x^{(i)}$ to denote the i-th block according to that block structure.

We will now describe four models to cross-lingual similarity computation, where the first three are based on methods introduced in Chapter 3 and the fourth one represents an original contribution.

6.5 k-means

We have introduced the k-means algorithm in Section 3.1 and we now discuss how to apply it to build a cross-lingual similarity function. In a nutshell, we apply the standard k-means procedure to the stacked multiview matrix and interpret the centroids as an aligned basis, which is used to build the CL-similarity model; see Figure 6.2 for an illustration of the approach.

In order to apply the algorithm, we first merge all the term-document matrices into a single matrix X by stacking the individual term-document matrices (as seen in Figure 6.1):

$$X := \left[X_1^T, X_2^T, \cdots, X_m^T\right]^T,$$

such that the columns respect the alignment of the documents (here MATLAB notation for concatenating matrices is used). Therefore, each document is represented by a long vector indexed by the terms in all languages.

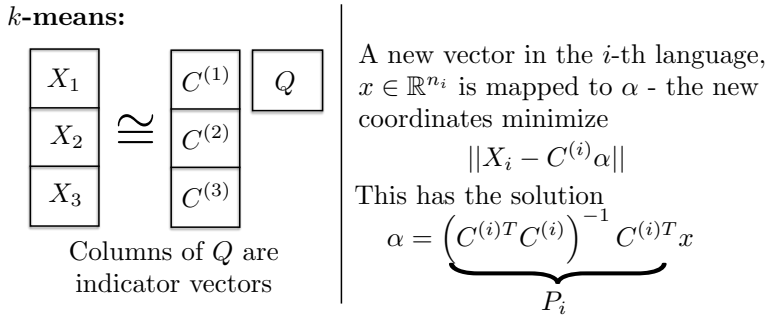


Figure 6.2: k-means algorithm and coordinate change.

We then run the k-means algorithm and obtain a centroid matrix $C \in \mathbb{R}^{N \times k}$, where the k columns represent centroid vectors. The centroid matrix can be split vertically into m blocks:

$$C = [C^{(1)T} \cdots C^{(m^T)}]^T,$$

according to the number of dimensions of each language, i.e., $C^{(i)} \in \mathbb{R}^{n_i \times k}$.

To reiterate, the matrices $C^{(i)}$ are computed using a multilingual corpus matrix X (based on Wikipedia for example).

To compute cross-lingual document similarities on new documents, we interpret each matrix $C^{(i)}$ as a vector space basis which can be used to map points in $\mathbb{R}^{n_i}$ into a k-dimensional space. Given a new vector $x \in \mathbb{R}^{n_i}$ we can express its coordinates as:

$$(C^{(i)T}C^{(i)})^{-1}C^{(i)T}x_i.$$

The resulting matrix for similarity computation between language i and language j is defined up to a scaling factor as:

$$C^{(i)}(C^{(i)T}C^{(i)})^{-1}(C^{(j)T}C^{(j)})^{-1}C^{(j)T}.$$

The matrix is a result of mapping documents in a language independent space using pseudo-inverses of the centroid matrices $P_i = (C^{(i)T}C^{(i)})^{-1}C^{(i)T}$ and then comparing them using the standard inner product, which results in the matrix $P_i^T P_j$. For the sake of presentation, we assumed that the centroid vectors are linearly independent. (An independent subspace could be obtained using an additional Gram-Schmidt step [11] on the matrix C, if this was not the case.)

6.6 Cross-Lingual Latent Semantic Indexing

Section 3.2 introduced two closely related approaches to pattern analysis based on spectral decompositions. This section summarizes how these approaches apply to cross-lingual document analysis. Truncated Singular value Decomposition (TSVD) applied to monolingual document analysis was introduced in [62] where it is referred to as Latent Semantic Indexing (LSI). An extension to cross-lingual document analysis was proposed in [49] and is referred to as Cross-Lingual Latent Semantic Indexing (CL-LSI). Now follows a description of a variation of CL-LSI, relevant to the thesis.

The method is based on computing a truncated singular value decomposition of multiview stacked matrix $X \approx USV^T$. See Figure 6.3 for the decomposition. Representing documents in "topic" coordinates is done in the same way as in the k-means case (see Figure 6.2), we will describe how to compute the coordinate change functions.

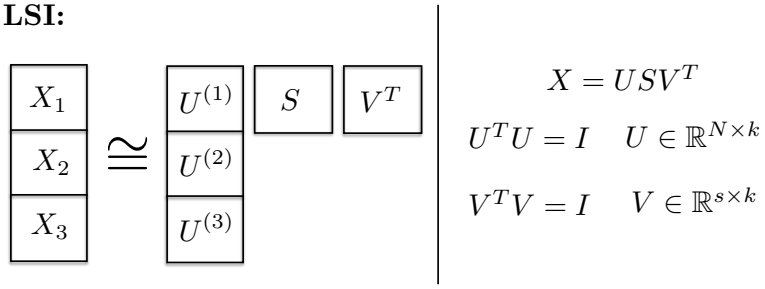


Figure 6.3: LSI multilingual corpus matrix decomposition.

The cross-lingual similarity functions are based on a rank-k truncated SVD: $X \approx U\Sigma V^T$, where $U \in \mathbb{R}^{N\times k}$ are basis vectors of interest and $\Sigma \in \mathbb{R}^{k\times k}$ is a truncated diagonal matrix of singular eigenvalues. An aligned basis is obtained by first splitting U vertically according to the number of dimensions of each language: $U = [U^{(i)T} \cdots U^{(m)T}]^T$. The standard CL-LSI approach is to use matrices U_i directly as maps to the common coordinate space - the cross-lingual similarity function between two documents x_i and x_j is given by $x_i^T U^{(i)} S^{-1} S^{-1} U^{(j)T} x_j$. We note that in general the blocks are not orthogonal and the norms of their columns may vary. Our modification to the standard approach is to treat $U^{(i)}$ as aligned bases, and use their pseudo-inverses to construct projection maps (exactly k-means centroids were treated). Assuming that the blocks $X^{(i)}$ have full rank (and thus so are $U^{(i)}$), the pseudo-inverses can be expressed as $P_i := (U^{(i)T}U^{(i)})^{-1}U^{(i)T}$. The matrices P_i are used to change the basis from the standard basis in $\mathbb{R}^{n_i}$ to the basis spanned by the columns of U_i.

Implementation note. Since the matrix X can be large we could use an iterative method like the Lanczos algorithm with re-orthogonalization [11] to find the left singular vectors (columns of U) corresponding to the largest singular values. In our experiments the Lanczos method converged slowly indicating problems with singular value separation. Moreover, the Lanczos method is hard to parallelize. Instead, we use a randomized version of the SVD [63] that can be viewed as a block Lanczos method. That enables us to use parallelization and speeds up the computation considerably.

To compute the matrices P_i we used the QR algorithm [11] to factorize $U^{(i)}$ as $U^{(i)} = Q_i R_i$, where $Q_i^T Q_i = I$ and R_i is a triangular matrix. P_i is then obtained by solving $R_i P_i = Q_i$.

6.7 Bi-Lingual Document Analysis CCA

CCA can be applied to bi-lingual document analysis given two languages. Let $X_1 \in \mathbb{R}^{p\times\ell}$ and $X_2 \in^{q\times\ell}$ denote the two multiview aligned matrices based on a bi-lingual aligned document collection vectorized using two vector space models. If $W^{(1)} \in \mathbb{R}^{p\times k}$ and $W^{(2)} \in \mathbb{R}^{q\times k}$ are solutions comprising of k canonical correlation vector pairs (corresponding to columns of $W^{(1)}$ and $W^{(2)}$) of the generalized eigenvalue problem in Equation 3.8, then the bi-lingual similarity function that we use is expressed as a cosine similarity after applying the canonical maps:

$$\text{sim}(x_1, x_2) \propto \frac{x_1^T W^{(1)} W^{(2)^T} x_2}{\|W^{(1)T} x_1\| \|W^{(2)T} x_2\|}.$$

The intuition behind using the matrices as projectors (even though their columns are not an orthonormal basis) is that the CCA is equivalent to minimizing the expected distance between $W^{(1)T}\mathcal{X}^{(1)}$ and $W^{(2)T}\mathcal{X}^{(2)}$, see Section 6.5 in [16].

6.8 Hub Language Based CCA Extension

Building cross-lingual similarity models based on comparable corpora is challenging for two main reasons. The first problem is related to missing alignment data: when a number of languages is large, the dataset of documents that cover all languages is small (or may even be empty). Even if only two languages are considered, the set of aligned documents can be small (an extreme example is given by the Piedmontese and Hindi Wikipedias where no inter-language links are available), in which case none of the methods presented so far are applicable.

The second challenge is scale - the data is high-dimensional (many languages with hundreds of thousands of features per language) and the number of multilingual documents may be large (over one million in case of Wikipedia). The optimization problem posed by CCA is not trivial to solve: the covariance matrices themselves are prohibitively large to fit in memory (even storing a 100,000 by 100,000 element matrix requires 80GB of memory) and iterative matrix-multiplication based approaches to solving generalized eigenvalue problems are required (the covariance matrices can be expressed as products of sparse matrices, which means we have fast matrix-vector multiplication).

We now describe an extension of CCA to more than two languages, which can be trained on large comparable corpora and can handle missing data. The extension we consider is based on a generalization of CCA to more than two views, introduced in [7], namely the Sum of Squared Correlations SSCOR, introduced in Equation SSCOR in Section 4.2, which we will re-state formally later in this section. Our approach exploits a certain characteristic of the data, namely the *hub language* characteristic (see below) in two ways: to reduce the dimensionality of the data and to simplify the optimization problem. We focus on the SSCOR formulation as opposed to the SCOR formulation which we explored in the previous chapters, since the Lagrangian problem is easier to solve (That is under the hub language assumption, which we present next).

Hub language characteristic. In the case of Wikipedia, we observed that even though the training resources are scarce for certain language pairs, there often exists indirect training data. By considering a third language, which has training data with both languages in the pair, we can use the composition of learned maps as a proxy. We refer to this third language as a hub language.

A *hub language* is a language with a high proportion of non-empty documents in $D = \{d_1, ..., d_\ell\}$. As we have mentioned, we only focus on multilingual documents that include at least two languages. The prototypical example in the case of Wikipedia is English. Our notion of the hub language could be interpreted in the following way. If a non-English Wikipedia page contains one or more links to variants of the page in other languages, English is very likely to be one of them. That makes English a hub language.

We use the following notation to define subsets of the multilingual comparable corpus: let $a(i, j)$ denote the index set of all multilingual documents with non-missing data for the i-th and j-th language:

$$a(i, j) = \{k \mid d_k = (u_1, ..., u_m), u_i \neq \emptyset, u_j \neq \emptyset\},$$

and let $a(i)$ denote the index set of all multilingual documents with non missing data for the i-th language.

We now describe a two step approach to building a cross-lingual similarity matrix. The first part is related to LSI and reduces the dimensionality of the data. The second step refines the linear mappings and optimizes the linear dependence between data.

Step 1: Hub language based dimensionality reduction. The first step in our method is to project $X_1, \ldots, X_m$ to lower-dimensional spaces without destroying the cross-lingual

structure. Treating the nonzero columns of X_i as observation vectors sampled from an underlying distribution $\mathcal{X}^{(i)} \in \mathbb{R}^{n_i}$, we can analyze the empirical cross-covariance matrices:

$$C^{(i,j)} = \frac{1}{|a(i,j)| - 1} \sum_{\ell \in a(i,j)} (X_i(:,\ell) - c_i) \cdot (X_j(:,\ell) - c_j)^T,$$

where $c_i = \frac{1}{|a_i|} \sum_{\ell \in a(i)} X_i(:,\ell)$. By finding low-rank approximations of $C^{(i,j)}$ we can identify the subspaces that are relevant for extracting linear patterns between $\mathcal{X}^{(i)}$ and $\mathcal{X}^{(j)}$. Let X_1 represent the hub language corpus matrix. The LSI approach to finding the subspaces is to perform the singular value decomposition on the full $N \times N$ covariance matrix composed of blocks $C^{(i,j)}$. If $|a(i,j)|$ is small for many language pairs (as it is in the case of Wikipedia), then many empirical estimates $C^{(i,j)}$ are unreliable, which can result in overfitting. For this reason, we perform the truncated singular value decomposition on the matrix $C = [C^{(1,2)} \dots C^{(1,m)}] \approx USV^T$, where $U \in \mathbb{R}^{n_1 \times k}, S \in \mathbb{R}^{k \times k}, V \in \mathbb{R}^{(\sum_{i=2}^{m} n_i) \times k}$. Define $T = [U^T V^T]^T \in \mathbb{R}^{N \times k}$, which is compatible with the block structure $b = (n_1, \dots, n_m)$. Note that the columns of $T^{(1)}$ are orthonormal, but that is not true in general for the other blocks. We proceed by reducing the dimensionality of each X_i by setting: $Y_i = T^{(i)T} \cdot X_i$, where $Y_i \in \mathbb{R}^{k \times s}$. To summarize, the first step reduces the dimensionality of the data and is based on CL-LSI, but optimizes only the hub language related cross-covariance blocks.

Step 2: Simplifying and solving SSCOR. The second step involves solving a generalized version of canonical correlation analysis on the matrices Y_i in order to find the mappings P_i. The approach is based on the sum of squares of correlations formulation by Kettenring [7], where we consider only correlations between pairs $(Y_1, Y_i), i > 1$ due to the hub language problem characteristic. We will present the original unconstrained optimization problem, then a constrained formulation based on the hub language problem characteristic. Then we will simplify the constraints and reformulate the problem as an eigenvalue problem by using Lagrange multipliers.

The original sum of squared correlations is formulated as an unconstrained problem:

$$\underset{w_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i<j}^{m} \rho(w_i^T Y_i, w_j^T Y_j)^2.$$

We solve a similar problem by restricting $i = 1$ and omitting the optimization over non-hub language pairs. Let $D_{i,i} \in \mathbb{R}^{k \times k}$ denote the empirical covariance of $\mathcal{Y}_i$ and $D_{i,j}$ denote the empirical cross-covariance computed based on $\mathcal{Y}_i$ and $\mathcal{Y}_j$. We solve the following constrained (unit variance constraints) optimization problem:

$$\underset{w_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i=2}^{m} \left(w_1^T D_{1,i} w_i\right)^2 \quad \text{subject to} \quad w_i^T D_{i,i} w_i = 1, \quad \forall i = 1, \dots, m. \tag{6.1}$$

The constraints $w_i^T D_{i,i} w_i$ can be simplified by using the Cholesky decomposition $D_{i,i} = K_i^T \cdot K_i$ and substitution: $y_i := K_i w_i$. By inverting the K_i matrices and defining $G_i := K_1^{-T} D_{1,i} K_i^{-1}$, the problem can be reformulated:

$$\underset{y_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i=2}^{m} \left(y_1^T G_i y_i\right)^2 \quad \text{subject to} \quad y_i^T y_i = 1, \quad \forall i = 1, \dots, m. \tag{6.2}$$

A necessary condition for optimality is that the derivatives of the Lagrangian vanish. The Lagrangian of (6.2) is expressed as:

$$L(y_1, \dots, y_m, \lambda_1, \dots, \lambda_m) = \sum_{i=2}^{m} \left(y_1^T G_i y_i\right)^2 + \sum_{i=1}^{m} \lambda_i \left(y_i^T y_i - 1\right).$$

Stationarity conditions give us:

$$\frac{\partial}{\partial y_1} L = 0 \quad \Rightarrow \quad \sum_{i=2}^{m} \left(y_1^T G_i y_i\right) G_i y_i + \lambda_1 y_1 = 0, \tag{6.3}$$

$$\frac{\partial}{\partial y_i} L = 0 \quad \Rightarrow \quad \left(y_1^T G_i y_i\right) G_i^T y_1 + \lambda_i y_i = 0, \ i > 1. \tag{6.4}$$

Multiplying each Equation (6.4) with y_i^T and applying the constraints, we can eliminate λ_i which gives us:

$$G_i^T y_1 = \left(y_1^T G_i y_i\right) y_i, \ i > 1. \tag{6.5}$$

Plugging this into Equation (6.3), we obtain an eigenvalue problem:

$$\left(\sum_{i=2}^{m} G_i G_i^T\right) y_1 + \lambda_1 y_1 = 0.$$

The eigenvectors of $\left(\sum_{i=2}^{m} G_i G_i^T\right)$ solve the problem for the first language. The solutions for y_i are obtained from (6.5): $y_i := \frac{G_i^T y_1}{\|G_i^T y_1\|}$. Note that the solution (6.1) can be recovered by: $w_i := K_i^{-1} y_i$. The original variables w are then expressed as:

$$Y_1 \ := \ \text{eigenvectors of } \sum_{i=2}^{m} G_i G_i^T, \tag{6.6}$$

$$W_1 \ = \ K_1^{-1} Y_1, \tag{6.7}$$

$$W_i \ = \ K_i^{-1} G_i^T Y_1 N, \tag{6.8}$$

where N is a diagonal matrix that normalizes $G_i^T Y_1$, with $N(j,j) := \frac{1}{\|G(_i Y_1(:,j)\|}$.

Remark. The technique is related to Generalization of Canonical Correlation Analysis (GCCA) [64], where an unknown group configuration variable is defined and the objective is to maximize the sum of squared correlations between the group variable and the others. The problem can be reformulated as an eigenvalue problem. The difference lies in the fact that we set the unknown group configuration variable as the hub language, which simplifies the solution. The complexity of our method is $O(k^3)$, where k is the reduced dimension from the LSI preprocessing step, whereas solving the GCCA method scales as $O(s^3)$, where s is the number of samples (see [65]). Another issue with GCCA is that it cannot be directly applied to the case of missing documents.

To summarize, we first reduced the dimensionality of our data to k-dimensional features and then found a new representation (via linear transformation) that maximizes directions of linear dependence between the languages. The final projections that enable mappings to a common space are defined as: $P_i(x) := W_i^T T^{(i)T} x$.

Remark on SUMCOR. The original sum of correlations is formulated as an unconstrained problem:

$$\underset{w_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i<j}^{m} \rho(w_i^T Y_i, w_j^T Y_j).$$

Under the hub language assumption, this problem is equivalent to (analogous to how we derived the formulation in Equation 6.2):

$$\begin{aligned} &\underset{y_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i=2}^{m} y_1^T G_i y_i \\ &\text{subject to} \quad y_i^T y_i = 1, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{6.9}$$

The Lagrangian of (6.9) is expressed as:

$$L(y_1, \ldots, y_m, \lambda_1, \ldots, \lambda_m) = \sum_{i=2}^{m} y_1^T G_i y_i + \sum_{i=1}^{m} \lambda_i \left(y_i^T y_i - 1 \right).$$

Stationarity conditions give us:

$$\begin{aligned} \frac{\partial}{\partial y_1} L = 0 \quad &\Rightarrow \quad \sum_{i=2}^{m} G_i y_i + \lambda_1 y_1 = 0, \\ \frac{\partial}{\partial y_i} L = 0 \quad &\Rightarrow \quad G_i^T y_1 + \lambda_i y_i = 0, \qquad\qquad i > 1. \end{aligned} \tag{6.10}$$

It follows that:

$$G_i y_i = -\frac{1}{\lambda_i} G_i G_i^T y_1,$$

and thus:

$$\sum_{i=2}^{m} -\frac{1}{\lambda_i} G_i G_i^T y_1 + \lambda_1 y_1 = 0.$$

Since the following equality can be derived:

$$\lambda_1 = \sum_{i=2}^{m} \lambda_i$$

the original problem reduces to

$$\sum_{i=2}^{m} -\frac{1}{\lambda_i} G_i G_i^T y_1 + \sum_{i=2}^{m} \lambda_i y_1 = 0,$$

but the solution to that problem does not seem obvious. This is in contrast to the SSCOR problem, where all λ_i as well as y_i could be eliminated simultaneously.

Summary. This chapter presented the problem of computing cross-lingual document similarities and then presented four computational approaches to solve the problem. The last one is an original contribution of the author. It is based on combining an efficient preprocessing step with a particular generalization of CCA (SSCOR), which under an additional assumption (hub language) can be solved efficiently (both the preprocessing step as well as the simplifying assumption are crucial for the approach). The next chapter will discuss an application of cross-lingual similarity learning to a large scale cross-lingual news analysis system.

Chapter 7

Applications to Cluster Linking

In online media streams – particularly news articles – there is often duplication of reporting, different viewpoints or opinions, all centering around a single event. Typically each event is covered by many articles and the question we address is how to find all the articles in different languages that are reporting on the same event. The current chapter describes an application of the cross-lingual similarity presented in Chapter 6 to cross-lingual cluster linking. The application is relevant for monitoring global news in multiple languages. Presented in the current chapter is our original approach to cross-lingual cluster linking, which was published in [9]. The main idea in our approach is to combine semantic information extraction with cross-lingual document analysis, which we proved to be effective in [10]. There we used a simpler set of features to decide which clusters to link and put greater emphasis on the manual evaluation of the quality.

To prepare the ground for the discussion of the cross-lingual approach, we will first describe how *events* are defined for our purposes and sketch a general approach to event tracking in a monolingual setting.

The term event is vague and ambiguous, but for the practical purposes, we define it as "any significant happening that is being reported about in the media". Examples of events would include shooting down of the Malaysia Airlines plane over Ukraine on July 18th, 2014 (see Figure 7.1) and HSBC's admittance of aiding their clients in tax evasion on February 9th, 2015. Events such as these are covered by many articles and the question is how to find all the articles in different languages that are describing a single event. Generally, events are more specific than general themes as the time component plays an important role – for example, the two wars in Iraq would be considered as separate events.

We take a pragmatic approach where the events are **identified** as the clusters discovered by stream clustering algorithm that is designed to cluster news articles together if their content is similar and they are published close in time.

The particular choice of a streaming clustering algorithm is not relevant to the discussion of our approach and we assume that given a monolingual news stream, we have at our disposal an online clustering component which assigns to each news article a cluster ID. In this work we used the clustering component of the Event Registry [66] and [67] system.

We now consider the case where we are dealing with several document streams, where each stream corresponds to a particular language. Running the monolingual streaming clustering on each stream results in clusters within streams that need to be matched across streams, which we will state formally.

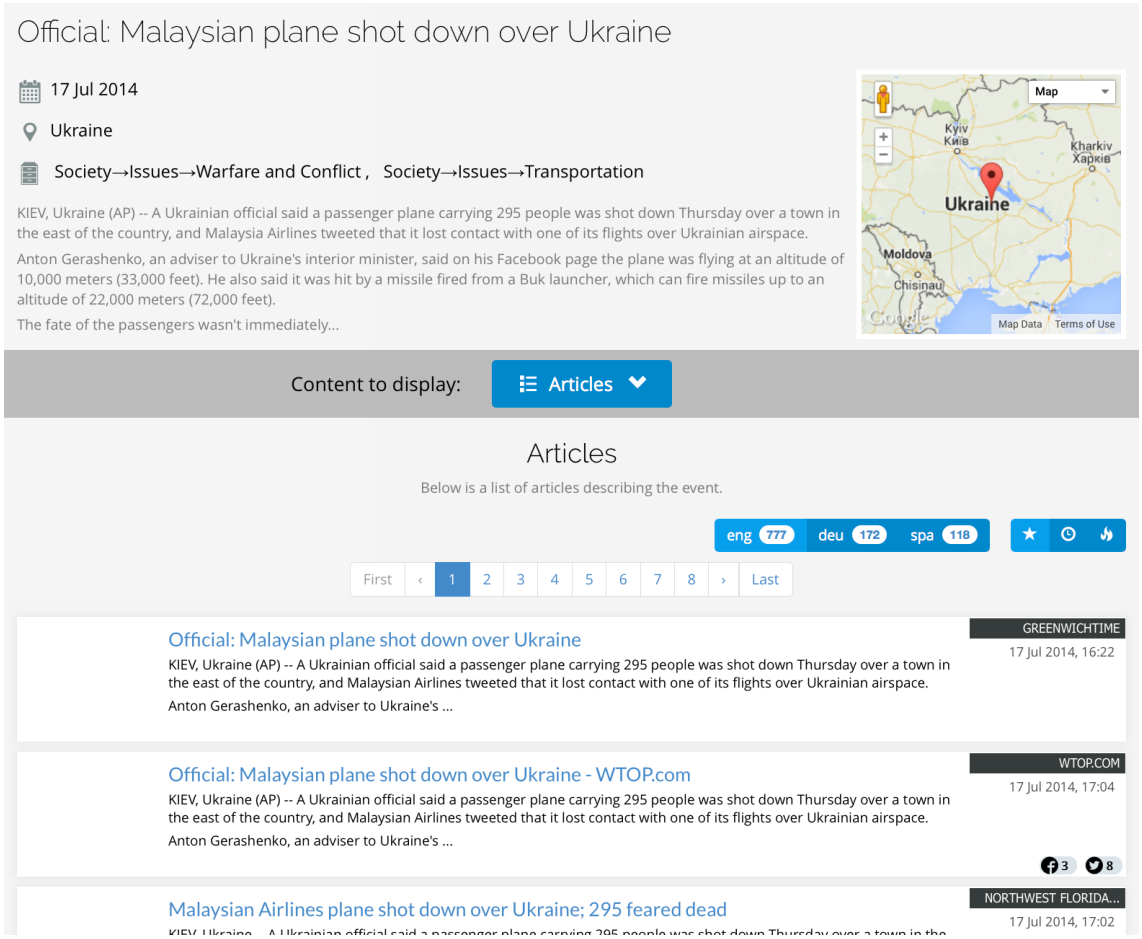


Figure 7.1: Events are represented by collections of articles about an event, in this case the Malaysian Airlines plane which was shot down over Ukraine. The results shown in the figure can be obtained using the query `http://eventregistry.org/event/997350#?lang=eng&tab=articles`. The content presented is part of the Event Registry system.

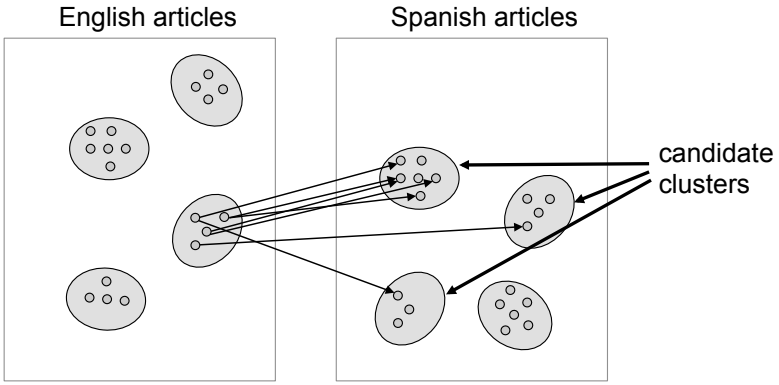


Figure 7.2: Clusters composed of English and Spanish news articles. Arrows link English articles with their Spanish k-nearest neighbor matches according to the cross-lingual similarity.

7.1 Problem Definition

The problem of cross-lingual event linking is to match monolingual clusters of news articles that are describing the same event across languages. For example, we want to match a cluster of Spanish news articles and a cluster of English news articles that both describe the same earthquake. We also refer to matched clusters as to being in *correspondence*. There are several possible ways of defining when events are in correspondence - our approach is to learn the correspondence based on a classification training dataset (the correspondence is thus captured by a dataset).

Each article $a \in A$ is written in a language ℓ, where $\ell \in L = \{\ell_1, \ell_2, ..., \ell_m\}$. For each language ℓ, we obtain a set of monolingual clusters C_ℓ. More precisely, the articles corresponding to each cluster $c \in C_\ell$ are written in the language ℓ. Given a pair of languages $\ell_a \in L$, $\ell_b \in L$ and $\ell_a \neq \ell_b$, we would like to identify all cluster pairs $(c_i, c_j) \in C_{\ell_a} \times C_{\ell_b}$ such that c_i and c_j describe the same event.

Matching of clusters is a *generalized matching* problem. We cannot assume that there is only one cluster per language per event, nor can we assume complete coverage – i.e., that there exists at least one cluster per event in every language. This is partly due to news coverage which might be more granular in some languages, partly due to noise and errors in the event detection process. This implies that we cannot make assumptions on the matching (e.g., one-to-one or complete matching) and excludes the use of standard weighted bipartite matching type of algorithms for this problem. An example is shown in Figure 7.2, where a cluster may contain articles which are closely matched with many clusters in a different language.

We also seek an algorithm which does not do exhaustive comparison of all clusters, since that can become prohibitively expensive when working in a real-time setting. More specifically, we wish to avoid testing cluster c_i with all the clusters from all the other languages. Performing exhaustive comparison would result in $O(|C|^2)$ tests, where $|C|$ is the number of all clusters (over all languages), which is not feasible when the number of clusters is on the order of tens of thousands. We address this by testing only clusters that are connected with at least one k-nearest neighbor (marked as *candidate clusters* in Figure 7.2).

Algorithm 7.1: Algorithm for identifying candidate clusters C that potentially correspond to c_i

input: test cluster c_i, a set of clusters C_ℓ for each language $\ell \in L$
output: a set of clusters C that potentially correspond to c_i
$C \leftarrow \{\}$;
for *article* $a_i \in c_i$ **do**
  for *language* $\ell \in L$ **do**
    /* use hub CCA to find 10 most similar articles to article a_i in language ℓ */
    $SimilarArticles = getCCASimilarArticles(a_i, \ell)$;
    for *article* $a_j \in SimilarArticles$ **do**
      /* find cluster c_j to which article a_j is assigned to */
      $c_j \leftarrow c$, such that $c \in C_\ell$ and $a_j \in c$;
      /* add cluster c_j to the set of candidates C */
      $C \leftarrow C \cup \{c_j\}$;
    end
  end
end

7.2 Algorithm

In order to identify clusters that correspond to cluster c_i, we have developed a novel two-stage approach. For a cluster c_i, we first efficiently identify a small set of candidate clusters and then find those clusters among the candidates, which correspond to c_i. An example is shown in Figure 7.2.

The details of the first step are described in Algorithm 7.1. The algorithm begins by individually inspecting each news article a_i in the cluster c_i. Using a chosen method for computing cross-lingual document similarity (see Chapter 6), it identifies the ten[1] most similar news articles to a_i in each language $\ell \in L$. For each similar article a_j, we identify its corresponding cluster c_j and add it to the set of candidates. The set of candidate clusters obtained in this way is several orders of magnitude smaller than the number of all clusters, and at most linear with respect to the number of news articles in cluster c_i. In practice, clusters contain highly related articles and as such similar articles from other languages mostly fall in only a few candidate clusters.

Although computed document similarities are approximate, our assumption is that articles in different languages describing the same event will generally have a higher similarity than articles about different events. While this assumption does not always hold, redundancy in the data should mitigate these false positives.

The second stage of the algorithm determines which (if any) of the candidate clusters correspond to c_i. We treat this task as a supervised learning problem. For each candidate cluster $c_j \in C$, we compute a vector of learning features that should be indicative of whether c_i and c_j are in correspondence or not and apply a binary classification model that predicts if the clusters are in correspondence or not. The classification algorithm that we used to train a model was a linear Support Vector Machine (SVM) method [16].

We use three groups of features to describe a cluster pair (c_i, c_j). The first group is based on **cross-lingual article links**, which are derived using cross-lingual similarity: each news article a_i is linked with its 10-nearest neighbors articles from all other languages (ten per each language). The group contains the following features:

[1] This parameter was manually selected based on the storage and speed requirements of a real system.

- `linkCount` is the number of times any of the news articles from c_j is among 10-nearest neighbors for articles from c_i. In other words, it is the number of times an article from c_i has a very similar article (i.e., is among 10 most similar) in c_j.

- `avgSimScore` is the average similarity score of the links, as identified for `linkCount`, between the two clusters.

Remark. The size of the nearest neighbor sets in our experiments is set to 10. The parameter was manually set with the goal of keeping computational and storage requirements low, while still capturing the cross-lingual correspondences.

The second group are **concept-related features**. Articles that are imported into Event Registry are annotated by disambiguating mentioned *entities* and *keywords* to the corresponding Wikipedia pages [68]. Whenever "David Bowie" is, for example, mentioned in the article, the article is annotated with a link to his Wikipedia page. In the same way, all mentions of entities (people, locations, organizations) and ordinary keywords (e.g., "bank", "tax", "ebola", "plane", "company") are annotated. Although the Spanish article about "Bowie" will be annotated with his Spanish version of the Wikipedia page, in many cases we can link the Wikipedia pages to their English versions. This can be done since Wikipedia itself provides information regarding which pages in different languages represent the same concept/entity. Using this approach, the word "avión" in a Spanish article will be annotated with the same concept as the word "plane" in an English article. Although the articles are in different languages, the annotations can therefore provide a language-independent vocabulary that can be used to compare articles/clusters. By analyzing all the articles in clusters c_i and c_j, we can identify the most relevant entities and keywords for each cluster. Additionally, we can also assign weights to the concepts based on how frequently they occur in the articles in the cluster. From the list of relevant concepts and corresponding weights, we consider the following features:

- `entityCosSim` is the cosine similarity between vectors of entities from clusters c_i and c_j.

- `keywordCosSim` is the cosine similarity between vectors of keywords from clusters c_i and c_j.

- `entityJaccardSim` is Jaccard similarity coefficient [69] between sets of entities from clusters c_i and c_j.

- `keywordJaccardSim` is Jaccard similarity coefficient between sets of keywords from clusters c_i and c_j.

The last group of features contains three **miscellaneous features** that seem discriminative but are unrelated to the previous two groups:

- `hasSameLocation` feature is a boolean variable that is true when the location of the event in both clusters is the same. The location of events is estimated by considering the locations mentioned in the articles that form a cluster and is provided by Event Registry.

- `timeDiff` is the absolute difference in hours between the two events. The publication time and date of the events is computed as the average publication time and date of all the articles and is provided by Event Registry.

- `sharedDates` is determined as the Jaccard similarity coefficient between sets of date mentions extracted from articles. We use extracted mentions of dates provided by Event Registry, which uses an extensive set of regular expressions to detect and normalize mentions of dates in different forms.

Summary. This chapter presented an application of the cross-lingual models on the task of linking news clusters across several languages. The approach is based on formulating the problem as a classification problem and using several sets of features, most notably: cross-lingual similarity based and semantic extraction (concepts-related) based features. In Chapter 8 we will presents some experiments and investigate the applicability of the proposed approach.

Chapter 8

Experiments

This chapter presents several experiments that explore how the theory relates to practice. The first set of experiments focuses on global optimality and convergence rates on synthetic data. We then move to cross-lingual document analysis and examine the performance of our approach on the task of information retrieval based on training sets with missing data. Finally we evaluate our approach to cross-lingual cluster linking.

8.1 Synthetic Experiments

We generated several multiview problem instances by varying the number of views and number of dimensions per view in order to compare the performance of local search methods and the proposed SDP relaxation. The goal of these experiments was to see under which conditions and how often do the global bounds provide useful information. The main observations are that the set of problems where the bounds are useful has a non-zero measure and that the difficulty of the optimization problem increases with the number of views and decreases with the number of dimensions per view.

8.1.1 Generating Synthetic Problem Instances

Let m denote the number of views (sets of variables) and n_i denote the dimensionality of i-th view and $N := \sum_i n_i$. In all cases, we used the same number of dimensions per view ($n_1 = n_2 = \cdots = n_m$). We used three different methods to generate random correlation matrices.

The first method, the **random Gram matrix** method (see [70], [71]), generates the correlation matrices by sampling N vectors $v_1, \ldots, v_n$ for an N-dimensional multivariate Gaussian distribution (centered at the origin, with an identity covariance matrix), normalizing them and computing the correlation matrix $C = [c_{i,j}]_{N \times N}$ as $c_{i,j} := v_i^T \cdot v_j$.

The second method, the **random spectrum** method, samples the eigenvalues $\lambda_1, \ldots, \lambda_N$ uniformly from a simplex ($\sum_{i=1}^{N} \lambda_i = N$) and generates a random correlation matrix with the prescribed eigenvalues (see [71]).

The final method, the **random 1-dim structure** method, generates a correlation matrix that has an approximately (due to noise) single-dimensional correlation structure. We generate a random m dimensional Gram matrix B, and insert it into an $N \times N$ identity matrix according to the block structure to obtain a matrix C_0. That is, we set $C_0(i,j) = \delta(i,j)$, where δ is the Kronecker delta. For $I, J = 1 \ldots, m$, we override the entries

$$C_0\left(1 + \sum_{i=1}^{I-1} n_i, 1 + \sum_{i=1}^{J-1} n_i\right) = B(I,J),$$

where we used 1-based indexing. We then generate a random Gram matrix $D \in \mathbb{R}^{N \times N}$ and compute the final correlation matrix as $C = (1 - \epsilon) C_0 + \epsilon D$. In our experiments, we set $\epsilon = 0.001$.

The purpose of using a random spectrum method is that as the dimensionality increases, random vectors tend to be orthogonal, hence the experiments based on random Gram matrices might be less informative. As the experiments show, the local method suffers when all $n_i = 1$ (an instance of a BQO problem). By using the approximately 1-dimensional correlation matrix sampling, we investigated how the problem behaves when $n_i > 1$.

In all cases, we perform a final step that involves computing the per-view Cholesky decompositions of variances and change of basis (as in QCQP).

8.1.2 Convergence of Horst's algorithm on Synthetic Data

We first present an empirical inquiry of the convergence rate of the Horst's algorithm. In Figure 8.1, we generated $1,000$ random instances of matrices A with block structure $b = (2, 2, 2, 2, 2)$, where we used the random Gram matrix method. For each matrix, we generated a starting point x_0 and ran the algorithm. The plot depicts the solution change rate on a logarithmic scale ($\log_{10} \frac{\|x_{old} - x\|}{\|x\|}$). We observe linear convergence over a wide range of rates of convergence (slopes of the lines). Figure 8.2 shows the convergence properties for a fixed matrix A with several random initial vectors x_0. The problem exhibits a global and a local solution. For 65% of the initial vectors, the global solution was reached (versus 35% for the local solution). Note that the global solution paths tend to converge faster (the average global solution path slope is -0.08, compared to -0.05 for the local solution paths).

8.1.3 SDP and Horst Solutions on Synthetic Problems

We also explored how the number of views and the dimensionality of the problems relates to the difficulty of finding globally optimal solutions, for all aforementioned methods of generating random problem instances. For each sampling scenario and each choice of m and n_i, we generated 100 experiments, and computed $1,000$ solutions based on Algorithm 4.1, the SDP solution (and respective global bounds), and examined the frequencies of the following events:

- a **relaxation gap** candidate detected (Tables 8.1, 8.2, 8.3 (a))
- **local convergence** detected (Tables 8.1, 8.2, 8.3 (b))
- when a local solution is worse than the SDP-based **lower bound** (Tables 8.1, 8.2, 8.3 (c)).

The possibility of a relaxation gap is detected when the best local solution is lower than 1% of the SDP bound. In this case, the event indicates only the possibility of relaxation gap – it might be the case that further local algorithm restarts would close the gap. Local convergence is detected when the objective value of two local solutions differs relatively by at least 10% and absolutely at least by 0.1 (both criteria must be satisfied simultaneously). Finally, the event of a local solution being below the SDP lower bound means that it is below $\frac{2}{\pi}$ of the optimal objective value of the SDP relaxation.

We find that regardless of the generation technique, the lower SDP bound is useful only when $n_i = 1$ (Table 8.1, 8.2, 8.3 (c)) and the results are similar for different choices of m. There are, however, rare instances (less than 0.1%) where the lower bound is useful for $n_i = 2$ and even rarer (less than 0.01%) for $n_i = 3$.

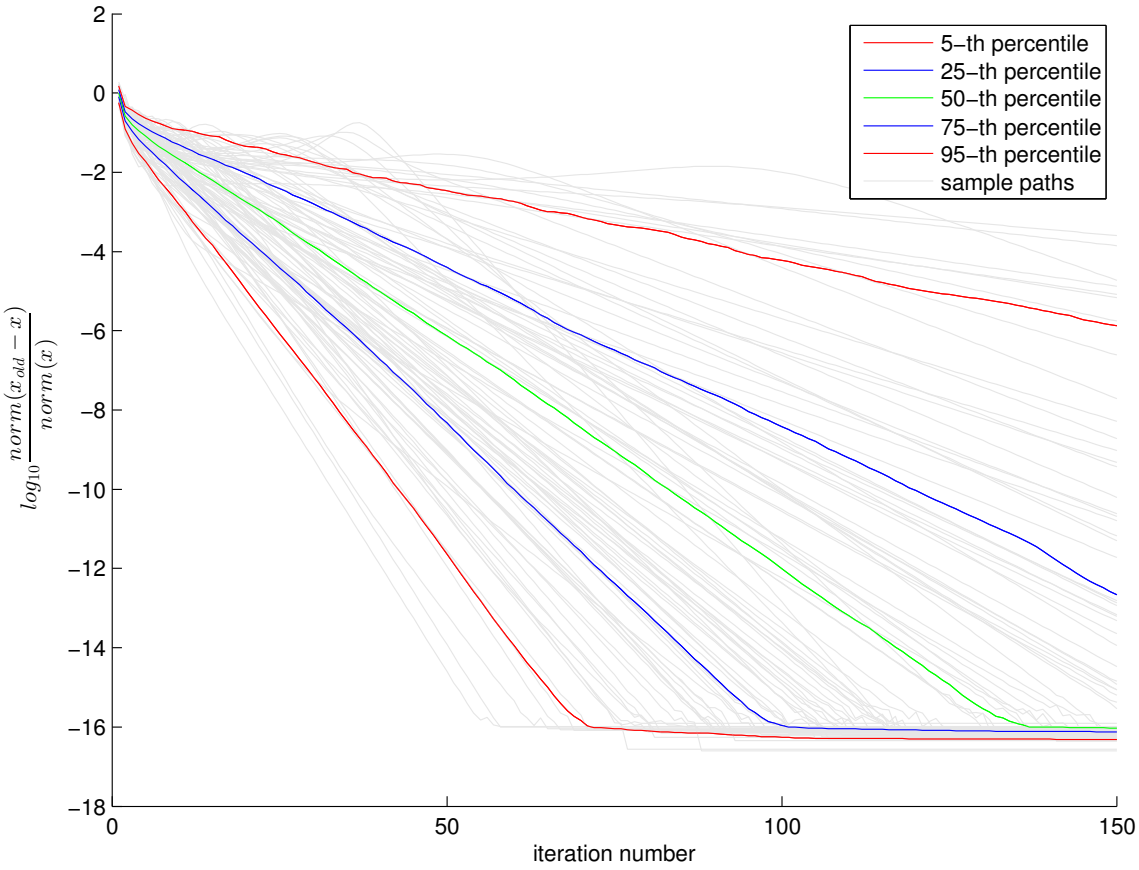


Figure 8.1: Convergence plot (1,000 random matrices A, one random x_0 per problem instance).

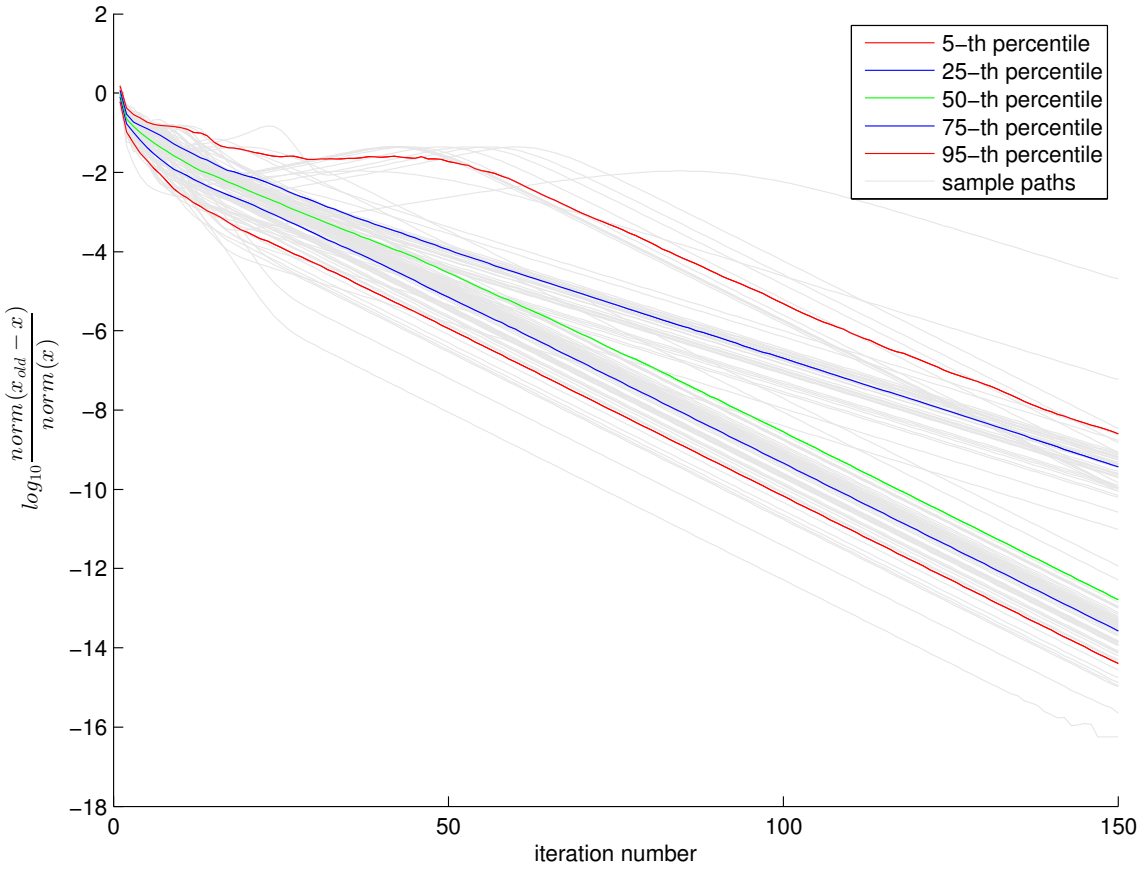


Figure 8.2: Convergence plot (single random A, $1,000$ random initial vectors x_0).

Table 8.1: Random Gram matrix.

(a) Possible relaxation gap.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	0%	5%	17%
$m = \mathbf{3}$	0%	0%	9%

(b) Local convergence.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	1%	5%	48%
$m = \mathbf{3}$	0%	1%	26%

(c) Local solution below lower SDP bound.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	0%	0%	14%
$m = \mathbf{3}$	0%	0%	12%

Table 8.2: Random spectrum sampling.

(a) Possible relaxation gap.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	0%	5%	36%
$m = \mathbf{3}$	0%	1%	20%

(b) Local convergence.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	1%	3%	50%
$m = \mathbf{3}$	0%	0%	31%

(c) Local solution below lower SDP bound.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	0%	0%	15%
$m = \mathbf{3}$	0%	0%	16%

The chance of local convergence increases as the number of views m increases which can be consistently observed for all choices of n_i and sampling strategies. Generating a generic problem where the local algorithm converges to a local solution is less likely as the dimensionality increases (Tables 8.1, 8.2).

In the case of noisy embeddings of a 1-dimensional correlation structures, the dependence on n_i behaves differently: the local convergence (see Table 8.3b) for the case $(m = 5, n_i = 3)$ is more likely than for the case $(m = 5, n_i = 2)$. This is unexpected as in the general case, increasing n_i reduces that chance of local convergence, see Table 8.2b, Table 8.1b.

The relationship between m and n_i and the possibility of a relaxation gap behaves similarly as local convergence - increasing m increases it and increasing n_i decreases it (Table 8.1a, Table 8.2a), except in the case of noisy 1-dim correlation structures, where we observe the same anomaly when $n_i = 2$ (Table 8.3a).

Therefore we have demonstrated that there exist sets of problems with nonzero measure where the SDP bounds give useful information.

8.2 Experiments on EuroParl Corpus

Applications of canonical correlation analysis on collections of documents include: dimensionality reduction, cross-lingual document retrieval and classification [18] and [72], multilingual topic extraction [73] and news bias detection [74]. In this section, we explore the behavior of Algorithm 4.1 with respect to the global bounds on real data, computed

Table 8.3: Random 1-dim structure sampling.

(a) Possible relaxation gap.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	24%	16%	23%
$m = \mathbf{3}$	7%	4%	7%

(b) Local convergence.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	9%	6%	51%
$m = \mathbf{3}$	0%	0%	31%

(c) Local solution below lower SDP bound.

	$n_i = \mathbf{3}$	$n_i = \mathbf{2}$	$n_i = \mathbf{1}$
$m = \mathbf{5}$	0%	0%	13%
$m = \mathbf{3}$	0%	0%	15%

based on Algorithm 5.1, introduced in Section 5.4. We start by describing the data and then describe a method to reduce the dimensionality of the data in order to apply the SDP bounds.

8.2.1 Dataset and preprocessing

The experiments were conducted on a subset of EuroParl, Release v3, [75], a multilingual parallel corpus, where our subset includes Danish, German, English, Spanish, Italian, Dutch, Portuguese and Swedish. We first removed all documents which had one translation or more missing. Documents (each document is a day of sessions of the parliament) were then arranged alphabetically and split into smaller documents, such that each speaker instance represented a separate document. Therefore, we ended up with $12,000$ documents per language. These roughly correspond to all speeches between $2/25/1999$ and $3/25/1999$. We then computed the bag of words (vector space) [39] model for each language, keeping unigrams, bigrams and trigrams that occurred more than thirty times. For example: "Mr", "President" and "Mr President" all occurred more than thirty times in the English part of the corpus and they each represent a dimension in the vector space.

This resulted in feature spaces with dimensionality ranging from $50,000$ (English) to $150,000$ (German). Finally, we computed the TF-IDF weighting and normalized every document for each language. Therefore, we obtained corpus matrices $X^{(i)}$ for each language, where each matrix has $12,000$ columns and the columns are aligned ($X^{(i)}(:,\ell)$ and $X^{(j)}(:,\ell)$ correspond to translations of the same text).

8.2.2 Local Versus Global Approaches

We now experimentally address two questions: does the random projection based approach introduced in Section 5.4 enable us to find *stable patterns* and how informative the SDP bounds are. Stable patterns are represented by highly correlated directions in both the training and test sets that were not a part of the parameter estimation procedure.

The experiments were conducted on the set of five EuroParl languages: English, Spanish, German, Italian and Dutch. We set $k = 10$ which corresponds to $n_i = 50$ dimensions per view, so the QCQP matrix will be of size 250×250. We randomly selected 5000 training documents and $1,000$ test documents. For a range of random projection regularization parameters γ, we computed the mappings P_i based on the train set, as defined in Equation 5.4. We then used the matrices P_i to reduce the dimensionality of the training and test sets. Then, for a range of QCQP regularization parameters κ, we set up the

Table 8.4: Train and test sum of correlation.

(a) Train set sum of correlations.

	γ =**0.1**	γ =**0.5**	γ =**0.9**	γ =**0.99**
κ =**0.01**	10.0	9.8	9.8	9.8
κ =**0.1**	10.0	9.8	9.8	9.8
κ =**0.5**	10.0	9.8	9.8	9.8
κ =**0.9**	10.0	9.8	9.8	9.8
κ =**0.99**	10.0	9.8	9.7	9.8

(b) Test set sum of correlations.

	γ =**0.1**	γ =**0.5**	γ =**0.9**	γ =**0.99**
κ =**0.01**	5.8	8.6	9.6	9.8
κ =**0.1**	6.2	8.6	9.6	9.8
κ =**0.5**	7.0	8.6	9.6	9.8
κ =**0.9**	7.4	8.8	9.6	9.8
κ =**0.99**	7.4	8.8	9.6	9.8

QCQP problem, computed 1,000 local solutions (by Horst algorithm) and solved the SDP relaxation. The whole procedure was repeated 10 times.

For each (γ, κ) pair, we measured the sum of correlations on the test and train sets. Table 8.4 shows the sums of correlations averaged over 10 experimental trials. The maximal possible sum of correlations for five datasets is $\binom{5}{2} = 10$. We observed that regularizing the whole optimization problem is not as vital as regularizing the construction of random projection vectors. This is intuitive, since finding the random projection vectors involves a regression in a high dimensional space, a harder estimation problem as opposed to solving a lower dimensional QCQP. The choice of $\gamma = 0.1$ lead to perfectly correlated solutions on the training set for all κ values. This turned out to be over-fitted when we evaluated the sum of correlations on the test set. Perfect correlations on the training set were not reproduced on the test set (the sum of correlations ranges between 5.8 and 7.4). Note that higher κ values improved the performance on the test set up to a certain level below 7.5. As we increased γ to 0.5, we saw a reduction in overfitting and with $\gamma = 0.9$ stable patterns were observed.

We conclude that using appropriate regularization parameters, we can reduce the dimensionality of the original QCQP problem and still find stable solutions.

The reduced dimensionality of the problem then enabled an investigation of the behavior of the SDP relaxation. For the SDP bounds, we observed behavior that was similar to the high-dimensional synthetic (generic) case. That is we found that the potential relaxation gap was very small and that the SDP and the Horst's algorithm yielded the same result. For this reason we omit the SDP results from Table 8.4.

8.3 Experiments on the Wikipedia Corpus

We will describe the main dataset for building cross-lingual models, which is based on Wikipedia, and then present two sets of experiments. The first set of experiments establishes that the hub based approach can deal with language pairs where little or no training data is available. The second set of experiments compares the main approaches that we

presented on the task of mate retrieval and the task of event linking. In the mate retrieval task we are given a test set of document pairs, where each pair consists of a document and its translation. Given a query document from the test set, the goal is to retrieve its translation in the other language, which is also referred to as its *mate* document. Finally, we examine how different choices of features impact the event linking performance.

8.3.1 Wikipedia Comparable Corpus

The following experiments are based on a large-scale real-world multilingual dataset extracted from Wikipedia by using inter-language links for alignment. Wikipedia is a large source of multilingual data that is especially important for the languages for which no translation tools, multilingual dictionaries (e.g. Eurovoc [44]), or strongly aligned multilingual corpora (e.g. Europarl [75]) are available. Documents in different languages are related with so-called *inter-language* links that can be found on the left of the Wikipedia page. The Wikipedia is constantly growing. At the time of writing the thesis there were twelve Wikipedias with more than one million articles, 52 with more than one hundred thousand articles, 129 with more than ten thousand articles, and 236 with more than one thousand articles.

We now present some details on how the dataset was processed to obtain the cross-lingual aligned corpus. Each Wikipedia page is embedded in the page tag. First, we ignored all the pages whose titles started with a Wikipedia namespace (which includes categories and discussion pages) and all redirection pages (but we stored the redirect link because inter-language links can point to redirection links also). We removed the markup of all the pages that we processed.

We constructed the inter-language link matrix using the previously stored redirection links and inter-language links. Inter-language links that pointed to the redirection links were replaced with the redirection target links. Since linking is not enforced to be consistent, we obtained a matrix M that was not symmetric. The existence of the inter-language link in one way (i.e., English to German) does not guarantee that there is an inter-language link in the reverse direction (German to English). To correct this we symmetrized the matrix M by computing $M + M^T$ and thus obtained an undirected graph. In the rare case that after symmetrization we had multiple links pointing from a document, we kept the first link. Our experiments were based on Wikipedia dumps available in 2013.

8.3.2 Experiments With Missing Alignment Data

In this subsection, we will present the empirical performance of hub CCA approach. We will demonstrate that this approach can be successfully applied even in the case of fully missing alignment information. To this purpose, we selected a subset of Wikipedia languages containing three major languages, English (4,212k articles)–*en* (hub language), Spanish (9,686k articles)–*es*, Russian (9,662k articles)–*ru*, and five minority (in terms of Wikipedia sizes) languages, Slovenian (136k articles)–*sl*, Piedmontese (59k articles)–*pms*, Waray-Waray (112k articles)–*war* (all with about 2 million native speakers), Creole (54k articles)–*ht* (8 million native speakers), and Hindi (97k articles)–*hi* (180 million native speakers). For preprocessing, we removed the documents that contained less than 20 different words (which are referred to as stubs[1]) and removed words that occurred in less than 50 documents as well as the top 100 most frequent words (in each language separately). We represented the documents as normalized TFIDF [39] weighted vectors. The

[1]Such documents are typically of low value as a linguistic resource. Examples include the titles of the columns in the table, remains of the parsing process, or Wikipedia articles with very little or no information contained in one or two sentences.

Table 8.5: Training – test sizes (in thousands). The first row represents the size of the training sets used to construct the mappings in low-dimensional language independent space using *en* as a hub. The diagonal elements represent the number of the unique training documents and test documents in each language.

	en	es	ru	sl	hi	war	ht	pms
en	671 - 4.6	463 - 4.3	369 - 3.2	50.3 - 2.0	14.4 - 2.8	8.58 - 2.4	17 - 2.3	16.6 - 2.7
es		463 - 4.3	187 - 2.9	28.2 - 2.0	8.7 - 2.5	6.9 - 2.4	13.2 - 2	13.8 - 2.6
ru			369 - 3.2	29.6 - 1.9	9.2 - 2.7	2.9 - 1.1	3.2 - 2.2	10.2 - 1.3
sl				50.3 - 2	3.8 - 1.6	1.2 - 0.99	0.95 - 1.2	1.8 - 1.0
hi					14.4 - 2.8	0.58 - 0.8	0.0 - 2.1	0.0 - 0.8
war						8.6 - 2.4	0.04 - 0.5	0.0 - 2.0
ht							17 - 2.3	0.0 - 0.4
pms								16.6 - 2.7

IDF scores were computed for each language based on its aligned documents with the English Wikipedia. The English language IDF scores were computed based on all English documents for which aligned Spanish documents existed.

The evaluation is based on splitting the data into training and test sets. We selected the test set documents as all multilingual documents with at least one nonempty alignment from the list: (*hi*, *ht*), (*hi*, *pms*), (*war*, *ht*), (*war*, *pms*), the remaining documents are used for training. The test set is suitable for testing the retrieval between language pairs with possibly empty alignment, since alignments with the hub language were available. In Table 8.5, we display the corresponding sizes of training and test sets for each language pair.

On the training set, we used the two step approach described in Section 6.8 to obtain the common document representation as a set of mappings P_i. The test set for each language pair, denoted by $test_{i,j} = \{(x_\ell, y_\ell) | \ell = 1 : n(i,j)\}$, consists of comparable document pairs (linked Wikipedia pages), where $n(i,j)$ is the test set size. We evaluated the representation by measuring the *mate retrieval* quality on the test sets as follows: for each ℓ, we ranked the projected documents $P_j(y_1), \ldots, P_j(y_{n(i,j)})$ according to their similarity with $P_i(x_\ell)$ and computed the rank of the mate document (aligned document) $r(\ell) = rank(P_j(y_\ell))$. The final retrieval score (between -100 and 100) was computed as: $\frac{100}{n(i,j)} \cdot \sum_{\ell=1}^{n(i,j)} \left(\frac{n(i,j) - r(\ell)}{n(i,j) - 1} - 0.5 \right)$. A score that is less than 0 means that the method performed worse than random retrieval and a score of 100 indicates perfect mate retrieval. The mate retrieval results are included in Table 8.6.

We observe that the method performs well on all pairs of languages, where at least 50,000 training documents are available(*en*, *es*, *ru*, *sl*). We note that taking $k = 500$ or $k = 1,000$ multilingual topics usually results in similar performance, with some notable exceptions: in the case of (*ht*, *war*) the additional topics result in an increase in performance, as opposed to (*ht*, *pms*) where performance drops, which suggests overfitting. The languages where the method performs poorly are *ht* and *war*, which can be explained by the quality of data (see Table 8.7 and explanation that follows). In case of *pms*, we demonstrate that solid performance can be achieved for language pairs (*pms*, *sl*) and (*pms*, *hi*), where only 2,000 training documents are shared between *pms* and *sl* and no training documents are available between *pms* and *hi*. Also observe that in the case of (*pms*, *ht*) the method still obtains a score of 62, even though training set intersection is zero and *ht* data is corrupted, which we will show in the next paragraph.

We further inspected the properties of the training sets by roughly estimating the fraction $\frac{rank(A)}{min(rows(A),\ cols(A))}$ for each English training matrix and its corresponding aligned

Table 8.6: Pairwise retrieval, 500 topics on the left – 1,000 topics on the right.

	en	es	ru	sl	hi	war	ht	pms
en		98 - 98	95 - 97	97 - 98	82 - 84	76 - 74	53 - 55	96 - 97
es	97 - 98		94 - 96	97 - 98	85 - 84	76 - 77	56 - 57	96 - 96
ru	96 - 97	94 - 95		97 - 97	81 - 82	73 - 74	55 - 56	96 - 96
sl	96 - 97	95 - 95	95 - 95		91 - 91	68 - 68	59 - 69	93 - 93
hi	81 - 82	82 - 81	80 - 80	91 - 91		68 - 67	50 - 55	87 - 86
war	68 - 63	71 - 68	72 - 71	68 - 68	66 - 62		28 - 48	24 - 21
ht	52 - 58	63 - 66	66 - 62	61 - 71	44 - 55	16 - 50		62 - 49
pms	95 - 96	96 - 96	94 - 94	93 - 93	85 - 85	23 - 26	66 - 54	

Table 8.7: Dimensionality drift. Each column corresponds to a pair of aligned corpus matrices between English and another language. The numbers represent the ratio between the numerical rank and the highest possible rank. For example, the column *en – ht* tells us that for the English-Creole pairwise-aligned corpus matrix pair, the English counterpart has full rank, but the Creole counterpart is far having full rank.

en – es	en – ru	en – sl	en – hi	en – war	en – ht	en – pms
0.81 – 0.89	0.8 – 0.89	0.98 – 0.96	1 – 1	0.74 – 0.56	1 – 0.22	0.89 – 0.38

matrix of the other language, where $rows(A)$ and $cols(A)$ denote the number of rows and columns respectively. The denominator represents the theoretically highest possible rank the matrix A could have. Ideally, these two fractions should be approximately the same - both aligned spaces should have reasonably similar dimensionality. We display these numbers as pairs in Table 8.7.

It is clear that in the case of the Creole language only at most 22% documents are unique and suitable for the training. Though we removed the stub documents, many of the remaining documents are nearly indistinguishable, as the quality of some smaller Wikipedias is low. This was confirmed for the Creole, Waray-Waray, and Piedmontese languages by manual inspection. The low quality documents correspond to templates about the year, person, town, etc. and contain very few unique words.

There is also a problem with the quality of the test data. For example, if we look at the test pair (*war*, *ht*) only 386/534 Waray-Waray test documents are unique but on the other side almost all Creole test documents (523/534) are unique. This indicates a poor alignment which leads to poor performance.

8.3.3 Evaluation Of Cross-Lingual Event Linking

We now turn our attention to the task of cross-lingual event linking and the evaluation of the novel approach presented in Chapter 7. In order to determine how accurately we can predict cluster correspondence, we performed two experiments in a multilingual setting using English, German and Spanish languages for which we had labelled data to evaluate the linking performance. In the first experiment, we tested how well the individual approaches for cross-lingual article linking perform when used for linking the clusters about the same event. In the second experiment we tested how accurate the prediction model is when trained on different subsets of learning features. To evaluate the prediction accuracy

for a given dataset we used 10-fold cross validation.

We created a manually labelled dataset in order to evaluate cross-lingual event linking using two human annotators. The annotators were provided with an interface listing the articles, their content from and top concepts for a pair of clusters and their task was to determine if the clusters were in correspondence or not (i.e., discuss same event). To obtain a pair of clusters (c_i, c_j) to annotate, we first randomly chose a cluster c_i, used Algorithm 7.1 to compute a set of potentially corresponding clusters C and randomly chose a cluster $c_j \in C$. The dataset provided by the annotators contains 808 examples, of which 402 are cluster pairs in correspondence and 406 are not. Clusters in each learning example are either in English, Spanish or German. Although Event Registry imports articles in other languages as well, we restricted our experiments to these three languages. We chose only these three languages since they have a very large number of articles and clusters per day which makes the cluster linking problem hard due to large number of possible links.

In Chapter 3 and Chapter 6, we described the three main algorithms for identifying similar articles in different languages. These algorithms were k-means, LSI and hub CCA. As a training set, we used common Wikipedia alignment for all three languages. To test which of these algorithms performed best, we made the following test. For each of the three algorithms, we analyzed all articles in Event Registry and for each article computed the most similar articles in other languages. To test how informative the identified similar articles are for cluster linking we then trained three classifiers as described in Section 7.2 – one for each algorithm. Each classifier was allowed to use as learning features **only** the cross-lingual article linking features for which values are determined based on the selected algorithm (k-means, LSI and hub CCA). The results of the trained models are shown in Table 8.8. We also show how the number of topics (the dimensions of the latent space) influences the quality, except in the case of the k-means algorithm, where only the performance on 500 topic vectors is reported, due to higher computational cost.

We observe that, for the task of cluster linking, LSI and hub CCA perform comparably and both outperform k-means.

We also compared the proposed approaches on the task of Wikipedia mate retrieval (the same task as in Section 8.3.2). In our evaluation we used a measure referred to as the Average Mean Reciprocal Rank (AMRR) [76]. Given a collection of Q queries $q_1, \ldots, q_Q$ and their corresponding mate documents $m_1, \ldots, m_Q$, the Mean Reciprocal Rank (MRR) is expressed as:

$$MRR = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{rank(i)},$$

where $rank(i)$ refers to the rank position of the mate document m_i with respect to the i-th query q_i. The rank is obtained by sorting the similarity scores $\{\text{sim}(q_i, r_j)\}_{j=1}^{Q}$ in descending order. Since we are dealing with more than two languages, we averaged the MRR scores over several language pairs and this aggregate score is referred to as AMRR.

We computed the AMRR performance of the different approaches on the Wikipedia data by holding out $15,000$ aligned test documents and using $300,000$ aligned documents as the training set.

Figure 8.3 shows AMRR score as the function of the number of feature vectors. It is clear that hub CCA outperforms LSI approach and k-means lags far behind when testing on Wikipedia data. The hub CCA approach with 500 topic vectors manages to perform comparably to the LSI-based approach with $1,000$ topic vectors, which shows that the CCA method can improve both model memory footprint as well as similarity computation time.

Furthermore, we inspected how the number of topics influences the accuracy of cluster linking. As we can see from Table 8.8 choosing a number of features larger than 500

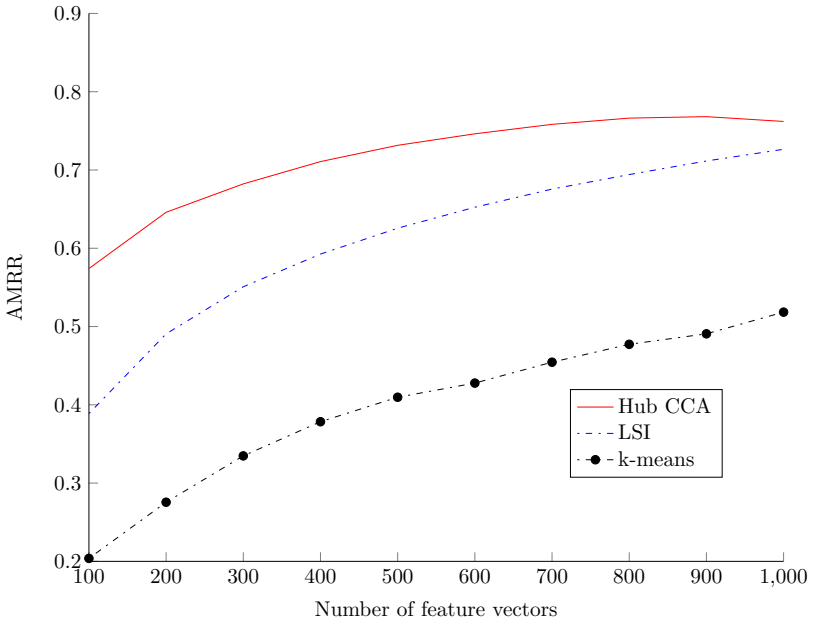


Figure 8.3: Average of mean reciprocal ranks.

barely affects linking performance, which is in contrast with the fact that additional topics helped to improve AMMR, see Figure 8.3. Such differences may have arisen due to different domains of training and testing (Wikipedia pages versus news articles).

We also analyzed how cluster size influences the accuracy of cluster linking. We would expect that if the tested pair of clusters has a larger number of articles, then the classifier should be able to more accurately predict whether the clusters should be linked or not. The reasoning is that the large clusters would provide more document linking information (more articles mean more links to other similar articles) as well as more accurately aggregated semantic information. In the case of smaller clusters, the errors of the similarity models have greater impact which should decrease the performance of the classifier, too. To validate this hypothesis we have split the learning examples into two datasets – one containing cluster pairs where the combined number of articles from both clusters is below 20 and one dataset where the combined number is 20 or more. The results of the experiment can be seen in Table 8.9. As it can be seen, the results confirm our expectations: for smaller clusters it is indeed harder to correctly predict if the cluster pair should be merged or not.

The hub CCA attains higher precision and classification accuracy on the task of linking small cluster pairs than the other methods, while LSI is slightly better on linking large cluster pairs. The gain in precision of LSI over hub CCA on linking large clusters is much smaller than the gain in precision of hub CCA over LSI on linking small clusters. For that reason we decided to use hub CCA as the similarity computation component in our system.

Remark. Computing hub CCA involves LSI as a preprocessing step which is followed by a low-dimensional eigen-decomposition, which is negligible when compared to the first step. The computational complexity of k-means is NP-hard, but given a fixed number of iterations, its cost is dominated by inner product computations (centroids-centroids and dataset-centroids), which is true for LSI as well when one uses a random-projection based approach [63].

In the second experiment, we evaluated how relevant individual groups of features are

Table 8.8: Accuracy of cluster linking with 500/800/1,000 topic vectors obtained from different cross-lingual similarity algorithms. The table shows for each of the algorithms the obtained classification accuracy, precision and recall. Some results related to k-means are omitted, since the computation took too long and the algorithm's performance was consistently lower than the other methods considered.

Models	Accuracy %	Precision %	Recall %	F_1 %
hub CCA	78.2/79.6/80.3	76.3/78.0/80.5	81.6/82.1/79.9	78.9/80.0/80.2
LSI	78.9/78.7/80.6	76.8/77.0/78.7	83.3/80.6/83.6	79.9/78.8/81.1
k-means	73.9/-/-	69.5/-/-	84.6/-/-	76.3/-/-

Table 8.9: Accuracy of cluster linking using 500 topic vectors on two datasets containing large (left number) and small (right number) clusters. The dataset with small clusters contained the subset of learning examples in which the combined number of articles from both clusters of the cluster pair were below 20. The remaining learning examples were put into the dataset of large clusters.

Models	Accuracy %	Precision %	Recall %	F_1 %
hub CCA	81.2 - 77.8	80.5 - 74.5	91.3 - 57.5	85.6 - 64.9
LSI	82.8 - 76.4	81.3 - 70.9	93.1 - 57.5	86.8 - 63.5
k-means	75.5 - 71.2	72.8 - 70.8	95.3 - 36.2	82.5 - 47.9

to correctly determine cluster correspondence. For this purpose, we computed accuracy using individual groups of features, as well as using different combinations of groups. Since hub CCA had the best performance of the three algorithms, we used it to compute the values of the cross-lingual article linking features. The results of the evaluation are shown in Table 8.10. We can see that using a single group of features, the highest prediction accuracy can be achieved using concept-related features. The classification accuracy in this case is 88.5%. By additionally including also the cross-lingual article linking features, the classification accuracy rises slightly to 89.4%. Using all three groups of features, the achieved accuracy is 89.2%.

To test if the accuracy of the predictions is language dependent we have also performed the evaluations separately on individual language pairs. For this experiment we have split the annotated learning examples into three datasets, where each dataset contained only examples for one language pair. When training the classifier all three groups of features were available. The results are shown in Table 8.11. We can see that the performance of cluster linking on the English-German dataset is the highest in terms of accuracy, precision, recall and F_1. The performance on the English-Spanish dataset is comparable to the performance on the English-German dataset, where the former achieves higher recall (and slightly higher F_1 score), while the latter achieves higher precision. A possible explanation of these results is that the higher quantity and quality of English-German language resources leads to a more accurate cross-lingual article similarity measure as well as to a more extensive semantic annotation of the articles.

Based on the performed experiments, we can make the following conclusions. The cross-lingual similarity algorithms provide valuable information that can be used to iden-

Table 8.10: The accuracy of the classifier for story linking using different sets of learning features. See Section 7.2 for a description of different feature sets that were used in classification: **Concepts** for concept-related features, **Misc** for miscellaneous features and **Hub CCA** for features based cross-lingual article links using hub CCA.

Features	Accuracy %	Precision %	Recall %	F_1 %
Hub CCA	78.3 ± 5.9	78.2 ± 7.0	78.9 ± 5.2	78.4 ± 5.5
Concepts	88.5 ± 2.7	88.6 ± 4.8	88.6 ± 2.2	88.5 ± 2.4
Misc	54.8 ± 6.7	61.8 ± 16.5	58.2 ± 30.2	52.4 ± 13.0
Hub CCA + Concepts	89.4 ± 2.5	89.4 ± 4.6	89.6 ± 2.4	89.4 ± 2.3
Hub CCA + Misc	78.8 ± 5.0	78.9 ± 7.1	79.4 ± 4.6	79.0 ± 4.5
Concepts + Misc	88.7 ± 2.6	88.8 ± 4.6	88.8 ± 2.2	88.7 ± 2.3
All	89.2 ± 2.6	88.8 ± 4.9	90.1 ± 1.9	89.3 ± 2.3

Table 8.11: The accuracy of the classifier for story linking on training data for each language pair separately using all learning features.

Language pair	Accuracy %	Precision %	Recall %	F_1 %
en, de	91.8 ± 5.5	91.7 ± 6.3	93.7 ± 6.3	92.5 ± 5.1
en, es	87.7 ± 5.4	87.7 ± 7.4	88.5 ± 9.8	87.6 ± 5.9
es, de	88.6 ± 4.3	89.7 ± 9.1	84.3 ± 11.9	85.9 ± 6.0

tify clusters that describe the same event in different languages. For the task of cluster linking, the cross-lingual article linking features are however significantly less informative compared to the concept-related features that are extracted from the semantic annotations. Nevertheless, the cross-lingual article similarity features are very important for two reasons. The first is that they allow us to identify for a given cluster a limited set of candidate clusters that are potentially in correspondence. This is a very important feature since it reduces the search space by several orders of magnitude. The second reason these features are important is that concept annotations are not available for all articles as the annotation of news articles is computationally intensive and can only be done for a subset of collected articles. The prediction accuracies for individual language pairs are comparable although it seems that the achievable accuracy correlates with the amount of available language resources.

8.3.4 Remarks on the Scalability of the Implementation

One of the main advantages of our two step approach to cross-lingual cluster linking is that it is highly scalable. It is fast, very robust to the quality of training data, easily extendable, simple to implement and has relatively small hardware requirements. The similarity pipeline is the most computationally intensive part and currently runs on a machine with two Intel Xeon E5-2667 v2, 3.30GHz processors with 256GB of RAM. This is sufficient to do similarity computation over a large number of languages if needed. It currently uses Wikipedia as a freely available knowledge base and experiments show that the similarity pipeline dramatically reduces the search space when linking clusters.

Currently, we compute similarities over 24 languages with tags: *eng*, *spa*, *deu*, *zho*, *ita*,

fra, *rus*, *swe*, *nld*, *tur*, *jpn*, *por*, *ara*, *fin*, *ron*, *kor*, *hrv*, *tam*, *hun*, *slv*, *pol*, *srp*, *cat*, *ukr* but we support any language from the top 100 Wikipedia languages. Our data streams come from the service called *Newsfeed* (`http://newsfeed.ijs.si/`) which provides 430k unique articles per day. Our system currently computes 2 million similarities per second, that means that we compute $16 \cdot 10^{10}$ similarities per day. We store one day buffer for each language which requires 1.5 GB of memory with documents stored as 500-dimensional vectors. We note that the time complexity of the similarity computations scales linearly with dimension of the feature space and does not depend on the number of languages. For each article, we compute the top 10 most similar ones in every other language.

For all linear algebra matrix and vector operations, we use high performance numerical linear algebra libraries as BLAS, OPENBLAS and Intel MKL, which currently allows us to process more than one million articles per day. In our current implementation, we use the variation of the hub approach. Our projector matrices are of size $500 \times 300,000$, so every projector takes about 1.1 GB of RAM. Moreover, we need proxy matrices of size 500×500 for every language pair. That is 0.5 GB for 24 languages and 9.2 GB for 100 languages. Altogether we need around 135 GB of RAM for the system with 100 languages. Usage of proxy matrices enables the projection of all input documents in the common space and handling language pairs with missing or low alignment. That enables us to do block-wise similarity computations further improving system efficiency. Our code can therefore be easily parallelized using matrix multiplication rather than performing more matrix - vector multiplications. This speeds up our code roughly by a factor of 4. In this way, we obtain some caching gains and ability to use vectorization. Our system is also easily extendable. Adding a new language requires the computation of a projector matrix and proxy matrices with all other already available languages.

8.3.5 Remarks on the Reproducibility of Experiments

We have made both the code and data that were used in the experiments publicly available at `https://github.com/rupnikj/jair_paper.git`. The manually labelled dataset used in the evaluation of event linking is available at in the "dataset" subfolder of the github repository. The included archive contains two folders: "positive" and "negative", where the first folder includes examples of cluster pairs in two languages that represent the same event and the second folder contains pairs of clusters in two languages that do not represent different events. Each example is a JSON file that contains at the top level information about a pair of clusters (including text of the articles) as well as a set of "meta" attributes, that correspond to features described in Section 7.2.

The "code" folder includes MATLAB scripts for building cross-lingual similarity models introduced in Chapter 6, which can be used with publicly available Wikipedia corpus to reproduce the cross-lingual similarity evaluation. We have also made available the similarity computation over 100 languages as a service at `xling.ijs.si`.

In addition, the Event Registry system (`http://eventregistry.org/`) comes with an API, documented at `https://github.com/gregorleban/event-registry-python`, that can be used to download events and articles.

Chapter 9

Conclusions

9.1 Discussion

In the thesis we study a generalization of CCA to more than two sets of variables. We present a new result that proves that the complexity of the **SUMCOR problem is NP-hard** and describe a **novel approach** to finding **several sets of nonlinear patterns**, based on a locally convergent method. Experimentally, we observed that the performance of the local method (with linear convergence) is generally good, although we identified problem settings where convergence to local optima occurs (based on synthetic experiments we observed, that increasing the number of views increases the likelihood of such events). We present a **novel SDP relaxation** of the problem, which can be used to obtain new local solutions and to provide several **new computationally tractable bounds** on global optimality of the SUMCOR problem solutions. The usefulness of the bounds is explored on synthetic problem instances and problems related to cross-lingual text-mining. We introduce a **new preprocessing step based on random projections** to reduce the dimensionality of high dimensional problems such as in document corpora, making memory requirements tractable. We demonstrate the applicability of the approach on high-dimensional text data.

We present an **application** of two generalizations of CCA, the SUMCOR and SSCOR formulations to **cross-lingual similarity function learning**. The cross-lingual similarity functions are **applied** to the task of **cross-lingual cluster linking**, where we present and evaluate a novel approach that combines features based on semantic and language analysis. The approach is shown to be scalable both in terms of number of articles and number of languages, while accurately linking events. The approach is used in a system for real-time monitoring of global news in multiple languages, where a strong server is used to compute two millions of similarities per second.

On the task of mate retrieval, we observe that refining the LSI-based projections with hub CCA leads to improved retrieval precision (hub CCA achieves 0.7 AMRR, whereas LSI achieves 0.6, when 500 dimensions are used), but the methods perform comparably on the task of event linking (see Table 8.8). Further inspection showed that the CCA-based approach reached a higher precision on smaller clusters. The interpretation is that the linking features are highly aggregated for large clusters, which compensates the lower per-document precision of LSI. Another possible reason is that the advantage that we show on Wikipedia is lost on the news domain. This hypothesis could be validated by testing the approach on documents from a different domain.

The experiments show that the hub CCA-based features present a good baseline, which can greatly benefit from additional semantic-based features (an increase in F_1 score from 0.78 to 0.88, see Table 8.10). Even though in our experiments the addition of CCA-based

features to semantic features did not lead to great performance improvements (marginal increase in performance when compared to concept based features, see Table 8.10), there are two important benefits in the approach. First, the linking process can be sped up by using a smaller set of candidate clusters. Second, the approach is robust to languages where semantic extraction is not available, due to scarce linguistic resources. For example in Tables 8.5 and 8.6 we demonstrate that similarity models based on our **SSCOR reformulation under the hub language assumption** can be built for language pairs with scarce and low quality linguistic resources. For example, the 0.85 retrieval score for the Piedmontese-Hindi pair reported in Table 8.6 is promising, since their bilingual training set was empty.

9.2 Future Work

Regarding the work on the SUMCOR formulation, the experiments indicate that the noisy 1-dimensional embeddings present difficulties for the Horst's algorithm, which is in contrast to the performance on random generic problem instances. A natural question is, are there other problem structures that result in suboptimal behavior of the local approach?

Our empirical results focus on text data, and an interesting direction is to extend the analysis to data from other modalities, such as images, sensor streams and graphs.

Also of interest is the complexity analysis of the other generalizations proposed in [7].

Regarding the work on cluster linking, the proposed cross-lingual analysis approaches represent an important building block in our approach to cross-lingual cluster linking. The language component is built independently from the cluster linking component. It is possible that better embeddings can be obtained by methods that jointly optimize a classification task and the embedding.

Another point of interest is to evaluate our approach to cluster-linking on languages with scarce linguistic resources, where semantic annotation might not be available. For this purpose, the labelled dataset of linked clusters should be extended first. The mate retrieval evaluation shows that even for language pairs with no training set overlap, the hub CCA recovers some signal.

In order to further improve the performance of the classifier for cluster linking, additional features should also be extracted from articles and clusters and checked if they can increase the accuracy of the classification. Since the amount of linguistic resources varies significantly from language to language, it would also make sense to build a separate classifier for each language pair. Intuitively, this should improve performance since weights of individual learning features could be adapted to the tested pair of languages.

References

[1] H. Hotelling, "The most predictable criterion," *Journal of educational Psychology JEP*, vol. 26, no. 2, p. 139, 1935.

[2] M. I. Jordan and F. R. Bach, "Kernel independent component analysis," *Journal of Machine Learning Research JMLR*, vol. 3, pp. 1–48, 2001.

[3] A. Vinokourov, N. Cristianini, and J. S. Shawe-Taylor, "Inferring a semantic representation of text via cross-language correlation analysis," in *Advances in Neural Information Processing Systems 15, NIPS*, Vancouver, British Columbia, Canada, 2002, pp. 1473–1480.

[4] P. Horst, "Relations among m sets of measures," *Psychometrika*, vol. 26, pp. 129–149, 1961.

[5] L.-H. Zhang, Moody, and T. Chu, "On a multivariate eigenvalue problem, part II: Global solutions and the Gauss-Seidel method," *Preprint*,

[6] L.-H. Zhang, L.-Z. Liao, and L.-M. Sun, "Towards the global solution of the maximal correlation problem," *Journal of Global Optimization JOGO*, vol. 49, no. 1, pp. 91–107, Jan. 2011.

[7] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, pp. 433–45, 1971.

[8] J. Rupnik, P. Skraba, J. Shawe-Taylor, and S. Guettes, "A comparison of relaxations of multiset cannonical correlation analysis and applications," *CoRR*, vol. abs/1302.0974, 2013. [Online]. Available: `http://arxiv.org/abs/1302.0974`.

[9] J. Rupnik, A. Muhič, G. Leban, P. Škraba, B. Fortuna, and M. Grobelnik, "News across languages - cross-lingual document similarity and event tracking," *Journal of Artificial Intelligence Research, JAIR*, vol. 55, pp. 283–316, 2016.

[10] E. Belyaeva, A. Košmerlj, A. Muhič, J. Rupnik, and F. Fuart, "Using semantic data to improve cross-lingual linking of article clusters," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 35, Part 2, pp. 64–70, 2015, Machine Learning and Data Mining for the Semantic Web, MLDMSW.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)* Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[12] N. Bourbaki, *General Topology: Chapters 1-4*, ser. Elements of mathematics. Springer, 1998.

[13] J. Hartigan, *Clustering algorithms*. New York: John Wiley & Sons Inc, 1975.

[14] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of euclidean sum-of-squares clustering," *Machine learning ML*, vol. 75, no. 2, pp. 245–248, 2009.

[15] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[16] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[17] D. R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis; an overview with application to learning methods," Tech. Rep., 2007.

[18] B. Fortuna, N. Cristianini, and J. Shawe-Taylor, "Kernel methods in bioengineering, communications and image processing," in. Idea Group Publishing, 2006, ch. A Kernel Canonical Correlation Analysis For Learning The Semantics Of Text, pp. 263–282.

[19] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol. 1*. Philadelphia,USA: Society for Industrial and Applied Mathematics, 2002.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[21] A. Barvinok, "Problems of distance geometry and convex properties of quadratic maps," *Discrete and Computational Geometry*, vol. 13, no. 1, pp. 189–202, Dec. 1995.

[22] A. Y. Alfakih and H. Wolkowicz, "On the embeddability of weighted graphs in euclidean spaces," Tech. Rep., 1998.

[23] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. Springer Publishing Company, Incorporated, 2015.

[24] M. T. Chu and J. L. Watterson, "On a multivariate eigenvalue problem, part I: Algebraic theory and a power method," *SIAM Journal on Scientific Computing SISC*, vol. 14, no. 5, pp. 1089–1106, Sep. 1993.

[25] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *Transactions on Signal Processing TSP*, vol. 57, no. 10, pp. 3918–3929, Oct. 2009.

[26] Y. Hua-Gang, H. Gao-Ming, and J. Gao, "Nonlinear blind source separation using kernel multi-set canonical correlation analysis," *International Journal of Computer Network and Information Security*, vol. 2, no. 1, p. 1, 2010.

[27] A. Lorbert and P. Ramadge, "Kernel hyperalignment," in *Advances in Neural Information Processing Systems 25, NIPS*, Lake Tahoe, Nevada, USA, 2012, pp. 1799–1807.

[28] T. Van Noorden and J. Barkeijer, "Computing optimal model perturbations: A constraint optimization problem," *Preprint*, vol. 1308, 2004.

[29] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.

[30] M. X. Goemans and D. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM*, vol. 42, pp. 1115–1145, 1995.

[31] D. Grigoriev and D. V. Pasechnik, "Polynomial-time computing over quadratic maps I: Sampling in real algebraic sets," *Computational Complexity*, vol. 14, no. 1, pp. 20–52, Apr. 2005.

[32] Y. Nesterov, *Global Quadratic Optimization via Conic Relaxation*. Université Catholique de Louvain. Center for Operations Research and Econometrics [CORE], 1998.

[33] J. Rupnik and B. Fortuna, "Regression canonical correlation analysis," in *Learning from Multiple Sources, Neural Information Processing Systems, NIPS, Workshop Book*, Whistler Canada, 2008, p. 119.

[34] M. J. Todd, "A study of search directions in primal-dual interior-point methods for semidefinite programming," *Optimization methods and software*, 2008.

[35] G. Wang and Y. Bai, "A new primal-dual path-following interior-point algorithm for semidefinite optimization," *Journal of Mathematical Analysis and Applications*, vol. 353, no. 1, pp. 339–349, 2009.

[36] S. Arora and S. Kale, "A combinatorial, primal-dual approach to semidefinite programs," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, ACM, 2007, pp. 227–236.

[37] Z. Lu, A. Nemirovski, and R. D. Monteiro, "Large-scale semidefinite programming via a saddle point mirror-prox algorithm," *Mathematical programming*, vol. 109, no. 2-3, pp. 211–237, 2007.

[38] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 673–696, 2000.

[39] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," 5, vol. 24, Elsevier, 1988, pp. 513–523.

[40] C. Peters and M. Braschler, *Multilingual Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[41] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Language Resources and Evaluation, LRE*, vol. 45, no. 1, pp. 45–62, Mar. 2011.

[42] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180.

[43] B. Pouliquen, R. Steinberger, and O. Deguernel, "Story tracking: Linking similar news over time and across languages," in *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, Association for Computational Linguistics, Manchester, United Kingdom, 2008, pp. 49–56.

[44] J. M. Á. Rodríguez, E. R. Azcona, and L. P. Paredes, "Promoting government controlled vocabularies for the semantic web: The eurovoc thesaurus and the cpv product classification system," *Semantic Interoperability in the European Digital Library SIEDL*, p. 111, 2008.

[45] J. C. Platt, K. Toutanova, and W.-t. Yih, "Translingual document representations from discriminative projections," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing EMNLP*, Association for Computational Linguistics, Massachusetts, USA, 2010, pp. 251–261.

[46] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL*, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 1128–1137.

[47] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP: Volume 2 - Volume 2*, ser. EMNLP '09, Singapore: Association for Computational Linguistics, 2009, pp. 880–889.

[48] M. Xiao and Y. Guo, "A novel two-step method for cross language representation learning," in *Advances in Neural Information Processing Systems 26 NIPS*, Sateline, NV, USA, 2013, pp. 1259–1267.

[49] S. Dumais, T. Letsche, M. Littman, and T. Landauer, "Automatic cross-language retrieval using latent semantic indexing," in *AAAI spring symposium on cross-language text and speech retrieval. American Association for Artificial Intelligence, vol. 16. 1997*, 1997, p. 21.

[50] M. Potthast, B. Stein, and M. Anderka, "A Wikipedia-based multilingual retrieval model," in *Advances in Information Retrieval , 30th European Conference on Information Retrieval Research ECIR*, Glasgow, UK, 2008, pp. 522–530.

[51] P. Sorg and P. Cimiano, "Exploiting Wikipedia for cross-lingual and multilingual information retrieval," *Data & Knowledge Engineering*, vol. 74, pp. 26–45, 2012.

[52] "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, Gothenburg, Sweden, 2014, pp. 462–471.

[53] J. Pozniak and R. Bradford, "Optimization of cross-lingual LSI training data," *Computer and Information Science 2015*, pp. 57–73, 2016.

[54] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributional semantics," in *Proceedings of ACL*, Jun. 2014. [Online]. Available: `http://arxiv.org/abs/1404.4641`.

[55] I. Vulic and M. Moens, "Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, Beijing, China, 2015, pp. 719–725.

[56] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu, "Deep multilingual correlation for improved word embeddings," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, 2015, pp. 250–256.

[57] S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth, "Cross-lingual models of word embeddings: An empirical comparison," *Arxiv*, vol. 1604.00425, 2016. [Online]. Available: `http://arxiv.org/pdf/1604.00425v1.pdf`.

[58] V. I. Spitkovsky and A. X. Chang, "A cross-lingual dictionary for english Wikipedia concepts," in *Proceedings of the Eight International Conference on Language Resources and Evaluation LREC'12*, Istanbul, Turkey: ELRA, May 2012.

[59] D. Rinser, D. Lange, and F. Naumann, "Cross-lingual entity matching and infobox alignment in Wikipedia," *Information Systems*, vol. 38, no. 6, pp. 887–907, 2013.

[60] A. Barrón-Cedeno, M. L. Paramita, P. Clough, and P. Rosso, "A comparison of approaches for measuring cross-lingual similarity of Wikipedia articles," *Advances in Information Retrieval*, pp. 424–429, 2014.

[61] A. Søgaard, Ž. Agić, H. M. Alonso, B. Plank, B. Bohnet, and A. Johannsen, "Inverted indexing for cross-lingual nlp," in *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, Beijing, China, 2015.

[62] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science JASIS*, vol. 41(6), pp. 391–407, 1990.

[63] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *Society for Industrial and Applied Mathematics SIAM Review*, vol. 53, no. 2, pp. 217–288, May 2011.

[64] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," *Proceedings of the American Psychological Association*, pp. 227–228, 1968.

[65] A. Gifi, *Nonlinear Multivariate Analysis*. Wiley Series in Probability and Statistics, 1990.

[66] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web WWW Companion*, ser. WWW Companion '14, Seoul, Republic of Korea: International World Wide Web Conferences Steering Committee, 2014, pp. 107–110.

[67] ——, "Cross-lingual detection of world events from news articles," in *Proceedings of the 13th International Semantic Web Conference ISWC*, Riva del Garda - Trentino, Italy, 2014, pp. 21–24.

[68] L. Zhang and A. Rettinger, "X-lisa: Cross-lingual semantic annotation," *Proceedings of the Very Large Data Bases VLDB Endowment*, vol. 7, no. 13, pp. 1693–1696, Aug. 2014.

[69] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, Nov. 5, 1971.

[70] R. B. Holmes, "On random correlation matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 12, no. 2, pp. 239–272, Mar. 1991.

[71] R. B. Bendel and M. R. Mickey, "Population correlation matrices for sampling experiments," *Communications in Statistics: Simulation and Computation*, vol. B7, no. 1, pp. 163–182, 1978.

[72] B. Fortuna, "Kernel canonical correlation analysis with applications," in *Proceedings of the 7th Multiconference on Information Society, IS*, Ljubljana, Slovenia, 2004.

[73] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analsysis," in *Proceedings of the 13th Multiconference on Information Society, IS*, Ljubljana, Slovenia, 2010, pp. 201–204.

[74] B. Fortuna, C. Galleguillos, and N. Cristianini, "Detecting the bias in media with statistical learning methods," in *Text Mining: Classification, Clustering, and Applications*, Chapman and Hall/CRC press, 2009.

[75] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *The Tenth Machine Translation Summit, MT Summit X*, vol. 5, Phuket, Thailand, 2005, pp. 79–86.

[76] E. M. Voorhees *et al.*, "The TREC-8 question answering track report," in *Proceedings of the 8th Text Retrieval Conference TREC-8*, vol. 99, Gaithersburg, MD, USA, 1999, pp. 77–82.

Bibliography

Publications Related to the Thesis

Journal Articles

J. Rupnik, A. Muhič, G. Leban, P. Škraba, B. Fortuna, and M. Grobelnik, "News across languages - cross-lingual document similarity and event tracking," *Journal of Artificial Intelligence Research, JAIR*, vol. 55, pp. 283–316, 2016.

E. Belyaeva, A. Košmerlj, A. Muhič, J. Rupnik, and F. Fuart, "Using semantic data to improve cross-lingual linking of article clusters," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 35, Part 2, pp. 64–70, 2015, Machine Learning and Data Mining for the Semantic Web, MLDMSW.

Conference Paper

J. Rupnik and B. Fortuna, "Regression canonical correlation analysis," in *Learning from Multiple Sources, Neural Information Processing Systems, NIPS, Workshop Book*, Whistler Canada, 2008, p. 119.

J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analsysis," in *Proceedings of the 13th Multiconference on Information Society, IS*, Ljubljana, Slovenia, 2010, pp. 201–204.

A. Muhič, J. Rupnik, and P. Škraba, "Cross-lingual document similarity," in *Proceedings of the 34th International Conference on Information Technology Interfaces, ITI*, IEEE, Cavtat / Dubrovnik, Croatia, 2012, pp. 387–392.

Jan Rupnik, Andrej Muhič, and Primoz Škraba, "Multilingual document retrieval through hub languages," in *Proceedings of the 15th Multiconference on Information Society, IS*, Ljubljana, Slovenia, 2012, pp. 201–204.

——, "Learning cross-lingual similarities," in *Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity, Neural Information Processing Systems, NIPS, Workshop Book*, Granada, Spain, 2011, p. 35.

Biography

Jan Rupnik was born in Ljubljana on 6 December 1982.

Following graduation at the Faculty of Mathematics and Physics, University of Ljubljana, with a degree in applied mathematics (Diploma) in 2006, he was employed as a researcher at the Artificial Intelligence Laboratory, Jožef Stefan Institute. In 2007 he enroled in the New Media and E-science doctoral study program at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia.

His research interests include Machine Learning, Data Mining, Data Fusion, Cross-Lingual Text Mining, Predictive Analytics, Applications of Data Mining in different domains. Most of his research work is connected to the development of statistical methods that enable cross-modal data integration, with a focus on scalability.

Jan Rupnik has been involved in a number of EU FP7 projects, including SMART (Statistical Multilingual Analysis For Retrieval And Translation), XLIKE (Cross- lingual Knowledge Extraction), EURIDICE (The Intelligent Cargo Concept in the European Project) and SOPHOCLES (Self-Organised information PrOcessing, CriticaLity and Emergence in multilevel Systems).