Department of Computer Science
Series of Publications A
Report A-2011-1

Data fusion and matching by maximizing statistical dependencies

Abhishek Tripathi

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public critism in Auditorium CK112, Exactum, on February 10th, 2011, at 12 noon.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.helsinki.fi (Internet)

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-6749-5 (paperback)
ISBN 978-952-10-6750-1 (PDF)
Computing Reviews (1998) Classification: G.0, I.0
Helsinki 2011
Helsinki University Print

Data fusion and matching by maximizing statistical dependencies

Abhishek Tripathi

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
abhishek.tripathi@cs.helsinki.fi
http://www.cs.helsinki.fi/abhishek.tripathi

PhD Thesis, Series of Publications A, Report A-2011-1
Helsinki, January 2011, 89 + 109 pages
ISSN 1238-8645
ISBN 978-952-10-6749-5 (paperback)
ISBN 978-952-10-6750-1 (PDF)

Abstract

Multi-view learning is a task of learning from multiple data sources where each source represents a different view of the same phenomenon. Typical examples include multimodal information retrieval and classification of genes by combining heterogeneous genomic data. Multi-view learning methods can be motivated by two interrelated lines of thoughts: if single view is not sufficient for the learning task, other views can complement the information. Secondly, learning by searching for an agreement between views may generalize better than learning from a single view. In this thesis, novel methods for unsupervised multi-view learning are proposed.

Multi-view learning methods, in general, work by searching for an agreement between views. However, defining an agreement is not straightforward in an unsupervised learning task. In this thesis, statistical dependency is used to define an agreement between the views. Assuming that the shared information between the views is more interesting, statistical dependency is used to find the shared information. Based on this principle, a fast linear preprocessing method that performs data fusion during exploratory data analysis is introduced. Also, a novel evaluation approach based on the dependency between views to compare vector representations for bilingual corpora is introduced.

Multi-view learning methods in general assume co-occurred samples for the

views. In many applications, however, the correspondence of samples is either not known in advance or is only partially known. Examples include probes used by different microarray platforms to measure genetic activities for the same set of genes and unaligned or partially aligned parallel documents in statistical machine translation. In this thesis, a novel approach is introduced for applying multi-view learning methods when sample correspondence between the views is not known.

A novel data-driven matching algorithm is proposed to infer a one-to-one matching of samples between two views. It is worth noticing that defining a similarity measure in such a case is not straightforward since the objects may have different sets of features. We assume that true matching of samples will maximize the statistical dependency between two views. A two-step iterative solution is proposed for the matching problem that uses canonical correlation analysis (CCA) to measure linear dependency. A non-linear version of the matching algorithm using kernel CCA is also presented. It is also shown how the prior information in the form of soft and hard constraints can be incorporated in the matching algorithm to improve the matching task.

The proposed matching algorithm is purely data-driven, hence it involves uncertainties due to measurement errors. The matching algorithm is extended to a more realistic setting where each view is represented by multiple instantiations. A concrete example is matching of metabolites between humans and mice, where each human-mouse pair will result in a different matching solution. A generalized matching algorithm, called *consensus matching*, is proposed which combines different matching solutions to a give a final matching of the samples.

Computing Reviews (1998) Categories and Subject Descriptors:
G.0 Mathematics of Computing
I.0 Computing Methodologies

General Terms:
bi-partite matching, (kernel) canonical correlation analysis, data fusion, dependency maximization, mutual information, multi-view learning

Additional Key Words and Phrases:
bioinformatics, bilingual corpus, statistical machine translation

Contents

Acknowledgements

This work has mainly been carried out at the Department of Computer Science, University of Helsinki, and partly at Department of Information and Computer Science at Aalto University School of Science and Technology. The work was financially supported by the Academy of Finland, University of Helsinki's project funds and also by Helsinki Institute for Information Technology (HIIT) and thesis writing fund from Department of Computer Science, University of Helsinki. I would also like to thank the Helsinki graduate school of computer science and engineering (HeCSE) and PASCAL, an EU network of excellence for Pattern Analysis, Statistical Modeling and Computational Learning for travel grants.

I sincerely thank my supervisor Professor Samuel Kaski who taught me every aspect of research, motivated me with his research ideas and discussions, and guided me throughout my PhD. I also extend my heartfelt gratitude to my instructor Dr. Arto Klami for his constant support, for patiently answering my doubts, and for encouraging new thoughts and ideas. Professor Kaski and Dr. Klami have contributed in most of my research papers, and this thesis is a result of their kind supervision. I would also like to thank Professor Juho Rousu for constantly mentoring, supervising my PhD studies, and for his guidance in thesis writing.

My gratitude to all my other co-authors, in particular, Dr. Suvi Savola, Mr. Sami Virpioja, Dr. Matej Orešič, Dr. Krista Lagus and Professor Sakari Knuutila for their invaluable contributions to the thesis. It was a great experience working with all of them.

I thank all my colleagues at the department, in particular, Greger Lindén and Professor Esko Ukkonen for their unconditional help and support, and for giving a friendly work environment. I thank Sourav and Anupam for their great friendship. I also thank all my colleagues and support staff at the Department of Information and Computer Science, Aalto, in particular, Janne Nikkilä, Leo Lahti, Merja Oja, Jaakko Peltonen, Jarkko Salojärvi, Jarkko Venna, and all other members of the MI group.

I express my deep gratitude to the pre-examiners of this thesis, Dr.

Craig Saunders and Dr. Tijl De Bie, for their time and invaluable feedback that helped improve this dissertation.

I dedicate my thesis to my parents who have been a constant support and motivational factor behind my PhD. I thank my brother for his guidance at every stage of my life. I specially thank Mrs. and Mr. Srikanthan, and Preethy for their affection, support, and also for proof-reading certain parts of my thesis. Not to forget my friends Surya, Ankur, Ankur Sinha, Saumyaketu, Aribam, Mradul and Rohit for their timely motivation, and encouragement not only during my PhD but since we are friends. I would also like to thank my Indian friends in Finland for the great time we shared together, and also for the excellent scientific and social discussions.

LIST OF PUBLICATIONS

The thesis consists of an introduction and the following publications:

1. Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

2. Sami Virpioja, Mari-Sanna Paukkeri, Abhishek Tripathi, Tiina Lindh-Knuutila, and Krista Lagus. Evaluating Vector Space Models with Canonical Correlation Analysis. Submitted in *Natural Language Engineering*.

3. Suvi Savola, Arto Klami, Abhishek Tripathi, Tarja Niini, Massimo Serra, Piero Picci, Samuel Kaski, Diana Zambelli, Katia Scotlandi and Sakari Knuutila. Combined use of expression and CGH arrays pinpoints novel candidate genes in ewing sarcoma family of tumors. *BMC Cancer*, 9:17, 2009.

4. Abhishek Tripathi, Arto Klami, and Samuel Kaski. Using dependencies to pair samples for multi-view learning. In *IEEE ICASSP 2009, the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1561-1564, 2009, IEEE Computer Society, Los Alamitos, CA, USA.

5. Abhishek Tripathi, Arto Klami, Sami Virpioja. Bilingual sentence matching using Kernel CCA. In *IEEE International Workshop on Machine Learning for Signal Processing, 2010 (MLSP 2010)*, pages 130–135, 2010, IEEE Computer Society, Los Alamitos, CA, USA.

6. Abhishek Tripathi, Arto Klami, Matej Orešič, and Samuel Kaski. Matching samples of multiple views *Data Mining and Knowledge Discovery*, 2010 (To appear).

SUMMARY OF PUBLICATIONS AND THE AUTHOR'S CONTRIBUTION

All the publications are joint efforts by the authors, and all participated in the writing. Exceptions and detailed contributions are mentioned below.

In Publication 1, a general framework of combining multiple data sources by maximizing mutual dependencies between them is presented. The idea and experimental design were jointly developed. The author has implemented the required methods and performed all the experiments.

In Publication 2, a novel approach to evaluate vectorial representations of sentence-aligned bilingual text is proposed. Assuming that any correlation between the documents in two languages will reflect the semantic similarity, we proposed an evaluation criterion based on statistical dependency to compare different feature representations. The idea is to select the vector representation that results in the highest dependency for a given bilingual corpus. The author participated in developing a dependency measure based on CCA, and in experimental design and writing of the manuscript.

In Publication 3, the author implemented the matching of mRNA and copy number probes based on chromosome location and sequence information. This work motivated us to develop a matching algorithm that could use measurement data for the matching.

In Publication 4, a novel problem of multi-view learning in a non-standard setting, when the sources are non-co-occurred, is introduced. The problem is solved by matching the samples between two data sets using a new matching algorithm. The idea and experimental design were developed jointly. The author contributed in deriving formulas, implemented the required methods and performed all experiments.

In Publication 5, the matching algorithm is extended to kernel matching by modeling non-linear dependencies. The kernel matching is shown to perform better than the linear counterpart in a bilingual sentence matching problem. The idea and experimental design were developed jointly. The author implemented the required methods and performed all experiments. The data were pre-processed by Sami Virpioja.

In Publication 6, a generalized matching algorithm, called *consensus matching*, is proposed. Here, the idea is to combine multiple matching solutions to find a consensus, and this is applied to find matching of metabolites between men and mice. A computationally feasible solution to combine multiple matching solutions is proposed. The idea and experimental design were developed jointly. The author contributed in deriving formulas, implemented the required methods and performed all experiments.

LIST OF ABBREVIATIONS

AIC	Akaike information criterion
ALL	Acute lymphoblastic leukemia
BIC	Bayesian information criterion
CCA	Canonical correlation analysis
CGH	Comparative genomic hybridization
COCO	Constrained covariance
DNA	Deoxyribonucleic acid
drCCA	Dimensionality reduction with canonical correlation analysis
EM	Expectation maximization
ESFT	Ewing sarcoma family of tumors
GCCA	Generalized canonical correlation analysis
HSIC	Hilbert-Schmidt independence criterion
IR	Information retrieval
KCC	Kernel canonical correlation
KCCA	Kernel canonical correlation analysis
KGV	Kernel generalized variance
KLDCCA	Kernel local discrimination CCA
KMI	Kernel mutual information
LDCCA	Local discrimination CCA
LPCCA	Locality preserving CCA
LPP	Linear programming problem
LSAP	Linear sum assignment problem
MKL	Multiple kernel learning
mRNA	Messenger RNA
PCA	Principal component analysis
RBF	Radial basis function
RKHS	Reproducing kernel Hilbert space
RNA	Ribonucleic acid
SVM	Support vector machine
SVM-2K	KCCA followed by SVM
VSM	Vector space model
WWW	World wide web

LIST OF SYMBOLS

In this thesis boldface symbols are used to denote matrices and vectors. Capital symbols (e.g. $\mathbf{W}$) signify matrices and lowercase symbols ($\mathbf{w}$) column vectors.

$\|.\|$	Norm operator
$\langle \mathbf{x}, \mathbf{y} \rangle$	Inner product of $\mathbf{x}$ and $\mathbf{y}$
$\text{cov}(\mathbf{X}, \mathbf{Y})$	Covariance of $\mathbf{X}$ and $\mathbf{Y}$
D	dimensionality of the data
$d_{KL}(p, q)$	Kullback-Leibler divergence between distributions p and q
$\det(\mathbf{C})$	Determinant of matrix $\mathbf{C}$
$I(\mathbf{X}, \mathbf{Y})$	Mutual information between $\mathbf{X}$ and $\mathbf{Y}$
$G = (V, E)$	A graph with vertices V and edges E
$G = (X \cup Y, E) = ((X, Y), E)$	Bipartite graph with node sets X and Y, and edge set E
$H(\mathbf{X})$	Entropy of $\mathbf{X}$
$H(\mathbf{X}, \mathbf{Y})$	Joint entropy of $\mathbf{X}$ and $\mathbf{Y}$
$H(\mathbf{X}\|\mathbf{Y})$	Conditional entropy of $\mathbf{X}$ given $\mathbf{Y}$
N	number of observations
$p(\mathbf{x}, \mathbf{y})$	Joint probability distribution
$\text{O}(.)$	Running time
$\mathbf{X}, \mathbf{Y}$	data matrices in $\mathbb{R}^{D \times N}$
$\mathbf{X}^T$	Transpose of $\mathbf{X}$
$\mathbf{x}$, $\mathbf{y}$	data samples, vectors in $\mathbb{R}^D$
$K(\mathbf{x}_i, \mathbf{x}_j)$	a kernel function
$\rho(\mathbf{x}, \mathbf{y})$	Pearson correlation between $\mathbf{x}$ and $\mathbf{y}$

Chapter 1

Introduction

1.1 General background

Machine learning is a field of science that lies in the intersection of computer science and statistics. From a computer science point of view, the defining question for machine learning is how to make computer programs that learn from experience. From a statistics point of view, machine learning not only deals what can be inferred from the data but also how we can effectively store, index, retrieve, or merge the data to make better statistical inferences (Mitchell, 2006). In this thesis, I focus on machine learning algorithms that learn from multiple independent views of the data. Typical examples include: identifying cancer-related genes by analyzing genetic-measurements obtained by different microarray platforms or under different biological conditions; classification of webpages based on their content and the contents of pages they link to; object recognition from color and shape.

Learning from data, in general, refers to summarizing the collection of observations, called the *data*, by finding regularities or patterns in it. The purpose of summarizing the data could be to predict the behavior of a new observation, or to simply understand the underlying phenomenon. The typical examples of learning from data include classification, regression, clustering or density estimation. These methods have been successfully used in many important fields, for example, bioinformatics, text categorization, and information retrieval.

Finding reliable regularities or patterns in data is, in principle, relatively easier if we have large collections of data. However, the amount of data is limited in many applications, for instance, the task of inferring differentially expressed genes in a typical microarray study where thousands of genes are measured over few arrays (Dudoit et al., 2002). One of the most

challenging problems in machine learning is to learn from a small number of observations. This is usually referred to as *large p, small n* problem. Here, n is the number of observations, and p is the number of features. In such cases, the model tends to over-learn the data, that is, the model describes the given data well, but does not capture the underlying process that generates the data. Also, the data is usually mixed with noise. This leads to poor learning and hence poor predictive performance.

There has been a lot of work on how to learn reliable models given limited amount of data, for instance, using Bayesian approaches (West, 2003). However, in this thesis, the problem is approached from a different perspective. Consider a situation in which the observations are represented in multiple views, where each view represents a different aspect of the data. Using multiple views for learning has been motivated through two different lines of thoughts: (i) Different views might contain partly independent information about the task at hand, and combining these complementary information increases the total information for the task at hand. (ii) In another setting when each view is sufficient, the model learned based on the agreement between the views may generalize better (Dasgupta et al., 2001). Another advantage of learning from multiple views is noise reduction. Assuming that noise is independent between the views, averaging over multiple views may reduce the effect of noise. Such approaches can be termed as *multi-view learning* approaches where the task is to learn from multiple views of the data.

The most important question in multi-view learning is how to combine different views. In other words, which information in multiple views is *relevant* to the problem that needs to be solved. This is rather well defined, though not trivial, in a supervised task. For instance, in a classification task, multiple views should be combined such that the classification accuracy is optimal. Hence, the information which improves the classification accuracy is relevant. Identifying the relevant information is, however, not straightforward in an unsupervised learning task. In the absence of a clear hypothesis, it is not easy to define how to combine multiple views. In this thesis, the problem of learning from multiple views in an unsupervised setting is considered. However, such approaches can also be extended to a semi-supervised or a supervised setting.

Recently, many approaches have been proposed towards learning from multiple views in both supervised and unsupervised settings. The principle behind learning from multiple views is to learn based on agreements between them. Views are said to agree with each other if the predictions made based on them are similar (Yarowsky, 1995; Blum and Mitchell, 1998). In

this thesis, *statistical dependency* between the views is used to define the agreement. Dependency between the views can be used to represent what is common or shared between them. These methods are suitable to the problems where the shared information between the views is more interesting than the information that is specific to a view. Hence, relevance is defined through the statistical dependency between the views. The information that is specific to a view is considered not interesting and can be discarded as noise. In other words, using dependency for multi-view learning methods helps separating relevant and irrelevant information for a given task.

Multi-view learning methods, in general, assume co-occurrence of samples, that is, each view consists of the same set of samples. For instance, while combining gene expression data from different platforms or species, the correspondence of probes should be known (Hwang et al., 2004); images must be paired with the corresponding texts in an image retrieval task (Vinokourov et al., 2003b); documents in two languages should be mapped at some level, for example at sentence-level or paragraph-level, in machine translation (Barzilay and McKeown, 2001) or cross-language information retrieval (Vinokourov et al., 2003a). The requirement of co-occurrence of views is quite hard in practice and limits the possibility of using vast amount of data available in different views. In many real world examples, correspondence of samples between two views is either not known or is only partially known in advance. Consider, for example, the huge amount of unaligned parallel or comparable texts in two languages in the WWW. Manual mapping of documents is cumbersome. This hinders the use of multi-view learning methods in such applications, in this case, in machine translation or cross-language information retrieval.

Standard multi-view learning methods can be applied to non-co-occurred views or data sources, if the correspondence of samples between the views is first inferred somehow. Since different views represent different aspects of the same concept or phenomenon, we can assume that there is an implicit one-to-one correspondence of samples between the views, and such correspondence is not known, or only partially known in advance. However, matching of samples between two views is not straightforward, because each view has a different set of features to represent a sample. Defining a measure of similarity, for instance, distance-based similarity between the samples in two different feature spaces, is far from trivial.

In this thesis, the problem of multi-view learning in a non-standard setting when the views are not co-occurred is considered. A novel data-driven method based on statistical dependency to match the samples between two views is introduced. The underlying assumption is that the correct match-

ing of samples will result in the maximal dependency between the views. Hence, the matching algorithm finds a matching of samples that maximizes the statistical dependency between the views. Given such a matching solution, any standard multi-view learning methods can be applied. In this thesis, the matching algorithm is empirically demonstrated on three examples: matching of probes of gene expression profiles from two microarray platforms; matching of sentences between bilingual parallel corpora using monolingual data; and matching of metabolites between humans and mice.

1.2 Contributions and organization of thesis

The main focus of this thesis is to study and develop methods to learn from multiple data sources. The multi-view learning methods proposed in this thesis are based on the following principle: Information that is shared by all the views or data sources is more interesting than source-specific information. We proposed the use of statistical dependency between the views to find what is shared between the views. The main contributions of this thesis are following:

1. An unsupervised data fusion approach that preserves the information shared between the data sources is introduced.

2. A novel evaluation method to compare vector space models for sentence-aligned parallel bilingual corpora is introduced. The evaluation approach provides a direct measure for evaluation based on statistical dependency.

3. A novel problem setting of multi-view learning when the correspondence of samples between the views is not known is introduced.

4. A data-driven matching algorithm to infer the matching of samples between two views is proposed. A two-step iterative solution to the matching problem is proposed. It uses CCA to model dependency in the first step and the assignment problem to infer the matching in the second step .

5. A non-linear extension of the matching algorithm using KCCA is proposed in order to utilize non-linear dependency for the matching. The empirical comparison of matching algorithms based on CCA and KCCA is demonstrated on a sentence-matching task for bilingual parallel corpora using monolingual data.

6. A generalized matching algorithm, called *consensus matching*, in a more realistic application when the two views have multiple representations is introduced. An approach to combine the matching solutions obtained using any two representations of the views is proposed. The consensus matching is implemented in a real research problem of inferring the matching of metabolites between humans and mice.

The organization of thesis is as follows. In Chapter 2, a brief overview of learning methods is given. The chapter describes the problem of learning from single data source, the challenges of learning and the different learning setups.

In chapter 3, the concept of learning from multiple data sources is defined and motivated through real world examples. The current state-of-the-art multi-view learning methods are also discussed. Section 3.1 describes the multi-view learning methods in the context of this thesis, and explains the principle behind using the statistical dependency for multi-view learning. Section 3.2 explains the notion of dependency in mathematical terms and gives a brief overview of measures of dependency. Sections 3.3.1 and 3.3.2 describe the methods to model dependencies between two data sources, and their variants are described in the following subsections.

The chapter is finally concluded by describing two novel methodologies introduced in this thesis. Section 3.4 explains an unsupervised data fusion approach based on maximizing statistical dependencies between views. Section 3.5 describes a direct evaluation approach to find an appropriate vector representation of sentences for a given bilingual corpora.

In Chapter 4, the problem of multi-view learning when the correspondence of samples between views is not known is described and a matching algorithm to infer the correspondence of samples between two views is introduced. Section 4.1 explains the matching problem in general, and describes the standard algorithms to solve the matching problem when the samples in two views are represented by same set of features. In Section 4.3, the matching problem when the samples in two views have non-comparable features sets is formulated. Section 4.4.1 presents a solution of the matching problem using CCA, and Section 4.4.2 presents the non-linear extension of matching algorithm using KCCA. The subsequent sections describe semi-supervised matching and the consensus matching methods, and their applications. In Section 4.6, few related matching methods are discussed and compared with the matching method described in this thesis.

Finally, Section 5 concludes the thesis and discusses possible future research directions based on the work proposed in this thesis.

Chapter 2

Learning from data

The core aim of machine learning is to build intelligent systems that can perform real world task in a robust and adaptive way. In order to build such a system, the basic idea is to provide a lot of examples to the system to make it learn the particular desired behavior. As an example, consider the task of face recognition where the task is to identify faces given an image (Turk and Pentland, 1991). The system is trained to identify a face by giving it examples; images with faces and images without faces. This is called *learning*. Once the system is trained, it should be able to identify a face given a new image.

Learning, in the context of this thesis, is a task of finding regularities or patterns in data, which is a collection of observations. In learning, we typically define a model to fit to the given data with respect to model parameters, for instance, maximizing likelihood or minimizing some cost function. The learned model can be used to predict the behavior of a new sample or to infer the underlying pattern of the given data. For instance, the task could be to label a new image by either of the pre-defined categories: The image has a face, or the image does not have a face. This is called *classification*. Such learning techniques can be applied in a variety of applications, for example, spam detection, speech recognition, financial prediction, fraud detection, medical diagnosis, and game playing.

In probabilistic or statistical learning, we assume that the data is generated by an underlying distribution. The generating distribution contains all the information we may need but it is not possible to access that distribution directly. The underlying distribution can be inferred through the training samples which are assumed to be independently and identically distributed. The task is to define a function space, also called a *model family*, according to the prior information, and to learn a function (or model) from the functional space that describes the test data well.

2.1 Notation

In this thesis, the collection of N observations is represented by the set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, also called a *training set*. Each observation is an instance of a random variable $\mathbf{x}$ drawn independently from an unknown probability distribution $p(\mathbf{x})$. Here, each observation $\mathbf{x}_i$ is represented as a D-dimensional vector, where each element of the vector is called a *feature*. The collection of observations (or samples) can be represented as a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with N rows and D columns. That is, each observation is represented as a row, and feature as a column in the data matrix.

2.2 Model complexity

The complexity of a good model depends on the dimensionality of the data. For instance, if the dimensionality D is small, the underlying pattern in the data is relatively simpler, and we need few parameters to define a model. When the data dimensionality D is higher, more number of parameters are needed to define a model in order to infer the underlying pattern. That is, a more complex model is needed for high dimensional data. In a simple case, when $N \gg D$, the task of learning is relatively easier. The problem of *large p, small n* appears when the dimensionality is high and the number of samples is smaller. Here, n represents the number of samples N, and p represents the dimensionality D.

2.3 Generalization ability of model

In practice, the available training data is often just a fraction of what would be needed to really explain the underlying phenomenon we aim to model. One of the main problems is thus to build a model that not only describes the data well but also explains the underlying distribution, or phenomenon. The trained model should also be able to correctly characterize the new samples that were not available for the learning; such samples comprise the *test set*. The ability of a model to correctly characterize the test sample is called *generalization*. For instance in the task of face recognition, the trained model should identify the faces in the new unseen images.

A model is called *over-fitted* or *overlearned* if it explains the training data well but does not generalize to the test data. In over-fitting, the model not only learns the underlying distribution but also learns the noise in the training data. Hence, the predictive ability of the model to the test data becomes poor. The models tend to overfit when the ratio of N to M_C be-

comes smaller, where M_c is the number of model parameters. The number of model parameters may also increase as the dimensionality D of the data grows. Typical solutions to avoid overlearning includes adding a regularization term, or restricting the complexity of the model, or preprocessing the data to reduce dimensionality.

Generalization ability of a model and the issue of overfitting are very important issues in machine learning, and many techniques have been developed, for instance, *cross-validation* and *bootstrapping* to solve these issues. The basic idea is to create a validation data out of the training data. The model can be trained on remaining data and the validation data can be used to check the generalization ability.

In cross-validation, the model is repeatedly validated on a subset of training data while trained on the remaining data. There are many ways of choosing a subset for the validation. In K-fold cross-validation (Bengio and Grandvalet, 2004, See), the training data is divided into K parts; one of the K parts is used for validation at a time while the rest are used for learning. In leave-one-out cross-validation, only one sample is left out for validation and the rest are used for training repeatedly; this is particularly useful if the data are scarce. The model with the best predictive performance based on cross-validation is selected.

In bootstrapping, the main idea is to create new data sets by re-sampling observations with replacement from the original data set. The generalization ability can be evaluated by looking at the variability of predictions between different bootstrap data sets. A detailed description of bootstrapping can be found in (Efron and Tibshirani, 1993).

One major drawback with methods like cross-validation or bootstrapping that repeatedly use a validation set to measure generalization ability is that the number of training runs may grow exponentially as the number of model parameters grows. Moreover, if the model is itself computationally expensive, multiple training runs may not be feasible. Approaches based on "information criteria" provide methods that do not need multiple training runs, and avoid overlearning by adding a penalty term for more complex models. For instance, *Akaike information criterion* (AIC) by Akaike (1974) and *Bayesian information criterion* (BIC) by Schwarz (1978) avoid the bias due to overfitting by adding a penalty to the maximum likelihood; hence they tend to favor a simpler model.

2.4 Learning setups

The problem of learning can be divided into different categories based on the application and available data. Suppose we are given a data set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ is a sample and y_i is the corresponding label. If the labels y_i of the training samples are known, the learning problem is known as *supervised learning* where the task is to predict the label of a new sample. If the labels are discrete, the problem is known as classification problem, and if the labels are continuous, it is called a regression problem.

The learning problem where the labels of training samples are not known or in some cases the notion of labels may not even exist is called *unsupervised learning*. Examples of unsupervised learning include finding groups of similar samples, also known as *clustering*, or determining the underlying distribution of the sample, known as *density estimation*. Another important learning setup is the *semi-supervised learning* which is a hybrid of supervised and unsupervised learning tasks. In semisupervised learning, labels are known only for a fraction of training samples. Chapelle et al. (2006) provides a detailed overview of semi-supervised learning and its applications.

Chapter 3

Multi-view learning

The fundamental aim of using more than one source of information is usually to achieve a more comprehensive understanding of a task, a concept or a phenomenon. With the rapid advancement of technology, huge amount of data is being generated and stored in different parts of world, and in many cases, the generated data concern a similar or related objective. Combining existing data sources about related concepts not only leads to a better understanding, but also saves valuable resources in terms of time, money and manpower for data generation. In biological science, for instance, several research groups might be working on a particular disease under different conditions and from different perspectives. Combining results and data from such studies may lead to a better understanding of the task.

Integrating information from different sources can be described in many contexts. For instance, it could refer to effectively storing, indexing, or retrieving data sets from different sources. As an example, combining multiple sources could refer to combining several databases into a single unified view. *Data warehousing* is a general term for combining different databases into a single queriable schema. Examples of database integration include combining databases of two merging companies, or web services that combine publicly available databases. Database integration is however not considered in this work. In this thesis, combining information from multiple sources refers to the task of *learning* from multiple data sets. The idea is to utilize relevant information from multiple data sets in order to improve the learning task.

Combining information from multiple sources has recently attracted a lot of interest in the machine learning community (Rüping and Scheffer, 2005; Hardoon et al., 2009; Cesa-Bianchi et al., 2010). Learning from multiple sources refers to the problem of analyzing data represented by multiple views of the same underlying phenomenon. It has been studied under dif-

ferent names, for instance, multi-view learning, multi-task learning, data fusion and transfer learning. These concepts differ according to the nature of the learning task and assumptions about the dependency between the sources. The underlying idea is to use information from multiple sources to improve the generalization ability of the learned model. One of the important questions is to decide on how the information from multiple sources can be combined, and it becomes more challenging in an unsupervised setting.

In this thesis, I consider the unsupervised learning task from multiple sources where each source represents a different view of the data. Such methods can be categorized under the term *multi-view learning*. An approach that uses statistical dependency between views to combine multiple views for learning is proposed. In the rest of this section, a general overview of multi-view learning methods is given. Next, different approaches to combine multiple views for learning has been described. Finally, the use of dependency between views for multi-view learning methods is discussed in Section 3.1.

Multi-view learning is a task of learning from instances represented by multiple independent views or sets of features. The underlying assumption in such methods is that additional views of the related concept can be incorporated in the task of learning to improve the predictive performance. Examples include: web pages can be classified based on their content, but the accuracy of classification can be improved when using the content on hyperlinks (Blum and Mitchell, 1998); a biological function can be better understood by combining heterogeneous biological data, for instance, gene measurements, protein measurements, metabolomics and interaction data; explicit or implicit feedback of user can be used to improve search result, or image ranking (Pasupa et al., 2009); the performance of automatic speech recognition can be improved using facial visual information (Potamianos et al., 2003); in collaborative filtering systems, the performance of a recommender system can be improved by combining movie ratings with the content data, genre of the movie (Williamson and Ghahramani, 2008).

Multi-view learning methods have been studied under different names and in different settings. Blum and Mitchell (1998) proposed a semi-supervised approach, called co-training, that allows using unlabeled data to augment the smaller set of labeled data for training when two distinct and redundant views for each sample are present. Here, it is assumed that either view is sufficient for learning if enough labeled data is available. Co-training was closely related to an earlier rule-based approach by Yarowsky (1995) that utilized unlabeled data in the context of the word-sense disambiguation problem. Both approaches assumed that a classifier generalizes

better if it is based on maximizing the agreement between two views, and this was later justified by Dasgupta et al. (2001).

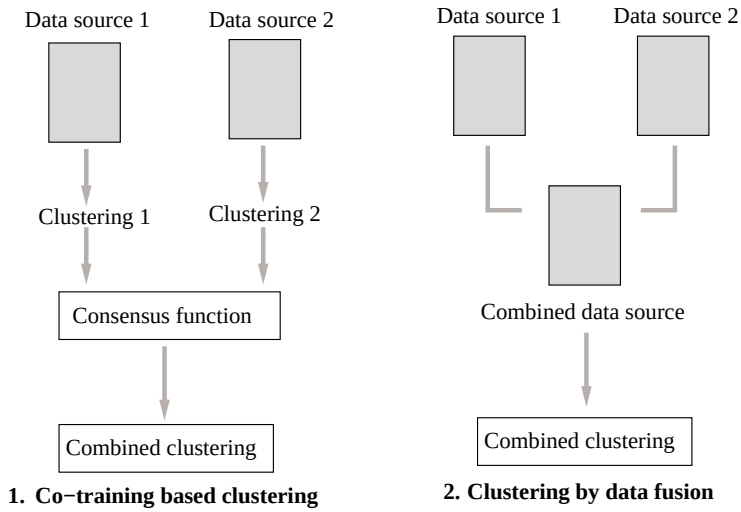

Figure 3.1: Illustration of two types of multi-view learning. (1) First subfigure shows co-training based learning for a clustering task. Each view is separately partitioned using a clustering algorithm, and then a consensus is defined by combining the two partitions. (2) Second figure shows multi-view learning based on data fusion. A combined representation of all views is obtained, and clustering is done on the combined representation.

Extending the concept of co-training, Bickel and Scheffer (2004) presented a multi-view clustering algorithm where class variables were replaced by mixture coefficients. Figure 3.1 illustrates the concept of co-training in the context of clustering. It was empirically shown in (Bickel and Scheffer, 2004) that learning based on agreement between different views, even if they are randomly partitioned, is better than learning based on a single view. Approaches based on co-training learn a model on each view separately, and combine the models by defining an agreement between them.

Another approach to learning from multiple sources is to combine the views together prior to applying any learning method as shown in Figure 3.1. The important question here is how to combine multiple views together. If all the views are equally relevant and useful, a combined view can be obtained by averaging over all the views. Another option will be to weight each view based on its *relevance* for the given task. In a supervised problem, for instance in classification, views can be combined such that the classification accuracy is improved. Lanckriet et al. (2004) used kernel ma-

trices to represent heterogeneous genomic data, and obtained the combined representation as a linear combination of kernel matrices for a classification task. The problem of learning the linear combinations of kernel matrices in the context of classification is known as *Multiple kernel learning* (MKL), and has been further studied by Bach et al. (2004); Sonnenburg et al. (2006); Rakotomamonjy et al. (2008). In a recent study, Pasupa et al. (2009) have empirically shown that the task of image search can be improved by using a linear combination of image features with the features extracted from eye movements.

In both types of multi-view learning approaches shown in Figure 3.1, the underlying assumption is to maximize the consensus between different views using some definition of consensus. Multi-view learning in an unsupervised setting is considered in this thesis, and a criterion based on statistical dependency to define consensus between multiple views is proposed. Section 3.1 describes our approach of multi-view learning by modeling statistical dependency between views. The subsequent sections give an overview of various measures of dependency and the methods to model statistical dependency. Section 3.4 describes the unsupervised data fusion approach proposed in the Publication 1 and section 3.5 describes the novel evaluation approach to compare different vector representations for bilingual corpora proposed in the Publication 2.

3.1 Multi-view learning using statistical dependency

In this thesis, multi-view learning methods in an unsupervised setting when the data are represented by multiple independent views are discussed. One of the important questions in multi-view learning is how to define the agreement between the views. Different approaches have used different criteria to define the agreement. In a supervised setting, defining agreement is rather well defined. For instance, in a classification task, the classifiers based on different views should agree with each other (Yarowsky, 1995; Blum and Mitchell, 1998; Dasgupta et al., 2001). In an unsupervised learning, it is, however, not straightforward to define or search for agreement between views due to not having a clear hypothesis. Multi-view learning methods in this thesis use statistical dependency between the views to define the agreement between them. This approach is useful when the information shared between the views is more interesting than the information specific to any of the views.

The concept of using statistical dependency to learn from multiple views

has recently attracted the attention of researchers in many application areas, but it has not been fully matured yet. Nikkilä et al. (2005) used CCA to find dependency between expression measurements of yeast under different stress treatments in order to study the environmental stress response in yeast. Kernel CCA is used to detect dependencies between images and their annotations by Hardoon et al. (2004); Farquhar et al. (2006) for content-based image retrieval. Li and Shawe-Taylor (2006) used KCCA to learn semantic representation between documents in Japanese and English for cross-language information retrieval and document classification. In this thesis, the concept of learning based on statistical dependencies is formally developed, and applied to several learning tasks.

Unlike the assumption in the unsupervised multi-view approach by (Bickel and Scheffer, 2004), the multi-view approach in this thesis does not assume that each view is sufficient for the task, and do not model everything that is present in each view. Here, it is assumed that each view may have many kind of regularities or information and the information which is shared by all the views is more interesting or relevant. In this thesis, statistical dependency is used as a definition of what is shared between multiple views. This setting is slightly different from traditional multi-view learning in that each view may not be sufficient for learning, and may lead to misleading models. Thus, combining multiple views by maximizing statistical dependencies helps the learning task by complementing information from each view. Formally, we study multi-view learning methods by maximizing statistical dependencies between views, hence maximizing agreement between them.

Another important reason for modeling dependencies between the views is the noise reduction. In practice, the data may contain noise that could be either due to measurement error, or some other kind of variation. The noise can be assumed to be independent between the samples. In multi-view setting when the sets of features are independent, the noise can also be assumed to be independent between the views. Looking for dependencies between the views is analogous to averaging over the several views; instead of simply averaging over views a more general feature mapping based on dependency maximization is adopted for noise reduction.

All the multi-view approaches proposed in this thesis are based on maximizing statistical dependency between views. In the next Section 3.2, various measures of dependency are described, and methods to model statistical dependencies between multiple views are discussed in 3.3.

3.2 Measures of dependency

In this thesis, by dependency we mean the relationship between two or more random variables, or the deviation from the condition of independence. Two random variables $\mathbf{x}$ and $\mathbf{y}$ are independent if and only if their joint probability $p(\mathbf{x}, \mathbf{y})$ can be written as a product of their marginal probabilities, that is, $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. Note that the independence of random variables is a binary quantity, while the dependence between random variables is a continuous quantity, that is, there are different degrees of dependence. The notion of independence can be easily generalized to more than two random variables.

3.2.1 Mutual information

Mutual information quantifies the amount of information shared between two random variables. In other words, it represents the reduction of uncertainty about the value of one random variable due the knowledge of value of other random variable.

In case of discrete random variables, mutual information is defined as

$$I(\mathbf{X}, \mathbf{Y}) = \Sigma_{\mathbf{x}} \Sigma_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}, \tag{3.1}$$

where the summation is over all possible values of $\mathbf{X}$ and $\mathbf{Y}$. In case of continuous random variables, the summation is replaced by integrals and $p(\mathbf{x}, \mathbf{y})$ denotes the joint probability density. It is clear that if the variables are independent, that is $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, the mutual information becomes zero, and vice versa. Also, mutual information is symmetric.

Mutual information can be interpreted through the concepts of *entropy* and *conditional entropy*. Entropy is a measure of uncertainty of a random variable. If $H(\mathbf{X})$ is the entropy of $\mathbf{X}$, then by $H(\mathbf{X}|\mathbf{Y})$ we mean the conditional entropy which is a measure of uncertainty about $\mathbf{X}$, given $\mathbf{Y}$; equivalently, the measure of information required to describe $\mathbf{X}$, given $\mathbf{Y}$. Intuitively, mutual information can be expressed in terms of entropies,

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}), \tag{3.2}$$

which is the reduction of uncertainty about $\mathbf{X}$ given $\mathbf{Y}$. Equivalently, it can be expressed as,

$$\begin{aligned} I(\mathbf{X}, \mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &= H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) \end{aligned}$$

Another interpretation of mutual information is through the concept of Kullback-Leibler divergence (Kullback and Leibler, 1951), which is a natural measure of difference between two distributions and is defined as

$$d_{KL}(p,q) = \Sigma_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})},$$

where p and q are two distributions. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ can be expressed as $d_{KL}(p(\mathbf{x},\mathbf{y}), p(\mathbf{x})p(\mathbf{y}))$, where one distribution assumes the variables to be dependent, and the other does not.

In this thesis, mutual information is considered as a standard measure of statistical dependence. Due to finite size of sample, it is however not possible to get an exact and accurate estimation of mutual information. Hence, other measures of dependency which can be reliably computed from small sample size are also considered, though they might not correspond to mutual information.

3.2.2 Correlation

Pearson's correlation (Pearson, 1896) is a measure of linear dependence, or degree of association between two univariate random variables. It is defined as the ratio of covariance of two variables and product of their standard deviations,

$$\rho_{xy} = \frac{\mathrm{cov}(\mathbf{X},\mathbf{Y})}{\sigma_x \sigma_y}.$$

In practice, the covariance and variances in the formula are replaced by their sample estimates giving sample correlation coefficient. The value of correlation coefficient is between -1 and 1. The sign of correlation tells the nature of association, and the absolute value signifies the strength of association. The correlation of -1 or 1 means the variables are linearly dependent. If the variables are independent, the correlation is 0, but the converse is not always true. However, for the multivariate Gaussian distribution, the correlation is zero if and only if the variables are statistically independent.

Also, there is a strong relationship between correlation and mutual information for multivariate normal distributions. As shown in (Borga, 2001; Bach and Jordan, 2002), the mutual information and correlation for Gaussian random variables are related as,

$$I(\mathbf{X},\mathbf{Y}) = -\frac{1}{2}\log(1-\rho_{xy}^2).$$

Hence, correlation can be used as a measure of dependency without loss of generality for Gaussian variables. This relationship does not hold for other

distributions, and correlation should merely be regarded as a measure of linear relationship in that case.

3.2.3 Kernel measure of dependency

There has recently been a lot of interest in using kernel methods to measure dependence between random variables. Kernel methods allow capturing higher order moments using functions in reproducing kernel Hilbert spaces to measure dependence. The underlying idea is that if there are non-linear dependencies between two variables, mapping the variables into kernel space transforms the nonlinear dependency into linear dependency which can thus be captured with standard correlation. Figure 3.2 illustrates the mapping of data into a kernel space where the nonlinear pattern transforms into a linear pattern.

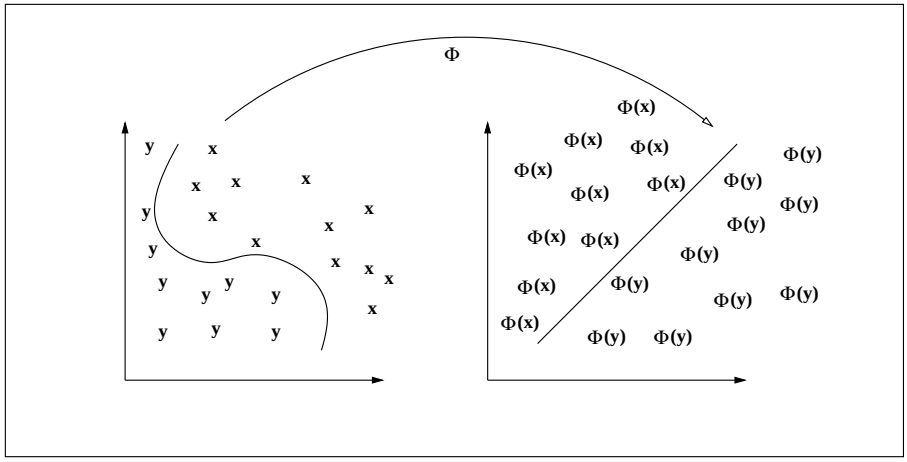

Figure 3.2: The mapping $\mathbf{\Phi}$ maps the data into a kernel space and transforms the nonlinear pattern into a linear pattern.

Rényi (1959) first suggested using the functional covariance or correlation to measure dependence of random variables. One such measure of statistical dependence between $\mathbf{x}$ and $\mathbf{y}$ can be defined as

$$\rho_{max} = \sup_{\mathbf{f},\mathbf{g}} \operatorname{corr}(\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x})), \tag{3.3}$$

where $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ have finite positive variance, and $\mathbf{f}$ and $\mathbf{g}$ are Borel measurable. This section gives a brief overview of kernel measures of dependency based on both the correlation and covariance operator.

Kernel canonical correlation (Bach and Jordan, 2002; Fukumizu et al., 2007; Leurgans et al., 1993) is a measure of dependency based on the correlation-operator. Bach and Jordan (2002) defined kernel canonical correlation (KCC) as a regularized spectral norm of the correlation-operator on reproducing kernel Hilbert spaces (RKHS), and showed that KCC can be empirically computed as a maximum eigenvalue solution to the generalized eigenvalue problem. Bach and Jordan (2002) extended the KCC to another measure of dependence, called *kernelized generalized variance* (KGV) by taking into account the whole spectrum of the correlation operator. While KCC is defined as maximum eigenvalue, KGV is defined in terms of the product of eigenvalues of the generalized eigenvalue problem. As shown in (Bach and Jordan, 2002), KGV approaches mutual information up to second order, expanding around independence.

Gretton et al. (2005b) proposed *constrained covariance* (COCO) based on covariance operator, which can again be empirically estimated as maximum eigenvalue solution to generalized eigenvalue problem. COCO is different from KCC in its normalization which is immaterial at independence. The regularization parameter in KCC is however not required in COCO, making it simpler yet equally good measure of dependency (Gretton et al., 2005b). COCO can be extended to *kernelized mutual information* (KMI) by taking into account the whole spectrum of the covariance operator and KMI is shown to be an upper bound near independence on a Parzen window estimate of the mutual information (Gretton et al., 2005b).

Another kernel measure of dependence based on a covariance operator is *Hilbert Schmidt Independence Criteria* (Gretton et al., 2005a; Fukumizu et al., 2008). The Hilbert Schmidt Independence Criteria (HSIC) is defined as the squared Hilbert-Schmidt norm of the entire eigen spectrum of the covariance operator, and the empirical estimate can be computed as the trace of the product of Gram matrices (Gretton et al., 2005a). While KGV and KMI depend on both the data distribution and the choice of kernel, Gretton et al. (2005a) showed that HSIC, in the limit of infinite data, depends only on the probability densities of variables assuming richness of the RKHSs, despite being defined in terms of kernel. The connection to mutual information is however not clear in the case of HSIC.

3.3 Maximization of mutual dependencies

This section describes the methods for modeling dependency between multivariate data. Canonical correlation analysis (CCA) and its kernel extension called kernel canonical correlation analysis (KCCA) are used to model de-

pendency in this thesis. The following subsections describe the two methods and discuss their properties.

3.3.1 Canonical correlation analysis

Canonical correlation analysis (Hotelling, 1936) is a classical method to find linear relationships between two sets of random variables. Given two random vectors $\mathbf{x}$ and $\mathbf{y}$ of dimensions D_x and D_y, CCA finds a pair of linear transformations such that one component within each set of transformed variables is correlated with a single component in the other set. The correlation between the corresponding components is called *canonical correlation*, and there can be at most $D = min(D_x, D_y)$ non-zero canonical correlations. The first canonical correlation is defined as: find linear transformations $\mathbf{x}^T\mathbf{w}_x$ and $\mathbf{y}^T\mathbf{w}_y$ such that the correlation between them is maximized,

$$\rho = \underset{\mathbf{w}_x,\mathbf{w}_y}{\text{argmax}}\, \text{corr}(\mathbf{x}^T\mathbf{w}_x, \mathbf{y}^T\mathbf{w}_y) \tag{3.4}$$

$$= \underset{\mathbf{w}_x,\mathbf{w}_y}{\text{argmax}}\, \frac{E[\mathbf{w}_x^T\mathbf{x}\mathbf{y}^T\mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T\mathbf{x}\mathbf{x}^T\mathbf{w}_x]E[\mathbf{w}_y^T\mathbf{y}\mathbf{y}^T\mathbf{w}_y]}}, \tag{3.5}$$

where ρ is the canonical correlation. The next canonical correlation can be computed recursively from the next pair of CCA components such that they are orthogonal to the previous pair of components, that is, $\langle \mathbf{a}^i, \mathbf{a}^j \rangle = \langle \mathbf{b}^i, \mathbf{b}^j \rangle = \delta_{ij}, i, j \in 1, \ldots, D$, where $\mathbf{a}^i = \mathbf{x}^T\mathbf{w}_x^i$, $\mathbf{b}^i = \mathbf{y}^T\mathbf{w}_y^i$ and $\langle \cdot, \cdot \rangle$ is the inner product of two vectors. In practice, the expectations in Eq. (3.5) are replaced by the sample-based estimates from the observation matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$. The samples $\mathbf{x}_i$ and $\mathbf{y}_i$ can be thought of as the measurements on N objects describing different views of these objects. Given sample-based estimates, Eq. (3.5) can be written as

$$\rho = \underset{\mathbf{w}_x,\mathbf{w}_y}{\text{argmax}}\, \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x}\sqrt{\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}}, \tag{3.6}$$

where $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between-set covariance matrix, and $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-set covariance matrices of random variables $\mathbf{x}$ and $\mathbf{y}$. The total covariance matrix is

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix}$$

. Note that the solution of Eq. (3.6) does not change by re-scaling $\mathbf{w}_x$, or $\mathbf{w}_y$ either separately or together, and hence the CCA optimization problem

in Eq. (3.6) can be solved by optimizing the numerator with respect to the conditions

$$\begin{aligned} \mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x &= 1, \\ \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y &= 1. \end{aligned}$$

As shown in (Hardoon et al., 2004), the corresponding Lagrangian is

$$Ł(\lambda, \mathbf{w}_x, \mathbf{w}_y) = \mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y - \frac{\lambda_x}{2}(\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x - 1) - \frac{\lambda_y}{2}(\mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y - 1),$$

which leads to the following eigenvalue problems

$$\begin{aligned} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{w}_x &= \lambda^2 \mathbf{w}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{w}_y &= \lambda^2 \mathbf{w}_y, \end{aligned}$$

assuming $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are invertible. It gives d positive eigenvalues $\lambda_1^2 \geq \cdots \geq \lambda_d^2$, and the canonical correlation is the square root of the eigen values, that is $\rho_i = \lambda_i$. We see that CCA reduces to the following generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{pmatrix} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = (1+\rho) \begin{pmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{pmatrix} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}, \tag{3.7}$$

which gives $D_x + D_y$ eigen values $1+\rho_1, 1-\rho_1, \ldots, 1+\rho_d, 1-\rho_d, 1, \ldots, 1$. Note that the problem of finding the maximal generalized eigenvalue $1 + \rho_{max}$ is equivalent to finding the minimal generalized eigenvalue $1 - \rho_{max}$, where ρ_{max} is the maximal canonical correlation. The quantity $1 - \rho_{max}$ is always bounded between zero and one, hence solving the minimal generalized eigenvalue problem provides a natural upgrade when extending CCA to more than two variables.

3.3.2 Kernel Canonical Correlation Analysis

CCA finds the linear relationship between two data sets using linear projections, but it is not able to capture non-linear relationships. Several extensions of CCA have been proposed that use non-linear projections to capture non-linear relationships. One of the approaches to extend CCA is to use kernel functions, called kernel canonical correlation analysis (Bach and Jordan, 2002; Hardoon et al., 2004; Kuss and Graepel, 2003). KCCA exploits the non-linear relationships by projecting the data onto a higher dimensional space before performing classical CCA; this process is known as the *kernel trick*. In this section, KCCA and its properties are briefly described.

Definition A *kernel* is a function k that for all $\mathbf{x}, \mathbf{z} \in X$ satisfies

$$k(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle,$$

where $\boldsymbol{\phi}$ is a mapping from X to a (inner product) feature space F

$$\boldsymbol{\phi} : \mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x}) \in F.$$

Let $\boldsymbol{\phi}_x : X \mapsto F_x$ and $\boldsymbol{\phi}_y : Y \mapsto F_y$ denote the feature space mappings with corresponding kernel functions $k_x(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}_x(\mathbf{x}_i), \boldsymbol{\phi}_x(\mathbf{x}_j) \rangle, \mathbf{x}_i, \mathbf{x}_j \in X$, and $k_y(\mathbf{y}_i, \mathbf{y}_j) = \langle \boldsymbol{\phi}_y(\mathbf{y}_i), \boldsymbol{\phi}_y(\mathbf{y}_j) \rangle, \mathbf{y}_i, \mathbf{y}_j \in Y$. Intuitively, performing KCCA is equivalent to performing CCA for variables in the feature space, that is, performing CCA on $\boldsymbol{\phi}_x(\mathbf{x})$ and $\boldsymbol{\phi}_y(\mathbf{y})$. Following the lines of Bach and Jordan (2002), the canonical correlation between $\boldsymbol{\phi}_x(\mathbf{x})$ and $\boldsymbol{\phi}_y(\mathbf{y})$ can be defined as

$$\rho = \underset{(f_x, f_y) \in F_x \times F_y}{\operatorname{argmax}} \operatorname{corr}(\langle \boldsymbol{\phi}_x(\mathbf{x}), f_x \rangle, \langle \boldsymbol{\phi}_y(\mathbf{y}), f_y \rangle), \tag{3.8}$$

where $f_x \in F_x$ and $f_y \in F_y$.

Given the samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$, the empirical estimate of Eq. (3.8) can be computed based on empirical covariances and the empirical canonical correlation is denoted as $\hat{\rho}$. The samples mapped to the feature spaces can be represented as $\boldsymbol{\Phi}_x = [\boldsymbol{\phi}_x(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_x(\mathbf{x}_N)]$ and $\boldsymbol{\Phi}_y = [\boldsymbol{\phi}_y(\mathbf{y}_1), \ldots, \boldsymbol{\phi}_y(\mathbf{y}_N)]$.

Given fixed f_x and f_y, the empirical covariance of the projections in feature spaces can be written as

$$\hat{\operatorname{cov}}(\langle \boldsymbol{\phi}_x(\mathbf{x}), f_x \rangle, \langle \boldsymbol{\phi}_y(\mathbf{y}), f_y \rangle) = \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{\phi}_x(\mathbf{x}_i), f_x \rangle \langle \boldsymbol{\phi}_y(\mathbf{y}_i), f_y \rangle. \tag{3.9}$$

Let S_x and S_y be the spaces linearly spanned by $\boldsymbol{\phi}_x(\mathbf{x}_i)$ and $\boldsymbol{\phi}_y(\mathbf{y}_i)$. The functions f_x and f_y can then be expressed as

$$\begin{aligned} f_x &= \sum_{i=1}^{N} \boldsymbol{\alpha}_x^{(i)} \boldsymbol{\phi}_x(\mathbf{x}_i) + f_x^{\perp} \\ f_y &= \sum_{i=1}^{N} \boldsymbol{\alpha}_y^{(i)} \boldsymbol{\phi}_y(\mathbf{y}_i) + f_y^{\perp}, \end{aligned}$$

where $f_x^{\perp}$ and $f_y^{\perp}$ are orthogonal to S_x and S_y, respectively, and $\boldsymbol{\alpha}_x, \boldsymbol{\alpha}_y \in \mathbb{R}^N$. Equation (3.9) becomes

$$\begin{aligned}
&\widehat{\mathrm{cov}}(\langle \boldsymbol{\phi}_x(\mathbf{x}), f_x \rangle, \langle \boldsymbol{\phi}_y(\mathbf{y}), f_y \rangle) \qquad (3.10)\\
&= \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{\phi}_x(\mathbf{x}_i), \sum_{j=1}^{N} \alpha_x^{(i)} \boldsymbol{\phi}_x(\mathbf{x}_j) \rangle \langle \boldsymbol{\phi}_y(\mathbf{y}_i), \sum_{q=1}^{N} \alpha_y^{(q)} \boldsymbol{\phi}_y(\mathbf{y}_q) \rangle \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{q=1}^{N} \alpha_x^{(i)} \mathbf{K}_x(\mathbf{x}_i, \mathbf{x}_j) \mathbf{K}_y(\mathbf{y}_i, \mathbf{y}_q) \alpha_y^{(q)} \\
&= \frac{1}{N} \boldsymbol{\alpha}_x \mathbf{K}_x \mathbf{K}_y \boldsymbol{\alpha}_y, \qquad (3.11)
\end{aligned}$$

where $\mathbf{K}_x$ and $\mathbf{K}_y$ are *Gram matrices* for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Similarly, we can compute within-set variances,

$$\begin{aligned}
\hat{\mathrm{var}}(\langle \boldsymbol{\phi}_x(\mathbf{x}) \rangle) &= \frac{1}{N} \boldsymbol{\alpha}_x \mathbf{K}_x \mathbf{K}_x \boldsymbol{\alpha}_x, \\
\hat{\mathrm{var}}(\langle \boldsymbol{\phi}_y(\mathbf{y}) \rangle) &= \frac{1}{N} \boldsymbol{\alpha}_y \mathbf{K}_y \mathbf{K}_y \boldsymbol{\alpha}_y.
\end{aligned}$$

The KCCA problem in Eq. (3.8) can be represented as

$$\hat{\rho} = \underset{\boldsymbol{\alpha}_x, \boldsymbol{\alpha}_y \in \mathbb{R}^N}{\mathrm{argmax}} \frac{\frac{1}{N} \boldsymbol{\alpha}_x \mathbf{K}_x \mathbf{K}_y \boldsymbol{\alpha}_y}{(\frac{1}{N} \boldsymbol{\alpha}_x \mathbf{K}_x^2 \boldsymbol{\alpha}_x)^{\frac{1}{2}} (\frac{1}{N} \boldsymbol{\alpha}_y \mathbf{K}_y^2 \boldsymbol{\alpha}_y)^{\frac{1}{2}}}. \qquad (3.12)$$

This is equivalent to performing CCA on data whose covariance matrices are the Gram matrices $\mathbf{K}_x$ and $\mathbf{K}_y$. Hence, Eq. (3.12) can be formulated as a generalized eigenvalue problem

$$\begin{pmatrix} 0 & \mathbf{K}_x \mathbf{K}_y \\ \mathbf{K}_y \mathbf{K}_x & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix} = \rho \begin{pmatrix} \mathbf{K}_x^2 & 0 \\ 0 & \mathbf{K}_y^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_\mathbf{y} \end{pmatrix}. \qquad (3.13)$$

We assume that the Gram matrices are computed over centered data. If $\mathbf{K}_x$ and $\mathbf{K}_y$ are Gram matrices of non-centered data, it is possible to compute Gram matrices of centered data as shown in (Schölkopf et al., 1998).

3.3.3 Regularization of (K)CCA

The KCCA implementation by applying CCA on Gram matrices has certain drawbacks. As shown in (Bach and Jordan, 2002; Hardoon et al., 2004), if the Gram matrices $\mathbf{K}_x$ and $\mathbf{K}_y$ are invertible, the KCCA will return a canonical correlation that is always one irrespective of what the Gram matrices are. The naive kernelization of CCA in Eq. (3.13) leads to trivial

learning in general. In order to avoid trivial learning, the norms of f_x and f_y are penalized to control the flexibility of the projection mappings. The regularized version of KCCA can be equivalently defined as

$$\underset{\boldsymbol{\alpha}_x,\boldsymbol{\alpha}_y\in\mathbb{R}^N}{\operatorname{argmax}} \frac{\frac{1}{N}\boldsymbol{\alpha}_x\mathbf{K}_x\mathbf{K}_y\boldsymbol{\alpha}_y}{(\frac{1}{N}\boldsymbol{\alpha}_x(\mathbf{K}_x+\frac{N\kappa}{2}\mathbf{I})^2\boldsymbol{\alpha}_x)^{\frac{1}{2}}(\frac{1}{N}\boldsymbol{\alpha}_y(\mathbf{K}_y+\frac{N\kappa}{2}\mathbf{I})^2\boldsymbol{\alpha}_y)^{\frac{1}{2}}}, \tag{3.14}$$

where κ is a small positive quantity. The optimization problem for regularized KCCA can be formulated as the following generalized eigenvalue problem

$$\begin{pmatrix} 0 & \mathbf{K}_x\mathbf{K}_y \\ \mathbf{K}_y\mathbf{K}_x & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_y \end{pmatrix} = \rho\begin{pmatrix} (\mathbf{K}_x+\frac{N\kappa}{2}\mathbf{I})^2 & 0 \\ 0 & (\mathbf{K}_y+\frac{N\kappa}{2}\mathbf{I})^2 \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha}_x \\ \boldsymbol{\alpha}_\mathbf{y} \end{pmatrix}. \tag{3.15}$$

The regularization parameter κ, in addition to inducing control of overfitting, also enhances the numerical stability of the solution. It has been shown that regularized KCCA generalizes better than the naive KCCA implementation (Bach and Jordan, 2002; Hardoon et al., 2004).

The similar approach of regularization can also be used in linear CCA, where the sample canonical correlation heavily depends on the number of samples and dimensionality of random variables. A standard assumption in classical CCA is that $N \gg D_x + D_y$. When the ratio $\frac{N}{D_x+D_y}$ is small, the sample estimate of covariance matrix of the vectors $(\mathbf{x}_i, \mathbf{y}_i), i \in 1,\ldots,N$ may be ill-conditioned which leads to the trivial CCA solution. Also, within-set covariance matrices $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are assumed to be invertible in Eq. (3.7), due to high dimensionality they can however be singular or near-singular which leads to unreliable estimates of their inverses. In (Leurgans et al., 1993), a regularized version of CCA is proposed based on smoothing the constraints $\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x = 1$ and $\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y = 1$ by adding a roughness penalty which is similar to the ridge-regression type regularization proposed by (Bie and Moor, 2003); the regularized version can be formulated as an eigenvalue problem by adding a small positive quantity γ to the diagonals to $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$

$$\begin{pmatrix} 0 & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \rho\begin{pmatrix} \mathbf{C}_{xx}+\gamma\mathbf{I} & 0 \\ 0 & \mathbf{C}_{yy}+\gamma\mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}. \tag{3.16}$$

3.3.4 Generalized canonical correlation analysis

Canonical correlation analysis, originally proposed for two multivariate random variables (Hotelling, 1936), can be extended to three or more sets. Kettenring (1971) comprehensively studied different generalizations

of CCA which were similar to the classical CCA in two ways: they reduce to Hotelling's classical CCA when considered for two sets of random variables, and they find canonical variables, one from each set, to optimize some function of their correlation matrix. More recently, Bach and Jordan (2002); Hardoon et al. (2004) described the generalization of (K)CCA to more than two random variables.

Bach and Jordan (2002) formulated the generalization as an eigenvalue problem analogous to classical CCA. Suppose $\mathbf{X}_i, i \in 1, \ldots, m$ denote a collection of m data sets, where each $\mathbf{X}_i$ is a matrix of size $N \times D_i$ such that $N \gg \sum_i D_i$. The generalized CCA is defined as

$$\begin{pmatrix} \mathbf{C}_{11} & \ldots & \mathbf{C}_{1m} \\ \vdots & & \vdots \\ \mathbf{C}_{m1} & \ldots & \mathbf{C}_{mm} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_m \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_{11} & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & \mathbf{C}_{mm} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_m \end{pmatrix} \quad (3.17)$$

which is a generalized eigenvalue problem of the form $\mathbf{Cw} = \lambda \mathbf{Dw}$, similar to the two-variables case in Eq. (3.7). Here, $\mathbf{C}$ is a covariance matrix of the column-wise concatenation of matrices $\mathbf{X}_i$, and $\mathbf{D}$ is a block-diagonal matrix with $\mathbf{C}_{ii}$, a within-set covariance matrix for each $\mathbf{X}_i$, as a diagonal element. Unlike the two-variable case, the connection between a generalized eigenvalue and the canonical correlation is not clear in the case of more than two variables. Also, the eigenvalues $\lambda_i \geq 0$ do not appear in pairs as in the case of CCA with two random variables. As shown by Bach and Jordan (2002), the minimal generalized eigenvalue is bounded between $[0, 1]$; analogous to the two-variable case we can define the first canonical correlation as the minimal generalized eigenvalue of $\mathbf{Cw} = \lambda \mathbf{Dw}$ in the case of more than two variables. Note that the maximal generalized eigenvalue depends on the size of the matrices $\mathbf{X}_i$. The generalized CCA can be regularized in the same way as in the two-variable case, by replacing $\mathbf{C}_{ii}$ by $\mathbf{C}_{ii} + \kappa \mathbf{I}$ in Eq (3.17) for all $i \in 1, \ldots, m$.

Generalization of kernel CCA to more than two variables can be described as above. Let $\mathbf{K}_i$ be the Gram kernel matrix corresponding to the data matrix $\mathbf{X}_i \in \mathbb{R}^{N \times D_i}$ given the feature space mapping $\boldsymbol{\phi}_i : X_i \mapsto F_i$, where $i \in 1, \ldots, m$. Let C_k denote the $mN \times mN$ matrix with blocks $C_k^{i,j} = \mathbf{K}_i \mathbf{K}_j$, and let D_k denote a block-diagonal matrix with diagonal elements $\mathbf{K}_i^2$. Analogous to KCCA in case of two variables, generalized KCCA can be formulated as an generalized eigenvalue problem

$$C_k \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha} D_k. \quad (3.18)$$

The *first kernel canonical correlation* can be defined as the minimal generalized eigenvalue, which is always bounded between zero and one; and it

is zero if and only if the m variables are pairwise independent (Bach and Jordan, 2002). The regularized version of generalized KCCA can be defined in the same way as the two-variable case by replacing $\mathbf{K}_i^2$ with $(\mathbf{K}_i + \frac{N\kappa}{2}\mathbf{I})^2$ in Eq. (3.18).

3.3.5 Properties of CCA

Canonical correlation has a very natural connection to mutual information if the data are Gaussian. In this section, the connection between CCA and mutual information in the case of two variables is described, and then it is extended to the generalized CCA. The relationship of kernel canonical correlation to the mutual information is also briefly discussed.

Let $\mathbf{x}$ and $\mathbf{y}$ be two random variables with dimensionality D_x and D_y respectively. As explained in Section 3.2.1, the mutual information between $\mathbf{x}$ and $\mathbf{y}$ can be defined in terms of Kullback-Leibler divergence (Kullback, 1959),

$$I(\mathbf{x}, \mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) \log \left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right) dx\, dy,$$

which can, for the normally distributed $\mathbf{x}$and $\mathbf{y}$, be easily computed as:

$$I(\mathbf{x}, \mathbf{y}) = -\frac{1}{2} log \left(\frac{det(\mathbf{C})}{det(\mathbf{C}_{xx})\, det(\mathbf{C}_{yy})} \right), \tag{3.19}$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are within-set covariance of random Gaussian variables $\mathbf{x}$ and $\mathbf{y}$ respectively, and $\mathbf{C}$ is covariance matrix of the the vector $(\mathbf{x}, \mathbf{y})$. The term $\frac{det(\mathbf{C})}{det(\mathbf{C}_{xx})\, det(\mathbf{C}_{yy})}$ is called *generalized variance*.

Now, the generalized eigenvalue problem for the two-variable case in Eq. 3.7 is of the form $\mathbf{Cx} = \lambda \mathbf{Dx}$. If $\mathbf{D}$ is invertible, then it can be written equivalently as an eigenvector problem $\mathbf{D}^{-1}\mathbf{Cx} = \lambda \mathbf{x}$. Thus, the product of generalized eigenvalues in Eq. 3.7 is equal to the ratio of determinants in Eq. 3.19, and we get

$$I(\mathbf{x}, \mathbf{y}) = -\frac{1}{2} \log \prod_{i=1}^{D} (1 + \rho_i)(1 - \rho_i) = -\frac{1}{2} \sum_{i=1}^{D} \log(1 - \rho_i^2), \tag{3.20}$$

where ρ_i is the canonical correlation and $D = min(D_x, D_y)$. Thus, in case of Gaussian variables, the mutual information can be computed using canonical correlation obtained from CCA.

The relation between CCA and mutual information can be easily extended to the generalized CCA for more than two variables. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$

be m random Gaussian variables of size $D_i, i \in 1, \ldots, \mathbf{m}$ respectively. The mutual information can be written in terms of generalized variance as

$$I(\mathbf{x}_1, \ldots, \mathbf{x}_m) = -\frac{1}{2} \log \left(\frac{det\mathbf{C}}{det\mathbf{C}_{11} \ldots det\mathbf{C}_{mm}} \right). \tag{3.21}$$

The generalized eigenvalue problem in Eq. 3.17 for CCA of more than two random variables can also be written as a generalized eigenvector problem $\mathbf{D}^{-1}\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$, if $\mathbf{D}$ is invertible which is a block-diagonal matrix with $\mathbf{C}_{ii}$ as elements. The generalized variance can again be defined as the ratio of determinants of matrices $\mathbf{C}$ and $\mathbf{D}$. Thus, we get

$$I(\mathbf{x}_1, \ldots, \mathbf{x}_m) = -\frac{1}{2} \sum_{i=1}^{i=D} \log(\lambda_i), \tag{3.22}$$

where λ_i are the generalized eigen values of $\mathbf{C}\mathbf{w} = \lambda\mathbf{D}\mathbf{w}$. The $I(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is known as multi-information.

Given the connection of generalized variance to mutual information in the case of Gaussian variable, Bach and Jordan (2002) defined the concept of kernel generalized variance(KGV), also discussed in the Section 3.2.3, as the product of the eigenvalues of the generalized eigenvalue problem for regularized KCCA as in Eq. 3.18, or equivalently as the ratio of determinants of its matrices

$$\text{KGV} = \frac{det\, C_k}{det\, D_k}.$$

The multi-information in the kernel case can be defined as

$$\hat{I}_k(\mathbf{K}_1 \ldots, \mathbf{K}_m) = -\frac{1}{2} \log(\text{KGV}), \tag{3.23}$$

which has its population counterpart $I_k(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ that is actually closely related to the mutual information between the original non-Gaussian variables in the input space.

In this thesis, classical CCA is used to compute the mutual information in the Gaussian case. In the case of kernel CCA, the quantity $\hat{I}_k(\mathbf{K}_1 \ldots, \mathbf{K}_m)$ is used as an approximation to mutual information.

3.3.6 Probabilistic CCA

In this section, probabilistic interpretation of CCA is briefly discussed. The section is concluded by listing advantages and drawbacks of using probabilistic models in the context of CCA.

Bach and Jordan (2005); Bie and Moor (2003) introduced the probabilistic interpretation of CCA by assuming an underlying generative model for the data. Bach and Jordan (2005) presented probabilistic CCA by extending the logic of probabilistic PCA presented by Tipping and Bishop (1999). The probabilistic PCA was explained as a factor analysis model with isotropic covariance matrix, graphical model of factor analysis is shown in Figure 3.3.

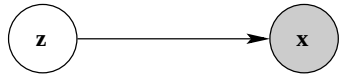

Figure 3.3: Graphical model for factor analysis.

Suppose the data matrix $\mathbf{X} \in \mathbb{R}^{M \times \mathbf{D}_x}$ represents M i.i.d samples for random vector $\mathbf{x}$. PCA is concerned with finding a linear transformation $\mathbf{A} \in \mathbb{R}^{D_x \times D}$ that makes the data uncorrelated with marginal unit variances. Tipping and Bishop (1999) showed that the posterior expectation of $\mathbf{z}$ given $\mathbf{x}$ based on the maximum likelihood estimates of the parameters $\mathbf{W}$, $\boldsymbol{\mu}$ and σ^2 for the model represented by Figure 3.3,

$$\begin{aligned} \mathbf{z} &= N(0, \mathbf{I}_D) \\ \mathbf{x}|\mathbf{z} &= N(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_D), \sigma > 0, \mathbf{W} \in \mathbb{R}^{M \times D} \end{aligned}$$

will yield the same linear subspace as PCA. Here, σ is a variance parameter, and $\boldsymbol{\mu}$ is a mean parameter.

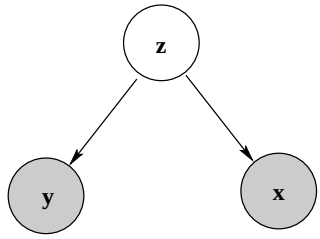

Figure 3.4: Graphical model for probabilistic CCA.

Similar to probabilistic PCA, Bach and Jordan (2005) gave a probabilistic interpretation of CCA as a latent variable model for two Gaussian random variables $\mathbf{x} \in \mathbb{R}^{1 \times D_x}$ and $\mathbf{y} \in \mathbb{R}^{1 \times D_y}$. Figure 3.4 shows the graphi-

cal model for probabilistic CCA, and the model can be described as follows

$$\begin{aligned}\mathbf{z} &= N(0, \mathbf{I}_D),\ \min D_x, D_y \geq D \geq 1\\ \mathbf{x}|\mathbf{z} &= N(\mathbf{W}_x\mathbf{z} + \boldsymbol{\mu}_x, \boldsymbol{\Psi}_x)\\ \mathbf{y}|\mathbf{z} &= N(\mathbf{W}_y\mathbf{z} + \boldsymbol{\mu}_y, \boldsymbol{\Psi}_y)\end{aligned}$$

where $\mathbf{W}_x$ and $\mathbf{W}_y$ are matrices of suitable dimensionalities, and μ and $\boldsymbol{\Psi}$ represent mean and covariance parameters for each variable. It is shown in (Bach and Jordan, 2005) that the maximum likelihood estimates of the parameters $\mathbf{W}_x$, $\mathbf{W}_y$, $\boldsymbol{\mu}_x$,$\boldsymbol{\mu}_y$, $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are given by

$$\begin{aligned}\hat{\mathbf{W}}_x &= \mathbf{C}_{xx}\mathbf{U}_x\mathbf{M}_x\\ \hat{\mathbf{W}}_y &= \mathbf{C}_{yy}\mathbf{U}_y\mathbf{M}_y\\ \hat{\boldsymbol{\Psi}}_x &= \mathbf{C}_{xx} - \hat{\mathbf{W}}_x\hat{\mathbf{W}}_x^T\\ \hat{\boldsymbol{\Psi}}_y &= \mathbf{C}_{yy} - \hat{\mathbf{W}}_y\hat{\mathbf{W}}_y^T\\ \hat{\boldsymbol{\mu}}_x &= \tilde{\boldsymbol{\mu}}_x\\ \hat{\boldsymbol{\mu}}_y &= \tilde{\boldsymbol{\mu}}_y\end{aligned}$$

Here, $\mathbf{C}_{xx}$ and $\tilde{\boldsymbol{\mu}}_x$ are sample covariance and mean of variable $\mathbf{x}$ given M samples $\mathbf{X} \in \mathbb{R}^{M\times\mathbf{D}_x}$; similarly for the variable $\mathbf{y}$. The $\mathbf{M}_x, \mathbf{M}_y \in \mathbb{R}^{D\times D}$ are arbitrary matrices with spectral norms smaller than one and $\mathbf{M}_x\mathbf{M}_y^T = \mathbf{P}_d$, which is a diagonal matrix of first D canonical correlations, and columns of $\mathbf{U}_x, \mathbf{U}_y$ contain corresponding canonical directions. The detailed proof can be found in the original publication (Bach and Jordan, 2005).

Probabilistic CCA has recently been studied further in many publications. Archambeau et al. (2006) presented an extension of probabilistic CCA using Student-t noise distribution; replacing Gaussian distributions with Student-t distributions makes the model more robust to outliers and atypical observations. Recently, Viinikanoja et al. (2010) extended it by presenting a Bayesian solution to the problem. In (Klami and Kaski, 2006, 2007, 2008), the probabilistic interpretation of CCA is motivated through a generative latent variable model to detect dependencies between two data sets. While Klami and Kaski (2006, 2008) used an expectation-maximization (EM) algorithm to compute maximum likelihood, Klami and Kaski (2007) presented a Bayesian treatment using Gibbs sampling. Also, a variational Bayes approach to solving probabilistic CCA has been proposed by Wang (2007).

The probabilistic CCA has certain advantages over traditional CCA. First, the probabilistic approach provides a better understanding of CCA

as a model-based method, which makes it possible to use CCA as a component in larger model families, for example as a part of a hierarchical model consisting of smaller probabilistic models. It also leads to generalizations of CCA to the members of exponential family other than the Gaussian distribution (Klami et al., 2010). Another advantage of probabilistic CCA is the possibility of using Bayesian methods which also provide a natural way of regularization by assigning prior probabilities to the model parameters. The quality of different solutions can be characterized through the likelihood given the observed data.

Despite all the advantages of having a probabilistic model, there are a few concerns in using probabilistic CCA. First, directly solving the eigenvalue problem in Section 3.3.1 is usually faster and guaranteed to have a global optimum, while probabilistic CCA is relatively slower. Furthermore, existing solutions to probabilistic CCA do not uniquely determine the projections and are invariant to arbitrary rotation which, in turn, may cause problems while interpreting the solutions. Second, the probabilistic approach as such models everything in the data which in a way violates the very basic essence of CCA that is to find what is common in data and ignore dataset-specific variation. As long as the task is to find the correlating subspaces, maximizing correlation does not seem to make any explicit assumption on the data and can hence be applied to more complex data. Probabilistic CCA makes explicit assumptions on the data distribution. The recent work by Klami et al. (2010) is a step forward to making probabilistic CCA work in a more general set up. Klami et al. (2010) also proposed a novel sampler that explicitly marginalizes out the components specific to data sets but still assumes normal distribution. Hence, probabilistic CCA models needs to be developed further in order to be applied in more complex data sets. In this thesis, the traditional eigenvalue problem is used for solving CCA.

3.4 Data fusion by maximizing dependencies

Data fusion, also known as *sensor fusion*, is a task of combining several data sources in order to improve the data analysis performance. In a supervised setting when it is possible to make sufficient modeling assumptions, data fusion is in principle straightforward. For instance in classification the task is to combine data sources such that the classification accuracy is improved (Girolami and Zhong, 2007; Lanckriet et al., 2004). Although the task is rather well defined in supervised settings, there are still practical challenges in this task.

This thesis focuses on the problem of unsupervised data fusion when the hypotheses are still vague and it is not straightforward how to combine data sources. An unsupervised data fusion approach is proposed in Publication 1. The proposed approach combines data sources by maximizing mutual dependencies between them such that the shared aspects between them is preserved. This kind of approach is useful in cases where the information shared by data sources is more interesting than the information specific to data sources.

In Publication 1, it has been shown that CCA can be used as a tool for data fusion. We assume vector-valued data sources such that each of them consists of measurements of the same entity but on different variables, that is, each source represents a different view on the data. Being a linear approach, CCA-based data fusion is simple, fast and easily interpretable, and provides a new way for dimensionality reduction. It also creates a direction towards building more complex data fusion models based on the idea of dependency maximization.

An alternative and intuitive approach to CCA can be described as a two-step procedure: The first step is to pre-process each data source separately to remove all within-source variation, and the second step is to extract the variation remained in the collection of all data sources. The first step makes sure that if the data sources are not dependent, the method in second step will not extract any information at all. Thus, any information extracted in the second step is due to the dependencies between data sources. The proposed data fusion approach is based on this two-step procedure for CCA.

In order to perform the first step, that is, to remove all within-source information, a traditional procedure called *whitening* is used. It linearly transforms the data source such that the transformed data has a unit covariance matrix. In practice, the whitening can be performed as

$$\bar{\mathbf{X}} = \mathbf{C}_x^{-\frac{1}{2}}\mathbf{X},$$

where $\mathbf{C}_x$ and $\bar{\mathbf{X}}$ denote the covariance and whitened version of $\mathbf{X}$, respectively. In $\bar{\mathbf{X}}$, all dimensions have equal unit variance and are uncorrelated. The second step is performed by applying principal component analysis (PCA) to the column-wise concatenation of all the whitened data sources. PCA (Hotelling, 1933) is a classical linear method to find projections of maximal variance. Given that all the within-source variation is removed, PCA can only find the variation which is shared between the data sources. Figure 3.5 illustrates the two step procedure to data fusion.

Suppose $\mathbf{Z} = \left[\bar{\mathbf{X}}_1, \ldots \bar{\mathbf{X}}_p\right]$ is the column-wise concatenation of p whitened data sets, where $\bar{\mathbf{X}}_i \in \mathbb{R}^{N \times \mathbf{D}_i}$. Applying PCA to $\mathbf{Z}$ will yield the factoriza-

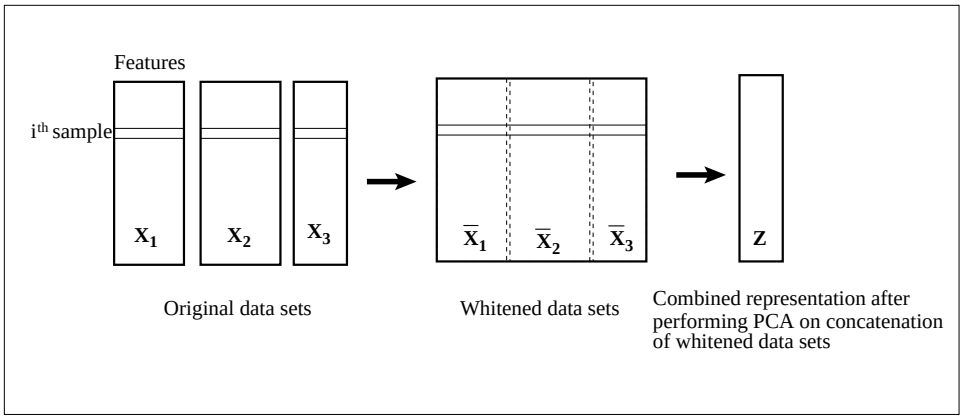

Figure 3.5: Illustration of data fusion as a two-step procedure that preserves the shared information between the data sources. The first step whitens each data source. Second step performs PCA on the column-wise concatenation of whitened data sources in order to create a combined representation.

tion

$$\mathbf{C}_z = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{3.24}$$

where $\mathbf{C}_z$ is the covariance matrix of $\mathbf{Z}$, the orthonormal matrix $\mathbf{V}$ contains the eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of projection variances. The combined representation can be obtained by projecting $\mathbf{Z}$ onto the first D eigenvectors $\mathbf{V}_D$ corresponding to the D largest eigenvalues

$$\mathbf{F}_D = \mathbf{Z}\mathbf{V}_D, \tag{3.25}$$

where $\mathbf{F}_D \in \mathbb{R}^{N \times D}$ is the fused representation.

Given the concatenation of non-independent whitened data sources, the PCA directions of highest variations must correspond to the dependencies between different preprocessed data sources. This is equivalent to modeling the mutual dependencies between the data sources. An alternative approach would be to model all the information by applying PCA on the concatenation of the original data sources, but that would not capture the dependencies between the data sources.

The two-step procedure described here is equivalent to applying classical CCA on the data sources. The equivalence of the two procedures has been shown in the Publication 1. Applying CCA to the data sources gives separate representations for each data source in the form of canonical variates. As shown in Publication 1, the final data fusion solution can be

obtained by summing the canonical variates as

$$\mathbf{F}_D = \mathbf{X}\mathbf{W}_x + \mathbf{Y}\mathbf{W}_y, \tag{3.26}$$

where $\mathbf{W}_x \in \mathbb{R}^{D_X \times D}$ and $\mathbf{W}_y \in \mathbb{R}^{D_y \times D}$ are the CCA projection matrices of the chosen dimensionality, D. The dimensionality of the combined representation $\mathbf{F}_D$ is $\mathbb{R}^{N \times D}$. The result can be easily generalized to more than two data sources as shown in the Publication 1.

The remaining task is to choose the optimal dimensionality for the projections. The total dimensionality of the combined representation will be the sum of dimensionalities of all data sources. In practice, the first few dimensions will, however, contain most of the shared information, and the rest of the dimensions may just represent noise. Intuitively, the optimal number of dimensions should be high enough to keep most of the shared information and low enough to avoid overfitting.

In Publication 1, a method based on randomization to choose the optimal dimensionality of projections is proposed. The method works by increasing the dimensionality one at a time by testing that the new dimension captures the shared variation. The randomization test compares the shared variance along the new dimension to the shared variance we would get under the null-hypothesis of mutual independency. The final dimensionality is detected when the shared variance does not differ significantly from the null-hypothesis. The details of the randomization test can be found in Publication 1.

Parallel and serial feature combination

Using CCA to combine data sets can be explained as follows: Find a new feature representation for each source that are mutually informative of each other, and then combine the extracted features together to get a fused representation. There are two ways in which the extracted features can be combined, namely, parallel combination and serial combination.

Suppose $\mathbf{u}_x = \mathbf{X}\mathbf{w}_x$ and $\mathbf{u}_y = \mathbf{Y}\mathbf{w}_y$ are the extracted features for $\mathbf{X}$ and $\mathbf{Y}$ respectively, where $\mathbf{w}_x$ and $\mathbf{w}_y$ represents a pair of CCA components. In parallel combination, the combined representation can be obtained by simply adding the two feature vectors, $\mathbf{f} = \mathbf{u}_x + \mathbf{u}_y$, where $\mathbf{f} \in \mathbb{R}^{N \times 1}$. If we use the first D pairs of CCA components, the combined representation will also be D-dimensional as shown in Eq. 3.26.

In serial combination, the features are serially concatenated to form a supervector. Using the first D pairs of CCA components, the serially combined feature representation is

$$\mathbf{F}_D = \begin{bmatrix} \mathbf{X}\mathbf{W}_x & \mathbf{Y}\mathbf{W}_y \end{bmatrix}, \tag{3.27}$$

where $\mathbf{F}_D \in \mathbb{R}^{N \times 2D}$. The dimensionality of the final representation in the case of serial combination is the sum of the dimensionalities of the projections for each data set.

In this thesis, the parallel combination of features is used for data fusion, and it comes naturally while projecting the concatenation of whitened data set onto the first D eigenvectors as in Eq. 3.25, which is equivalent to adding the CCA projections in Eq. 3.26. Peng et al. (2010) compared the parallel and serial combination of features for classification purposes using various CCA models including new CCA algorithms *Local Discrimination CCA* (LDCCA) and its kernel version KLDCCA.

Although the main task in (Peng et al., 2010) was to compare (K)LDCCA against other methods like locality preserving CCA (LPCCA), KCCA, CCA and SVM-2K (Farquhar et al., 2006), the classification accuracy in the experiment was also computed for both of the combinations strategies, parallel and serial. The results in Peng et al. (2010) show that the performance of both fusion methods was comparable. While parallel combination (94.81%, classification accuracy) slightly outperforms serial combination (93.33%) in KLDCCA, the serial combination (93.61%) was slightly better than the parallel (91.99%) in LDCCA. For other methods, the difference between the serial and parallel feature combinations was small, and each time serial combination was little better than the parallel combination. The numbers for the classification accuracies are borrowed from Peng et al. (2010).

Based on the results in Peng et al. (2010), it can be concluded that any strategy to combine features can be chosen because the differences in performances are small. The parallel combination should, however, be preferred because the dimensionality of combined representation is lower in the parallel combination than the serial combination.

Regularizing CCA through whitening matrices

As explained in Section 3.3.3, CCA overlearns in case of high-dimensional data and there are ways to regularize it by adding a small quantity to the diagonal of variance matrix or equivalently, through ridge-regression type regularization. In Publication 1, a different regularization based on the whitening matrix is used. The whitening matrix, $\mathbf{C}^{-\frac{1}{2}}$, is the square root of the inverse of corresponding covariance matrix $\mathbf{C}$. This approach, in our opinion, is a novel contribution in the context of regularizing CCA.

The problem in the form of numerical instability arises due to singular or near-singular covariance matrix during its inversion. The regularization methods described in Section 3.3.3 add a small quantity in the diagonal

of covariance matrix to avoid singularity. The proposed approach ignores the dimensions with zero or near-zero contribution to the total variance of the the covariance matrix during matrix inversion. The dimensions in covariance matrix that do not contribute significantly to the total variance can also be seen as less informative or noise. In practice, a threshold is defined as a proportion of contribution to the total variance, and the dimensions which collectively contribute less than the defined threshold are ignored during inversion. This regularization strategy is implemented in the R-package for data fusion released along the Publication 1.

3.5 Evaluating sentence representation using CCA

The representation of documents in a vector space model (VSM) is one of the fundamental tasks in information retrieval (IR) (Salton et al., 1975). Vector space model represents each document as a feature vector where each feature is a term, for instance, a word. The performance in many information retrieval tasks, such as ranking of documents based on a query and document classification, depends on how well the VSM represents the documents. The parameterization of VSM involves a number of crucial choices, for instance, the type of distributional information to construct the vector space (Lavelli et al., 2004), weighting and normalization strategies (Nakov et al., 2001) and the dimensionality reduction method for the space (Bingham and Mannila, 2001). Although the choice of parameters in a VSM has been widely studied, methodologies to evaluate different vector representations are still lacking. In this thesis, a novel evaluation approach to compare vector space models for sentence-aligned bilingual corpora is presented.

Consider a situation where we want to find an optimal vector representation for a given task, for instance, mate retrieval as a cross-language IR task (Vinokourov et al., 2003a). A simple approach is to compare all possible vector representations based on the performance in the task at hand, and choose the vector representation which gives the best performance. Such approaches are called indirect evaluation (Sahlgren, 2006). The indirect evaluations are rather time consuming, if the application setting is complex. Hence, it is not possible to compare a large number of parameterizations using indirect approaches. A direct evaluation method is needed which can quickly compare a large number of parameterizations to evaluate the goodness of different representations for a given application.

In Publication 2, an evaluation method based on CCA to compare different vector representations of sentences in parallel bilingual corpora is

presented. The parallel documents in two languages can be seen as two different views of the same underlying semantics. It is assumed that a good vector representation should reflect the underlying semantics. For example, vector representation for sentence-aligned parallel documents should capture the meaning of the sentences. The proposed evaluation method uses statistical dependency between parallel documents to capture the underlying semantics. The idea is to compute the linear dependency between parallel documents using CCA for different vector representations, and choose the one with highest linear dependency.

KCCA has already been used to infer semantic representation between multimodal information sources (Vinokourov et al., 2003a; Hardoon et al., 2004). Given feature spaces for parallel documents in two languages, Vinokourov et al. (2003a) used KCCA to infer semantic representations by finding subspaces that are maximally correlated. They assumed that any correlation between two feature spaces is due to the underlying semantic similarity. In this work, the correlation, that is, the underlying semantic similarity, is used as an evaluation measure to compare different parameterizations of feature spaces for the bilingual corpora. In Publication 2, the VSMs for sentences are considered to demonstrate the proposed evaluation approach. Many natural language processing tasks, such as machine translation and question answering, use sentences as the basic units. However, there is relatively little research on vector representations of sentences. It is worth noticing that the proposed method can also be used for vector space models in general, and is not restricted to the vector representation of sentences only. The method is language-independent and only requires feature representations extracted independently for each language.

In Publication 2, a comprehensive set of parameterizations for sentence-aligned bilingual corpora is compared using the proposed evaluation method. For example, given a bag-of-words representation of sentences for the bilingual corpora, different dimensionality reduction methods, weighting and normalizing schemes have been compared. The experiments used bilingual corpora for different language pairs. In order to demonstrate the effectiveness of the proposed evaluation approach, the results were validated in three different settings:

- The results of the evaluation approach are compared against known facts based on previous studies. For instance, SVD and PCA are often perform better than other dimensionality reduction methods when the target dimensionality is relatively low (Deerwester et al., 1990; Bingham and Mannila, 2001). Among different global weighting schemes, entropy weighting and logarithmic inverse document frequency (idf)

are often better when SVD is used for dimensionality reduction (Dumais, 1991).

- The results of the proposed evaluation method are compared to the results of an indirect evaluation in two sentence matching tasks. It is shown that the vector representation which gets higher score based on the proposed evaluation method also performs better in the task of matching sentences between two languages. The two sentence matching tasks are also used as an indirect evaluation for each other. The results suggested that indirect evaluation based on two very similar tasks could not perform better than the direct evaluation method proposed in this thesis.

- The results of the evaluation approach are also validated by manually finding word translations using canonical factor loadings for different vector representations. It is shown in Publication 2 that in almost all of the cases, the higher the evaluation score, the better the translation accuracy.

CCA-based evaluation setup

In this section, the evaluation setup to compare different vector representations using CCA is explained. In order to avoid overlearning, the evaluation measure is computed using a test set. Figure 3.6 shows the flowchart of the evaluation setup.

The data is divided into three parts in the evaluation setup. The data sets $\mathbf{S}_0$ and $\mathbf{T}_0$ are the training data set for the monolingual feature extraction methods F_s and F_t, which are then applied to the rest of the data sets. The data sets $\mathbf{S}_0$ and $\mathbf{T}_0$ represents the collection of N aligned sentences in two languages, and are transformed into matrices of real vectors using F_s and F_t as follows:

$$\mathbf{X}_0 := F_s(\mathbf{S}_0) \in \mathbb{R}^{D_x \times N} \tag{3.28}$$

$$\mathbf{Y}_0 := F_t(\mathbf{T}_0) \in \mathbb{R}^{D_y \times N}, \tag{3.29}$$

where D_x and D_y are the dimensionalities of the vector representations.

The CCA is applied to the development data set $(\mathbf{S}, \mathbf{T})$ is also sentence-aligned. The trained feature extraction from $\mathbf{S}_0$ and $\mathbf{T}_0$ are used to transform $\mathbf{S}$ and $\mathbf{T}$ into matrices of real vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Finally, $(\widetilde{\mathbf{S}}, \widetilde{\mathbf{T}})$ represent the test data sets. After applying the feature extractions, the samples are projected into the common space using the

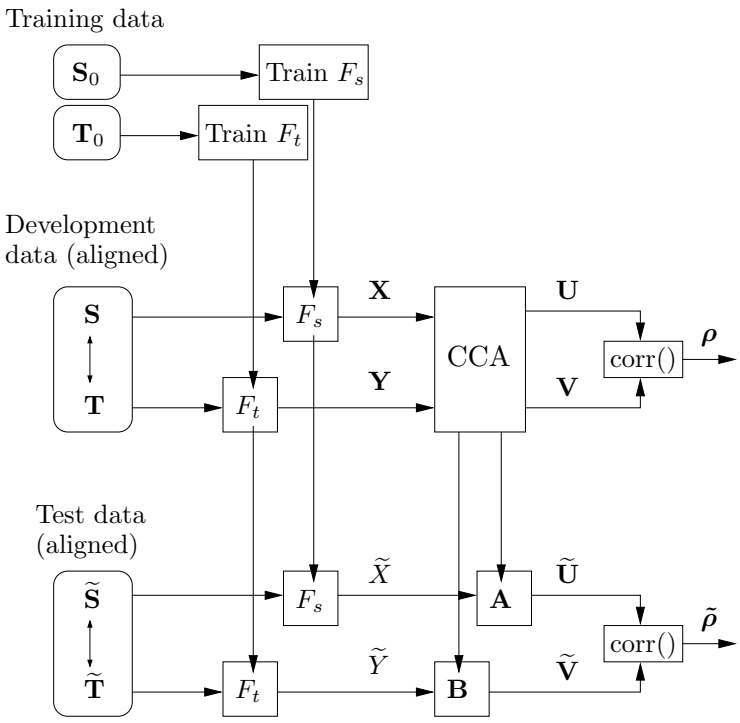

Figure 3.6: Diagram of the evaluation setup.

learned projection matrices $\mathbf{A}$ and $\mathbf{B}$. Assuming zero-mean data, the test correlations $\tilde{\boldsymbol{\rho}} = [\tilde{\rho}_1, \ldots, \tilde{\rho}_D]$ are calculated as follows:

$$\tilde{\boldsymbol{\rho}} = \operatorname{diag}\left(\frac{\mathbf{A}^T \widetilde{X} \widetilde{Y}^{\mathbf{T}} \mathbf{B}}{\sqrt{\mathbf{A}^T \widetilde{X} \widetilde{X}^{\mathbf{T}} \mathbf{A}} \sqrt{\mathbf{B}^T \widetilde{Y} \widetilde{Y}^{\mathbf{T}} \mathbf{B}}} \right). \tag{3.30}$$

We define the evaluation measure as the sum of correlations obtained from the test data,

$$R(\widetilde{X}, \widetilde{Y}) = \sum_i \tilde{\rho}_i. \tag{3.31}$$

For completely correlated sets R equals to D, and for uncorrelated sets, $R = 0$. Note that the held-out set correlations $\tilde{\rho}_i$ can also have negative values.

The evaluation setup requires a bilingual corpora of paired samples, such as sentences or documents. The proposed evaluation is based on CCA and finds linear dependency between the parallel documents, hence simple and fast to compute. However, linear dependency may not be able to capture the true dependency. The evaluation setup may be improved by

replacing linear dependency with non-linear dependency. A direct solution is to use kernel CCA instead of classical CCA in the evaluation setup but the additional kernel parameters will make the setup more complicated.

3.6 Discussion

In this chapter, novel methods for unsupervised multi-view learning are studied and developed. In an unsupervised setting, defining the agreement between views is challenging due to not having a clear hypothesis. In this thesis, statistical dependency is used to define the agreement between the views. Assuming that the shared information between the views is more interesting than the information specific to any of the views, statistical dependency can find the shared information between the views. Based on this principle, two methods are proposed in this chapter.

The problem of combining data sources has attracted the attention of researchers in many application domains. In this chapter, an unsupervised data fusion method for finding shared information between several data sources is proposed. The method is simple, fast and easily interpretable, and uses CCA in a new way for dimensionality reduction.

The second method is a simple and direct evaluation criterion based on the statistical dependency to compare several vector space models for the sentence-aligned bilingual corpora. CCA is used to find the shared information between the documents of the two languages. The novelty of this approach is to use the shared information as a measure to quantify the goodness of a vector space model.

Chapter 4

Matching problem in multiple views

In this chapter, a novel problem of matching the samples between two views is introduced. The multi-view learning methods and research problems discussed in Chapter 3 assume that the two views represent different aspects of the same set of samples, that is, the correspondence of samples between two views is known. Examples include the following: aligned corpora at a sentence or paragraph-level are usually needed for statistical machine translation; images must be paired with their captions in multimodal retrieval; and the correspondence of metabolites should be known in order to compare metabolic profiles of two species. Such strict correspondence is, however, not always known. For example, there are plenty of unaligned parallel or comparable multi-lingual documents available in various public databases or internet, and aligning such documents automatically would provide a valuable learning resource to many multi-lingual tasks such as machine translation. Another example is the matching of metabolites between two species. Metabolic profiling is a non-trivial task, and due to measurement errors, it is difficult to find the correspondence of metabolites between two species. The metabolic identities may not even be known between the two species but the functions can be similar. In order to apply traditional multi-view learning methods in such cases, it is necessary to infer the correspondence of samples between the two views. In this thesis, a novel data-driven approach to infer the matching of samples between different views by modeling the mutual dependency is proposed.

A concrete example of learning the matching of samples in two views can be explained through the research problem considered in the Publication 3. The aim of the research was to identify chromosomal regions and novel target genes involved in the tumorogenesis by combining data from two

microarray platforms, Agilent and Affymetrix. The correspondence of the samples between the two data types was not known in advance. Thus, in order to jointly analyze the two data types, it was important to infer the matching of samples between the two platforms.

Ewing sarcoma family of tumors (ESFT) is one of the most common tumors of bone in children and young adults. The progression of tumor manifests genetic alterations at chromosomal level, and identification of these chromosomal targets and markers help the diagnosis and management of patients. Rapid development of microarray technologies makes it possible to measure genomic data at different levels, for instance at DNA and RNA level; and hence leads to more sophisticated analyses. In Publication 3, array comparative genomic hybridization (CGH) data from the Agilent platform and gene expression data from the Affymetrix platform are combined to pinpoint novel candidates genes in ESFT.

High-resolution array CGH contains the probes corresponding to the genomic DNA which is utilized to identify novel genetic alterations such as copy number changes of the genes. In gene expression arrays, probes are designed corresponding to the messenger RNA (mRNA) to detect the changes in expression level of the genes. However, integrating the two platforms allows us to identify the impact of genomic changes in terms of the expression level of the genes. Since the probes are designed differently in the two platforms, it is important to identify the corresponding probes between the two platforms in order to perform the joint analyses.

In Publication 3, sequences of the probes were matched to the genome to get the chromosomal location and then each Agilent probe was matched to the nearest Affymetrix probe based on the chromosomal location. This is usually a standard practice of the matching in biological studies.

Such matching approaches, however, cannot match all of the probes due to lack of information. For instance, a probe sequence may not match to any location in the reference genome (Mecham et al., 2004b), it is not possible to match such probes based on sequence information. Also, many probes may have overlapping chromosomal locations which makes the matching more difficult. One of the solutions to overcome such issues is to use measurement values of the probes for matching. It is assumed that the probes can be matched based on the similarity of their measurement values. The genomic measurements in Publication 3 were taken for the same set of 16 patients for both platforms. Hence, measurement value of a probe in each platform is vector-valued with the same set of features. In this case when the probes have comparable measurement values, any standard measure of similarity can be used for matching. For example, given a Agilent probe, find the

nearest Affymetrix probe based on the Euclidean distance.

In practice, we may not always have the same set of features in both views. For example, if the probes in two platforms use different set of patients to measure genomic activity, the features in two data types will not be comparable. It is not straightforward to define a similarity measure between two probes that are represented by a different sets of features. Motivated by the problem, a novel data-driven matching algorithm to match the samples between two views is proposed in this thesis. The matching algorithm use the associated data values to match the samples in two views. Any prior information can also be incorporated to the matching algorithm to get more reliable matching of samples. For example, probe-sequence and chromosomal locations can be used as prior information to match probes between two microarray platforms. The multi-view matching algorithm is demonstrated in similar biological experiments, and is discussed in the Section 4.4.1.

In the Section 4.1, the matching problem is discussed in general. Section 4.2 then discusses the assignment problem and a few state-of-the-art solutions. In Section 4.3, the concept of the matching problem in multi-view learning is defined and mathematically formulated. Section 4.4 describes the solution to the matching problem using statistical dependency, both linear dependency and non-linear dependency. In Section 4.6, few related approaches for matching in multiple views are discussed. Finally, the extension of the matching problem to a more general and realistic scenario is introduced. Section 4.7 describes the *generalized matching problem* where each view is represented by multiple realizations, and the matching of samples can be computed using any pair of realizations, one realization from each view.

4.1 The matching problem

This section starts with a brief introduction to standard matching problems, in particular matching in bipartite graphs. It then describes the *assignment problem*, which is a weighted bipartite matching problem, and one of the famous solutions to the assignment problem, called the *Hungarian algorithm*. In the following section, it is shown that the problem of matching in multiple views is an instance of the assignment problem. Matching in multiple views, however, can not be directly solved using the Hungarian algorithm, which requires a cost of assignment for matching. Defining the cost of assignment for objects in two different views is a non-trivial task.

The problem of matching is a well-known concept in graph theory.

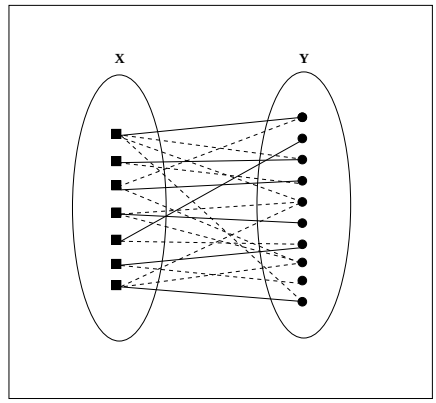

Figure 4.1: Example of a bipartite graph matching. The lines between the nodes represent the set of edges E, and the solid lines represent an instance of the maximum cardinality matching M.

Given a graph $G = (V, E)$, where V is the set of vertices and E is set of edges; a matching M is a subset of edges E, $M \subseteq E$, such that no two edges share a common vertex. Let $|M|$ denote the *cardinality* of the matching. In many applications, the matching problem can be defined on a bipartite graph. A graph is *bipartite* if its set of vertices V can be divided into two sets X and Y such that every edge in E has one end point in X and other end point in Y. A bipartite graph is represented as $G = (X \cup Y, E)$ (or simply $G = (X, Y)$). The *bipartite matching problem* is to find a matching of maximal cardinality, and is known as maximum cardinality bipartite matching. Figure 4.1 shows a bipartite graph where the lines between the nodes of X and Y represent the set of edges E and the solid lines represent an instance of the maximum cardinality matching M. Several algorithms for bipartite matching run in $O(\sqrt{n}m)$ time, and Hopcroft and Karp (1973) first proposed an algorithm to achieve this bound. Here, n denotes the number of nodes and m denotes the number of edges.

A *weighted bipartite graph* is a bipartite graph in which each edge has an associated weight. Let $c : E \mapsto \mathbb{R}$ be such a weight function. In the *weighted bipartite matching problem*, the task is to find a matching M of maximal cardinality with the maximum (or minimum) weight. The weight of matching M is defined as the sum of weights of edges in M. An example of the maximum weight bipartite matching problem is an assignment problem, which is formally defined in the Section 4.2.

4.2 Assignment problem – Hungarian algorithm

The assignment problem is a combinatorial optimization problem that consists of finding a matching in a weighted bipartite graph. As an example, suppose there are a number of agents and an equal number of tasks. Each agent can perform any task with an associated cost. The assignment problem consists of assigning each agent a task such that the sum of the costs of assignment is minimized. It is also called Linear sum assignment problem (LSAP) due to the linear cost function. There is also a *quadratic assignment problem* with a quadratic cost function, but in this work only linear assignment problem is considered.

Let $G = (X \cup Y, E)$ be a weighted bipartite graph with node sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$, edge set $E = \{[x_i, y_j], i, j \in 1, \ldots, n\}$ and the associated cost $c : E \mapsto \mathbb{R}, c(x_i, y_j) < \infty$. The assignment problem (LSAP) is then to find a perfect matching on G in terms of a mapping $f : X \mapsto Y$ such that

$$\sum_{x \in X} c(x, f(x)) \tag{4.1}$$

is minimized. The LSAP can easily be transformed into an equivalent maximization problem by using a simple transformation $c(x_i, y_j) = c_0 - c(x_i, y_j)$, where $c_0 = \max_{i,j} c(x_i, y_j)$.

The LSAP can also be formulated as a primal *linear programming problem*:

$$\text{maximize} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} z_{ij} \tag{4.2}$$

$$\text{subject to} \sum_{i=1}^{n} z_{ij} = 1, \forall j = 0, \ldots, n, \tag{4.3}$$

$$\sum_{j=1}^{n} z_{ij} = 1, \forall i = 0, \ldots, n,$$

$$z_{ij} = \{0, 1\}, \forall i, j = 0, \ldots, n.$$

The solution matrix $[z_{ij}]$ is called the primal solution, where $z_{ij} = 1$ if and only if the edge $\{x_i, y_j\}$ is in the matching M. Here, c_{ij} represents the cost $c(x_i, y_j)$.

The dual problem corresponding to the primal LSAP in Equation 4.2 can be obtained by associating dual variables u_i and v_j to the equality

constraints in Equation 4.2

$$\text{minimize} \sum_{i=1}^{n} u_i + \sum_{j=1}^{n} v_j \tag{4.4}$$
$$\text{subject to } u_i + v_j \geq c_{ij}, \forall i, j = 0, \ldots, n.$$

Many algorithms have been proposed to solve the assignment problem based on either the primal or the dual version, or using both in the primal-dual version. The first algorithm developed specifically for solving the assignment problem was the *Hungarian method* by (Kuhn, 1955) based on the primal-dual method of the linear programming. The original solution had an $O(n^4)$ run-time, but later implementations reduced it to $O(n^3)$. Kennedy (1995); Dell'amico and Toth (2000) give a nice overview of the bipartite matching and the state-of-the-art algorithms for the assignment problem, respectively. Burkard et al. (2009) give a recent comprehensive treatment to the assignment problem.

The assignment problem can be extended to the case where the node sets X and Y do not have the same number of nodes. It is called the *generalized assignment problem* (GAP) and the optimization problem can be formulated the same way as in Eq. 4.1. The GAP can also be solved using the Hungarian algorithm. The only difference to the AP is that some nodes will remain unused in the node set that has more nodes. In this work, we used the R (R Development Core Team, 2009) implementation of Hungarian method by (Hornik, 2005) which requires $O(n^3)$ run-time and can also solve the generalized assignment problem.

4.3 Matching between two different views

This section considers the problem of matching the samples between two data sets, where samples in each data set have different feature representations. The task is to match the samples using associated data values. Here, the multi-view matching problem is formally described, and formulated as an assignment problem. It is also explained how multi-view matching is different from standard matching problems and why standard solutions to the assignment problems cannot be directly applied. In Section 4.4, a novel algorithm to solve the multi-view matching problem is introduced.

Consider a simple case, $\mathbf{X} \in \mathbb{R}^{N \times D}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D}, M \geq N$, are two data matrices with the same sets of feature representations, that is, the samples of both data matrices lie in the same data space. Here, each row represents a sample and each column represents a feature. Assuming

one-to-one matching of samples, the matching problem consists of finding a permutation $\mathbf{p}$ of samples in $\mathbf{Y}$ such that the $i^{(th)}$ sample $\mathbf{x}_i \in \mathbf{X}$ is matched with the sample $\mathbf{y}_{p_i} \in Y$. The matching of samples between $\mathbf{X}$ and $\mathbf{Y}$ can be formulated as

$$\underset{\mathbf{p}}{\operatorname{argmin}} \sum_{i=1}^{N} C(\mathbf{x}_i, \mathbf{y}_{p_i}), \tag{4.5}$$

where $C(\mathbf{x}_i, \mathbf{y}_j)$ is the cost of matching $\mathbf{x}_i$ with $\mathbf{y}_j$. This is equivalent to the formulation of the assignment problem in Eq. 4.1. The idea here is to define the cost of matching or assignment using vectorial representations of samples. Since the observations $\mathbf{x}_i$ and $\mathbf{y}_j$ lie in the same data space, the distance, $d(\mathbf{x}_i, \mathbf{y}_j)$, between them can be assumed to be the measure of the likelihood of $\mathbf{y}_j$ matching the $\mathbf{x}_i$; the smaller the distance, the more likely the two samples match to each other. The problem of matching samples between the two data matrices can be written as

$$\underset{\mathbf{p}}{\operatorname{argmin}} \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_{p_i}), \tag{4.6}$$

which is the assignment problem when using distance between samples, $d(\mathbf{x}_i, \mathbf{y}_j)$, as the cost of assignment. Instead of using the distance between samples, it is possible to use other measures of similarity, for instance, the correlation between samples as the cost of assignment. In case of correlation, $\mathbf{x}_i$ is matched with the sample $\mathbf{y}_j$ that is maximally correlated, and the Eq. 4.6 becomes the maximization problem. The matching problem can be easily solved with the Hungarian method when the two views are assumed to be the same, for instance, when repeated measurements are with the same sensor. In this thesis, however, the matching problem, when $\mathbf{X}$ and $\mathbf{Y}$ consist of different sets of features representing two different views, is considered.

The task in multi-view learning, by definition, is to learn from multiple views where each view represents a different aspect. In this case, the samples in the two views $\mathbf{X}$ and $\mathbf{Y}$ will not have comparable feature representations. It is not trivial to define the cost of assignment when the samples in each view are represented by a different sets of feature representation. Even if the samples in both views have a vectorial representation of the same dimensionality, the features may not be directly comparable, and it is not possible to use similarity measures like Euclidean distance or correlation to define the cost. Hence, the matching solution described above cannot be used.

4.4 Matching between two views by maximizing dependencies

In this section, a novel approach to solving the matching problem in a multi-view setting is proposed. The approach uses the statistical dependency between the two views to infer the matching of samples. Section 4.4.1 proposes an iterative algorithm that uses CCA to model linear dependency between the views to solve the matching problem. A non-linear version of the matching algorithm is also proposed in the Section 4.4.2. Finally, Section 4.6, discusses the related matching algorithms.

Let $\mathbf{X} \in \mathbb{R}^{N \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D_y}, M \geq N$, be two data matrices representing two different views with different sets of features. We assume that each sample in $\mathbf{X}$ is matched with exactly one sample in $\mathbf{Y}$. Here, each row in $\mathbf{X}$ and $\mathbf{Y}$ represents a sample. The matching task is to infer a permutation $\mathbf{p}$ of samples in $\mathbf{Y}$ such that each $\mathbf{x}_i \in \mathbf{X}$ is matched with the sample $\mathbf{y}_{p_i} \in \mathbf{Y}$.

In Publication 4, a new approach to finding the matching of samples between two views based on the statistical dependency is introduced. Given a random permutation $\mathbf{p}$, the views will necessarily be statistically independent, that is, $p(\mathbf{X}, \mathbf{Y}(\mathbf{p})) = p(\mathbf{X})p(\mathbf{Y}(\mathbf{p}))$, where $\mathbf{Y}(\mathbf{p}) \in \mathbb{R}^{N \times D_y}$ represents a matrix obtained by picking rows indicated by $\mathbf{p}$. Hence, maximizing the dependency between the views should be a good solution to infer the permutation $\mathbf{p}$ that correctly matches the samples. It is proposed that maximizing the dependency, measured as the mutual information

$$I(\mathbf{X}, \mathbf{Y}(\mathbf{p})) = \int p(\mathbf{x}, \mathbf{y}_{p_i}) \log \frac{p(\mathbf{x}, \mathbf{y}_{p_i})}{p(\mathbf{x})p(\mathbf{y}_{p_i})} d\mathbf{x}d\mathbf{y}, \tag{4.7}$$

with respect to the permutation $\mathbf{p}$ finds a good matching. In practice, mutual information is difficult to compute and hence cannot be directly used as the cost function.

However, a lower bound to mutual information can be computed using any transformations $\mathbf{f}$ and $\mathbf{g}$, that is, $I(\mathbf{f}(\mathbf{X}), \mathbf{g}(\mathbf{Y}(\mathbf{p}))) \leq I(\mathbf{X}, \mathbf{Y}(\mathbf{p}))$. Hence, instead of the complete mutual information the lower bound can be maximized to search for dependency. The matching problem can then be formulated as

$$\max_{\mathbf{p},\mathbf{f},\mathbf{g}} I(\mathbf{f}(\mathbf{X}), \mathbf{g}(\mathbf{Y}(\mathbf{p}))) \leq \max_{\mathbf{p}} I(\mathbf{X}, \mathbf{Y}(\mathbf{p})), \tag{4.8}$$

which is maximized over $\mathbf{f}$ and $\mathbf{g}$ to make the bound as tight as possible.

A two-step iterative algorithm to solve the optimization problem in Eq. 4.8 is proposed. The first step assumes a fixed permutation $\mathbf{p}$ and

learns the projections $\mathbf{f}$ and $\mathbf{g}$, and in the second step the permutation $\mathbf{p}$ is updated given the projections learnt in the first step. Note that both steps maximize the dependency between views by optimizing the same cost function in Eq. 4.8. Section 4.4.1 describes the matching algorithm based on linear projections, and section 4.4.2 extends the matching algorithm to non-linear dependencies using the kernel approach.

4.4.1 Maximizing linear dependencies

In this section, a two-step iterative algorithm that uses linear projections to solve the matching problem in Equation 4.8 is described. The (canonical) correlation is used as a measure of dependency. It has a direct relation to mutual information, given Gaussian data, as explained in Section 3.3.5. For other distributions the relationship is only approximative. Thus, finding linear projections that maximize (canonical) correlation will also maximize the mutual information. Although correlation is not able to detect higher order dependencies, it is faster to compute.

Let $\mathbf{f}(\mathbf{x}) = \mathbf{x}\mathbf{w}_x^T$, where $\mathbf{x}, \mathbf{w}_x \in \mathbb{R}^{1\times D_x}$ and $\mathbf{g}(\mathbf{y}) = \mathbf{y}\mathbf{w}_y^T$, where $\mathbf{y}, \mathbf{w}_y \in \mathbb{R}^{1\times D_y}$ are the linear projections. Using correlation as a measure of dependency the optimization problem in Eq. 4.8 becomes

$$\max_{\mathbf{p},\mathbf{w}_x,\mathbf{w}_y} \operatorname{corr}(\mathbf{X}\mathbf{w}_x^T, \mathbf{Y}(\mathbf{p})\mathbf{w}_y^T). \tag{4.9}$$

Equation 4.9 can be solved using a two-step iterative approach: assuming fixed $\mathbf{p}$, the projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ are computed, and then the permutation $\mathbf{p}$ is updated, given fixed projection vectors.

Assuming fixed projection vectors, the optimization problem can be formulated as an *assignment problem*. Using the sample estimate of correlation for the cost in Eq. 4.9, we get

$$\max_{\mathbf{p}} \frac{\mathbf{w}_x\mathbf{X}^T\mathbf{Y}(\mathbf{p})\mathbf{w}_y^T}{\|\mathbf{X}\mathbf{w}_x^T\|\|\mathbf{Y}(\mathbf{p})\mathbf{w}_y^T\|}, \tag{4.10}$$

where the numerator can be expressed as

$$\frac{1}{2}(|\mathbf{X}\mathbf{w}_x^T\|^2 + \|\mathbf{Y}(\mathbf{p})\mathbf{w}_y^T\|^2 - \|\mathbf{X}\mathbf{w}_x^T - \mathbf{Y}(\mathbf{p})\mathbf{w}_y^T\|^2). \tag{4.11}$$

Assuming one-to-one matching and $N = M$, the first two terms in Eq. 4.11 and the denominator in Eq. 4.10 are constant with respect to $\mathbf{p}$, since the order of samples does not affect the norm. Ignoring the constant terms, the optimization problem for $\mathbf{p}$ can be written as

$$\min_{\mathbf{p}} \|\mathbf{X}\mathbf{w}_x^T - \mathbf{Y}(\mathbf{p})\mathbf{w}_y^T\|^2 = \min_{\mathbf{p}} \sum_{i=1}^{N} \|\mathbf{x}_i\mathbf{w}_x^T - \mathbf{y}_{p_i}\mathbf{w}_y^T\|^2, \tag{4.12}$$

which is an assignment problem where the cost of assignment is the Euclidean distance between the samples in the projected space. The first step of the iterative algorithm can thus be solved using the Hungarian method to infer the permutation $\mathbf{p}$. Note that Eq. 4.12 is equivalent to Eq. 4.10 only when $N = M$. When $N < M$, the term $\|\mathbf{Y}(\mathbf{p})\mathbf{w}_y^T\|^2$ is not constant, since some of the samples of $\mathbf{Y}$ will be ignored. In order to avoid this potential bias, the distance between samples is normalized as shown in Algorithm 1; for details see Publication 6.

Assuming fixed permutation $\mathbf{p}$, the task in the second step of iterative algorithm is to compute linear projections $\mathbf{w}_x$ and $\mathbf{w}_y$ to optimize the same cost function as in Eq. 4.9. This is equivalent to solving CCA for $\mathbf{X}$ and $\mathbf{Y}(\mathbf{p})$ to get the pair of projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ such that $\text{corr}(\mathbf{X}\mathbf{w}_x^T, \mathbf{Y}(\mathbf{p})\mathbf{w}_y^T)$ is maximized. CCA, however, returns projection matrices $\mathbf{W}_x \in \mathbb{R}^{D\times\mathbf{D}_x}$ and $\mathbf{W}_y \in \mathbb{R}^{D\times D_y}$, where $D = min(D_x, D_y)$, consisting of consecutive pairs of projection vectors such that $\text{corr}(\mathbf{X}\mathbf{w}_x^{(i)T}, \mathbf{X}\mathbf{w}_x^{(j)T}) = 0, \forall j \neq i$. These additional components can naturally be used while solving the matching by extending the distance measure in Eq. 4.12 to multidimensional projections.

Each pair of CCA components, $(\mathbf{w}_x^i, \mathbf{w}_y^i)$, is associated with the corresponding canonical correlation, ρ_i, which signifies the contribution of the particular pair. Since correlation is scale-invariant, each component can be re-scaled for maximum informativeness. For normal distributions, mutual information decomposes additively over components as $I(\mathbf{X}, \mathbf{Y}) = -\frac{1}{2}\sum_i(1-\rho_i^2)$. Taking the same analogy, each component is, slightly heuristically, re-scaled with corresponding canonical correlation while computing the distance, giving

$$\min_{\mathbf{p}} \sum_{i=1}^{i=N}\sum_{j=1}^{j=D} \rho_j^2 \|\mathbf{x}_i\mathbf{w}_x^{(j)T} - \mathbf{y}_{p_i}\mathbf{w}_y^{(j)T}\|^2 \tag{4.13}$$

as the final cost function to compute the matching. Here, $\mathbf{w}_x^{(j)}$ and $\mathbf{w}_y^{(j)}$ represent the jth rows of corresponding projection matrices, and ρ_j is the the associated canonical correlation.

In practice, the matching algorithm starts with a random permutation $\mathbf{p}$ and computes the linear projections based on it. Both the steps are then repeated until convergence. The algorithm depends on the initialization, and converges to a locally optimal solution. The initialization of permutation $\mathbf{p}$ is thus an important step. Section 4.4.3 describes how prior information about the matching can be incorporated in the permutation $\mathbf{p}$ to get a better initialization, and hence better convergence of the algorithm. The concept of *candidate sets* based on the prior information about

Input: Matrices $\mathbf{X} \in \mathbb{R}^{N \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D_y}$. Candidate sets S_i for each row $\mathbf{x}_i$ of $\mathbf{X}$, consisting of sets of indices for the samples in $\mathbf{Y}$. Each element in S_i is an index from 1 to M.

Output: A match between the objects in $\mathbf{X}$ and $\mathbf{Y}$, given as a vector $\mathbf{p} \in [1..M]^N$. All the elements in $\mathbf{p}$ must be unique and $p_i \in S_i \; \forall i$.

1 Initialization: Choose random $\mathbf{p}$ that satisfies the candidate sets.
2 **repeat**
3 Find the projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ and the canonical correlations $\{\rho_j\}_{j=1}^{D}$, where $D = \min(D_x, D_y)$, by maximizing the correlation between $\mathbf{X}$ and $\mathbf{Y}(\mathbf{p})$.
4 Compute pair-wise distances $d(i,k)$ between samples in $\mathbf{X}$ and samples in $\mathbf{Y}$ where $k \in S_i$, $d(i,k) = \{\sum_{j=1}^{D} \rho_j^2 \|\mathbf{x}_i \mathbf{w}_x^{(j)T} - \mathbf{y}_k \mathbf{w}_y^{(j)T}\|^2\} / \{\sum_{j=1}^{D} \rho_j \|\mathbf{x}_i \mathbf{w}_x^{(j)T}\| \|\mathbf{y}_k \mathbf{w}_y^{(j)T}\|\}$.
5 Set $d(i,k) = \infty$ for all pairs (i,k) for which $k \notin S_i$.
6 Find the match in the subspace defined by $\mathbf{W}_x$ and $\mathbf{W}_y$ by optimizing $\min_{\mathbf{p}} \sum_{i=1}^{N} d(i, p_i)$, taking into account the constraint of unique values for the elements of $\mathbf{p}$.
7 **until** $\mathbf{p}$, $\mathbf{W}_x$, *and* $\mathbf{W}_y$ *do not change* ;

Algorithm 1: Summary of the matching algorithm.

matching is defined in Section 4.4.3. The two iterative steps of the matching algorithm along with implementation of candidate sets are summarized in Algorithm 1.

Matching probes between different microarray platforms

The matching algorithm based on CCA is demonstrated by matching the probes between two microarray platforms in Publication 4. Different microarray technologies use different sets of probes to measure the genomic activities, for example, measuring expression levels of the same set of genes. Integrating information from different microarray platforms can improve the task of gene expression analysis in following ways: first, identical observations in more than one platform are supposed to be more robust when validated by biology and second, jointly analyzing data sets from multiple sources may produce more reliable and significant results. However, combining data from different platforms requires the correspondence of probes.

Standard practices of matching probes between two microarray platforms are usually gene identifier-based matching or sequence-based match-

ing (Mecham et al., 2004a). Such methods, however, ignore the measurement values of the probes in matching. In this thesis, it is shown how the measurement values of probes can also be used to find the matching while gene identifier and sequence information can be used as *prior* information. Using measurement values to matching is advantageous in cases where the probe sequence may not match to any location in the reference genome (Mecham et al., 2004b), hence gene identifier or the chromosomal location for the probe cannot be found.

As a demonstration, the matching algorithm is applied to match the probes between two different versions of the Affymetrix oligonucleotide arrays, HG-U95 and HG-U133, in Publication 4. The gene expression profiles of pediatric acute lymphoblastic leukemia (ALL) patients from (Yeoh et al., 2002; Ross et al., 2003) are used as measurement data on both HG-U95 and HG-U133 platforms for the same set of 133 patients. This setup provides an excellent test-bed for the algorithm for two reasons: first, the ground truth is known, and second, it is possible to compare the proposed matching algorithm against the alternative approach of directly using the assignment problem in the original space since both the data sets have paired features, that is the same sets of 133 patients.

The CCA-based matching is able to correctly match 72.6% of the probes between the two platforms, and clearly outperforms the comparison approach that uses the assignment problem in the original space. Both the correlation and Euclidean distance are used as the cost of assignment for the comparison. It is also empirically shown that weighting CCA components with corresponding canonical correlations, as in Eq. 4.13, performs better than any lower-dimensional subspace. The details of the experimental setup and the results can be found in Publication 4.

4.4.2 Maximizing non-linear dependencies

The matching algorithm described in Section 4.4.1 is easy to understand and implement, but it makes the strong assumption of linear dependency which might affect its performance when linear dependencies are not sufficient to capture the relationship between the two views. Often the relationships between the views are non-linear, and the matching can be better inferred using the non-linear functions $\mathbf{f}$ and $\mathbf{g}$ instead of their linear counterparts. In other words, the bound in Eq. 4.8 can be made tighter by relaxing the linearity assumption. The matching algorithm is modular in nature, and it can be easily extended to incorporate non-linear dependencies. In this section, the matching algorithm based on maximizing non-linear dependencies is described.

Kernel methods provide a good approach to detect non-linear patterns in data. The main idea behind kernel methods is to map the features of data into a new feature space such that non-linear patterns can be represented in linear form, and then any state-of-the-art method to detect linear patterns can be applied in the new feature space. This is called the *kernel trick*. A detailed overview of kernel methods can be found in (Shawe-Taylor and Cristianini, 2004). Using the kernel trick, the matching algorithm is extended as follows: the features of two data matrices $\mathbf{X}$ and $\mathbf{Y}$ are mapped into a kernel space, and then linear projection vectors and the matching are learned in the kernel space.

As described in Section 3.3.3, let $\mathbf{K}_x$ and $\mathbf{K}_y$ be the Gram matrices for the data matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. The kernelized matching algorithm works directly with the Gram matrices instead of the original data sets, and it otherwise remains the same as in Algorithm 1.

Assuming fixed permutation $\mathbf{p}$, the Gram matrices $\mathbf{K}_x$ and $\mathbf{K}_{y(\mathbf{p})}$ are used to compute projection vectors in the kernel space. Here, $\mathbf{K}_{y(\mathbf{p})}$ is the Gram matrix corresponding to $\mathbf{Y}(\mathbf{p})$. As shown in (Bach and Jordan, 2002), this is equivalent to performing KCCA on the original data sets $\mathbf{X}$ and $\mathbf{Y}(\mathbf{p})$ giving the projection vectors in terms of expansion coefficients $\boldsymbol{\alpha}_x, \boldsymbol{\alpha}_y \in \mathbb{R}^{N\times 1}$. Analogously to the projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ in classical CCA, KCCA returns the projection matrices in terms of expansion coefficients: $A_x = [\boldsymbol{\alpha}_x^1, \ldots, \boldsymbol{\alpha}_x^q]$ and $A_y = [\boldsymbol{\alpha}_y^1, \ldots, \boldsymbol{\alpha}_y^q]$, where q is the minimum of the ranks of $\mathbf{K}_x$ and $\mathbf{K}_y$.

In the second step of the algorithm, given the KCCA projection matrices, the matching can again be solved using the Hungarian method with the distances computed in the kernel space as the cost of assignment. The optimization problem to compute the matching can be written as

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \|\mathbf{k}_x^i A_X - \mathbf{k}_y^i A_y\|^2, \tag{4.14}$$

where the $\mathbf{k}_x^i$ and $\mathbf{k}_y^i$ represent the $i^{(th)}$ row of the kernel Gram matrices. The Eq. 4.14 is similar to the optimization problem in the case of matching with classical CCA. Further, each KCCA projection dimension can be rescaled with the corresponding kernel canonical correlation giving

$$\min_{\mathbf{p}} \sum_{i=1}^{N} \sum_{j=1}^{1=q} \rho_j^2 \|\mathbf{k}_x^i \boldsymbol{\alpha}_x^j - \mathbf{k}_y^i \boldsymbol{\alpha}_y^j\|^2 \tag{4.15}$$

as the final cost function to compute the matching in the kernel space.

The kernel matching algorithm makes it possible to detect nonlinear dependencies between the two views. However, it introduces some drawbacks in terms of learning more parameters. The first drawback is the regularization parameter for KCCA; as discussed in Section 3.3.3, plain KCCA overlearns badly due to high dimensionality of data and results in a poor generalization. A proper regularization is needed to avoid trivial learning. Another set of potential parameters is associated with the kernel functions used to compute kernel matrices. For example, in the case of using Gaussian kernel function to compute kernel matrices, selecting an appropriate Gaussian width is important. The regularization parameter and other potential parameters associated with the kernel function can be learnt using a validation set.

4.4.3 Incorporating prior information

The proposed two-step iterative algorithm to the matching problem is completely unsupervised in nature[1]. It is, however, also possible to incorporate any prior information about the matches into the method. For instance, in the problem of matching probes between two microarray platforms, we used chromosomal location of probes to rule out highly unlikely matches: if the two probes represent the same gene, their chromosomal locations should not be far away. Another example is the task of aligning the bilingual documents by matching their sentences. In this case, if the partial alignment at the document-level or paragraph-level is known, the matching of sentences can be restricted within the same paragraph or document in the corresponding languages.

Such prior information about the matching can be incorporated in the algorithm as additional constraints on the permutation matrix. The concept of *candidate sets* is defined in order to exclude certain matches from the set of possible solutions. For each sample $\mathbf{x} \in \mathbf{X}$, a subset of samples in $\mathbf{Y}$ is defined as candidates. The matching algorithm is allowed to find a match within the candidate set only, by giving an infinite cost to the samples outside the candidate set. This also makes the algorithm faster by avoiding the need of computing all possible distances. The use of candidate sets can make the algorithm even faster if a good implementation of sparse assignment problem such as Jonker and Volgenant (1987); Duff and Koster (2001) is used, instead of just giving infinitely high cost to the samples outside the candidate sets.

[1] Although for the KCCA case, a supervised setting is needed to choose kernel parameters

Yet another kind of prior information is the known matching for some of the samples between the two views. This is a realistic scenario in bilingual document alignment where the matching of some of the sentences could be known, and the task is to infer the matching for the rest. The known matching can be used to supervise the matching algorithm while inferring the matching for the remaining samples between two views. In Publication 5, the *semi-supervised matching* algorithm incorporating both the hard constraint of the candidate sets and the soft constraint of known partial matching is proposed.

The semi-supervised matching also works in an iterative two-step manner similar to its unsupervised counterpart. In the semi-supervised matching, both of the data matrices are complemented with the samples of known matching. The complemented data matrices are then used to compute the projections in the first step of the algorithm. However, in the second step, while inferring the matching of samples given the projections, the part with known matching is kept fixed. Using the samples with known matching along with unmatched samples helps improve the quality of projections computed at each iteration. Besides, if the matching is known for a sufficient number of samples, it can be used to initialize the KCCA projections. Such projections can in turn be used to initialize the matching of samples instead of using random initial matching. The part with matched samples can also be used as a validation set to compute the KCCA parameters.

4.5 Sentence matching in parallel bilingual documents

Several natural language processing systems used for cross-lingual information retrieval and statistical machine translation need parallel or comparable bilingual documents as learning resources. Documents in two languages which are exact translations of each other are called parallel bilingual documents, while the comparable documents are not strict translations of each other but convey the same information. Before the bilingual documents can be used for learning, it is important to get them aligned at some level, say, at sentence or document-level, depending on the task. As a concrete application of the matching algorithm, the task of matching sentences in bilingual parallel documents is considered in Publication 5. The matching of sentences is done using only monolingual data.

The problem of aligning documents has been widely studied. Earlier methods were primarily based only on "anchor" cues such as speaker's identifiers, paragraph markers and sentence lengths (Gale and Church, 1991;

Brown et al., 1991). Later, models based on translation lexicons (Wu, 1994), part-of-speech taggers (Papageorgiou et al., 1994), statistical translation models and more complex models based on co-occurrence (Melamed, 1999) have also been applied. In this thesis, sentence-alignment is addressed as a matching problem which is based on vectorial representations of the sentences only. Unlike typical sentence-alignment methods, the matching algorithm does not use information like anchor cues, sentence length, part-of-speech taggers or translation lexicons. However, it is possible to incorporate such sources of information, if available, in the form of prior information as explained in Section 4.4.3.

In Publication 5 the matching algorithm is applied to align the Finnish and the English text from Europarl corpus consisting of proceedings of the European Parliament meetings in 11 languages (Koehn, 2005). The sentences in two languages are represented as vectors, and the matching is done purely based on the vectorial-representations of the sentences. Parallel documents in two languages can be seen as two different views of the same underlying semantics. The vector representation in two languages are, however, not comparable as such since each language has its own sets of words, and the representations will capture the frequencies and co-occurrences of these words. Unless the bilingual lexicons are given, the two views represented by two languages will have non-comparable sets of features. The two views are, however, statistically dependent due to the same semantic content. The proposed matching algorithm is used to infer the matching of sentences between parallel documents using the monolingual data. In Publication 5, the non-linear matching algorithm is empirically shown to outperform the linear matching algorithms. The semisupervised approach by taking partial alignment into account further improves the matching accuracy in both the linear and non-linear matching algorithm. The experimental setup and detailed results can be found in Publication 5.

4.6 Related approaches to matching

Recently, a few other studies have also been carried out to learn from non-commensurable data sources, that is, from data sources representing different views. While some of methods directly solve the matching of objects between the two views similarly to the proposed matching approach, others try to find a common subspace for the two views where the objects can be compared. The latter kind of approaches can also be used to solve the bi-partite matching using the costs stemming from the distances in the common subspace. In this section, the related approaches for finding the

matching of objects are briefly described. In Publication 6, these related approaches are compared to the matching algorithm proposed in this thesis. The summary of the comparison is presented at the end of this section.

4.6.1 Matching with probabilistic CCA

Haghighi et al. (2008) proposed a matching method based on CCA and demonstrated its use for bilingual lexicon induction from monolingual corpora. The method resembles the matching method proposed in this thesis, although it was developed independently. The translations are induced using a generative model based on CCA. Haghighi et al. (2008) proposed an EM algorithm to solve the model that is analogous to the proposed two-step iterative method. In the M-step, the CCA parameters are computed for a given matching, and in the E-step, the matching of objects is updated given the CCA parameters.

Although the M-step in (Haghighi et al., 2008) is motivated through the probabilistic formulation of CCA. This is identical to the corresponding step in Algorithm 1, since the maximum likelihood estimate of the probabilistic CCA is shown to be equivalent to the solution of classical CCA (Bach and Jordan, 2005). The main difference between the two methods lies in the second step of the algorithm. Haghighi et al. (2008) used the marginal likelihood weights approximating pointwise mutual information as a cost for the maximum weighted bipartite matching, and pointed out that a simple proxy of using the distance between objects in the latent space is more efficient. For the matching approach proposed in this thesis, the distance between the objects for the cost function in the maximum weighted bipartite matching is directly derived from the objective function. Hence, a mathematical basis for using the distance as a cost function in the bipartite matching is provided.

4.6.2 Kernelized sorting

Quadrianto et al. (2009) proposed another method, *kernelized sorting*, for matching of objects using the same philosophy of maximizing the dependency between the views, but used a different criterion as a measure of dependency, Hilbert-Schmidt Independence Criteria (HSIC). Given the two sets of objects, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, the idea is to find a permutation matrix $\pi \in \Pi_m$ on m terms such that the pairs $\mathbf{Z}(\pi) = \{(\mathbf{x}_i, \mathbf{y}_{\pi_i})\}$ for $1 \leq i \leq m$ are maximally dependent. Here,

$$\Pi_m = \left\{\pi | \pi \in 0, 1^{m \times m} \text{ and } \pi 1_m = 1_m, \pi^T 1_m = 1_m\right\}, \tag{4.16}$$

and $1_m \in \mathbb{R}^m$ is a vector of all ones. The permutation matrix π in (Quadrianto et al., 2009) and $\mathbf{p}$ in this work are related. The i^{th} element in vector $\mathbf{p}$ is the position of 1 in the the i^{th} row of π. As shown in (Quadrianto et al., 2009), the non-parametric sorting problem of $\mathbf{X}$ and $\mathbf{Y}$ can be defined as

$$\pi^* = \underset{\pi \in \Pi_m}{\operatorname{argmax}} \, Dep(\mathbf{Z}(\pi)), \tag{4.17}$$

where $Dep(\mathbf{Z}(\pi))$ is the dependency between the random variables $\mathbf{x}$ and $\mathbf{y}$. That is, the cost function resembles the formulation in Eq. 4.8, the only difference being the measure of dependency. It has also been shown by Quadrianto et al. (2009) that if mutual information is used as a dependency measure instead of HSIC, their matching method is related to the algorithm of Jebara (2004).

Using HSIC as a dependency measure has the advantage that it does not require density estimation and requires only the kernel matrices; hence it should be faster and easier to compute. On the other hand, kernelized sorting using HSIC does not provide the explicit low-dimensional representations. One drawback with kernelized sorting is that it is highly sensitive to the initialization as shown in Publication 6.

4.6.3 Manifold Alignment

Wang and Mahadevan (2009) proposed an approach for manifold alignment that enables comparing objects from different manifolds. Manifold alignment builds a connection between non-commensurable datasets by aligning their underlying manifolds. The manifold alignment method of Wang and Mahadevan (2009) learns a low-dimensional mapping for objects in both of the datasets by matching local geometries between them, and preserving the neighborhood relationships within each dataset. Hence, the objects in the two manifolds can be compared by defining a distance in the learned low-dimensional space.

Given datasets $\mathbf{X} \in \mathbb{R}^{N \times \mathbf{D}_x}$ and $\mathbf{Y} \in \mathbb{R}^{M \times \mathbf{D}_y}$, the idea is to represent each object, $\mathbf{x}_i$ and $\mathbf{y}_j$, using its relationship to the k-nearest neighbors. This makes the comparison of $\mathbf{x}_i$ and $\mathbf{y}_j$ possible. The distance between $\mathbf{x}_i$ and $\mathbf{y}_j$ is based on alignment of the k-nearest neighbors in both of the data spaces. One drawback in this method is that the comparison of objects requires going through all possible $k!$ combinations, and therefore only small neighborhoods can be used. The eigenvectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, obtained from the eigenvalue decomposition of the joint manifold are used to compute low-dimensional projections of $\mathbf{X}$ and $\mathbf{Y}$.

The manifold alignment approach does not directly solve the one-to-one matching problem, but it provides a distance measure for objects in

the two manifolds, or views. The method can, however, be used to solve the bi-partite matching by using the distance measure as a cost function, for instance, in the assignment problem. The method proposed by Wang and Mahadevan (2009) has the advantage of not being iterative but it is computationally very heavy, and requires a lot of parameter tuning.

4.6.4 Comparison of the matching algorithms

An empirical comparison of kernelized sorting, manifold alignment and the proposed matching algorithm is conducted in Publication 6. The task was to match the metabolites between two collections of humans. The performances of methods could be compared since the ground truth of true matching was known. The CCA-based matching algorithm outperformed both the kernelized sorting using the two initialization strategies suggested in (Quadrianto et al., 2009) and the manifold alignment as proposed in (Wang and Mahadevan, 2009) in the task of metabolite matching. When coupled with the assignment problem, the performance of manifold alignment was comparable to the CCA-based matching algorithm, though. It was also pointed out in Publication 6 that manifold alignment was considerably slower than both the kernelized sorting and CCA-based matching algorithm.

4.7 Generalized matching problem

The matching algorithm proposed in this thesis and the ones proposed by Haghighi et al. (2008); Quadrianto et al. (2009) learn the matching from a single observation of samples in the two views. While the task of matching is a well-defined optimization problem, there are still uncertainties involved. None of the algorithms guarantee global optimality, and the matching solutions depend on the initialization. A different initialization may lead to a different matching solution. Also, the solution will probably change if another set of realizations for both views is used for matching. Since the matching algorithm is purely data-driven, it may be infeasible to obtain an accurate and reliable matching solution, given a single realization of the views. In Publication 6, a generalized matching algorithm is proposed for the more realistic scenario where each view is represented by multiple realizations. The final matching is inferred based on all available data as a consensus of all possible matching solutions.

To clarify the terminology, the term *realization* is used to represent a data matrix. Given a single realization for each of the two views, a *match* or *matching* represents the set of all *pairs* of samples $(\mathbf{x}_i, \mathbf{y}_{p_i})$ for a

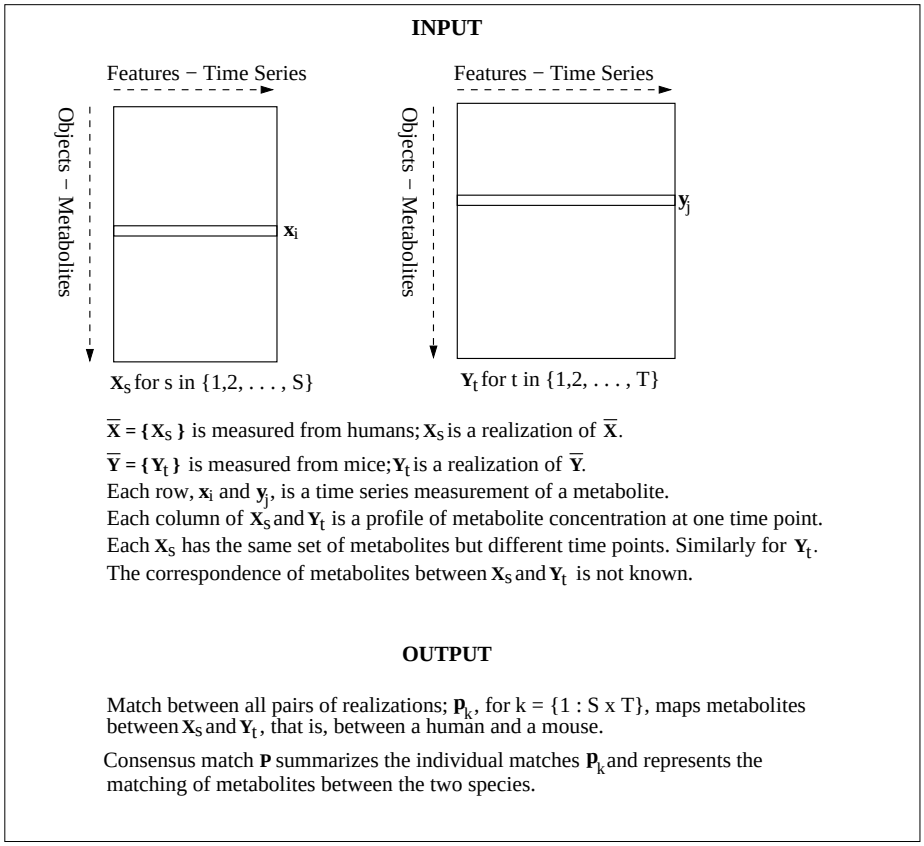

Figure 4.2: Mapping between the abstract terminology and the application of translational metabolomics. The task is to learn a *consensus match* between metabolite identities of humans and mice, *pairing* each human metabolite with one mouse metabolite. The consensus match is found by combining *individual matches* of several *realizations* of the two species. Each realization is a data matrix measuring the metabolic activity of a single individual, human or mouse. The rows of the data matrices correspond to the *objects* being matched, in this case the metabolites. The columns, in turn, are *features* that are used for learning the match, and they are the metabolic concentrations at different time points.

permutation $\mathbf{p}$. A consensus can be learned from a collection of matches (or matchings) computed from several such realizations. The terminology and the generalized matching algorithm are explained in Figure 4.2.

A realistic example is the matching of metabolites between two species in translational metabolomics. The purpose of translational metabolomics

is to study the differences and the commonalities in metabolic processes between two populations such as healthy and diabetic (Orešič et al., 2008), male and female (Nikkilä et al., 2008) or between human and a model organism such as mouse. This is an important task in order to find which properties of one population can be generalized to another population, for instance, generalizing metabolic phenotypes in mice to humans. Metabolites are small molecules that appear as an intermediate or end product of cellular processes in living organisms and the study of metabolites provides the best chance to find translational biomarkers, as has been previously demonstrated in metabolic syndrome (Damian et al., 2007). The comparative metabolome analysis is commonly performed by mass spectrometry. The identities of metabolites may not be clear due to various technical reasons of measurement process, and the functions of metabolites may be different in different tissues or species. In order to compare metabolic profiles of two species, the correspondence of functionally and structurally related metabolites should be known between the species.

In this thesis, a computationally feasible solution to combine the individual matching solutions for sufficiently many realizations of both views is proposed. The approach to compute the consensus match does not hold any assumptions about how the individual matching solution is computed. The consensus matching approach can thus be applied on top of any matching algorithm.

Consensus match

In this thesis, a novel concept of finding the matching of objects in the more realistic situation where both views are represented by multiple realizations is introduced. The task is not just to find the matching of objects given any two realizations of the two views but rather to find the underlying global match between the two views. The individual matches based on single realization of the two views are assumed to represent the instances of the underlying global match of the two views. A computationally feasible solution to compute a global match between the two views based on the consensus of the individual matching solutions is described here.

Let $\bar{\mathbf{X}} = \{\mathbf{X}_s\}, s \in \{1 : S\}$ and $\bar{\mathbf{Y}} = \{\mathbf{Y}_t\}, t \in \{1 : T\}$ be the T and S realizations of the two views, where each $\mathbf{X}_s \in \mathbb{R}^{N \times D_{xs}}$, $\mathbf{Y}_t \in \mathbb{R}^{M \times D_{xt}}$, and $M \geq N$. The number of features in each realization can be different. The individual matching solution between any two realizations $\mathbf{X}_s$ and $\mathbf{Y}_t$ can be computed using Algorithm 1. The idea is to find matches between sufficiently many realizations and then to find a consensus among them to learn a global permutation $\mathbf{P}$.

The proposed approach to learning the global match makes two simplifying assumptions: first, all pairs of realizations are assumed to be independent samples, which holds approximately assuming the total number of realizations is large, and second, while combining the matchings to compute the consensus, the information provided by the individual matching is sufficient to compute the global match. That is, the original observations are no longer needed.

In order to elaborate on the second assumption, let $\mathbf{p}_k, k \in \{1 : (S \times T)\}$ be a match between the samples of $\mathbf{X}_s$ and $\mathbf{Y}_t$ which is computed using Algorithm 1. The individual solutions $\mathbf{p}_k$ are combined through a contingency table $\mathbf{C} \in \mathbb{N}^{N \times M}$. The cell $\mathbf{C}(i, j)$ of the contingency table is the count of matchings where the ith sample of $\mathbf{X}_s$ is matched with the jth sample of $\mathbf{Y}_t$. Intuitively, if two objects are matched with each other in many individual matchings $\mathbf{p}_k$, the corresponding cell value will have a high count. According to the second assumption, only the information provided by the contingency table is used to compute the consensus match.

The N rows of the contingency table $\mathbf{C}$ represent the N samples of $\mathbf{X}_s$, and the M columns of C represent the samples of Y_t. The problem of finding the consensus match through the contingency table $\mathbf{C}$ can be formulated as finding a maximum-weight bipartite matching for the rows and columns of $\mathbf{C}$. The cost (weight) of assignment is defined based on the value of $\mathbf{C}(i, j)$, where $i \in \{1 : N\}, j \in \{1 : M\}$ and $\sum_{j=1}^{M} \mathbf{C}(i, j) = (S \times T), \forall i$. Let $\mathbf{P}$ denote the consensus matching based on the contingency table. The consensus matching can be computed by solving the following optimization problem

$$\max_{\mathbf{P}} \sum_{i=1}^{N} \mathbf{C}(i, \mathbf{P}(i)). \tag{4.18}$$

Interestingly, this is of the same form as Eq. 4.1, and hence can be solved by the Hungarian method.

In addition, an approach to characterize the potential alternative matches for each sample is also proposed. Note that the matching approach assumes the one-to-one matching of samples. However, in some applications we might be interested in one-to-many matching solutions. In order to find potential deviations from the one-to-one match and to find an alternative solution, the following approach is proposed: The contingency table $\mathbf{C}$ is re-arranged such that the pairs in the consensus match appear on the diagonal in the decreasing order of the count. Let $\mathbf{D}$ denote the re-arranged table. The approach is motivated by a crude measure of reliability of any given matched pair. It is assumed that the pairs occurring in the beginning of the diagonal are more likely to be the correct than those at the end of

the diagonal of $\mathbf{D}$.

In order to characterize the potential alternative pairs of each object, a simple approach based on randomization is proposed. The count $\mathbf{C}(r, i)$ is used as a test statistic for all $i \in [1, M]$, and the p-values for each pair of objects are estimated as the proportion where $\mathbf{Z}(r, i) > \mathbf{C}(r, i)$. The null distribution $\mathbf{Z}(r, i)$ is generated by drawing 1000 random matches that satisfy the candidate sets, and counting in how many of the random matches each of the potential pairs occur. Thus, the null distribution is constructed in the same way as the matching algorithm is being initialized. For a given object $\mathbf{x}_i$, all those $\mathbf{y}_j, j \in [1, N]$ are deemed potential alternate matches for which the pair $(\mathbf{x}_i, \mathbf{y}_j)$ has a low p-value, with any user-defined threshold.

4.8 Discussion

Multi-view learning methods have recently gained popularity in the machine learning community, and have been applied in many application domains, for example, bioinformatics, natural language processing and information retrieval. In a standard setting, multi-view learning methods assume that the correspondence of samples between the views is known. Such correspondence of samples is, however, not known in many cases. In this chapter, a generalized concept of multi-view learning methods is proposed when the correspondence of samples between the views is not known or only partially known.

The concept of a matching problem in multiple views is introduced in this chapter. A novel contribution of this thesis is the matching algorithm to find the correspondence of samples between two views. Given the solution of the matching algorithm, any standard multi-view learning method can be applied. The matching algorithm is demonstrated on three different applications. The chapter also described a few related methods to the proposed matching algorithm. The matching algorithm proposed in this thesis is empirically shown to perform better than the related methods in a metabolomics application.

Chapter 5

Summary and conclusions

In this thesis, I have considered the problem of learning from multiple views where each view represents a different aspect of the the same concept or phenomenon. The underlying assumption in multi-view learning is that learning by searching for agreement between views may improve the generalization ability. One of the important questions is how to define an agreement between views. In this work, multi-view learning methods that use statistical dependencies to find the agreement between views in an unsupervised setting have been studied and developed. Statistical dependency between views reflects what is shared or mutually informative between them, and hence provides an intuitive definition for the agreement. The methods discussed in this thesis can be easily extended to supervised or semi-supervised learning problems.

Based on the principle of using statistical dependencies between the views for multi-view learning, an unsupervised data fusion approach is proposed. The data fusion approach combines multiple data sources with co-occurring samples such that the shared information between them is preserved. It is shown how CCA can be used as a pre-processing tool for the data fusion. A randomization-based approach to determine the optimal dimensionality of the combined representation is also proposed.

Another novel contribution based on the statistical dependency between views is an evaluation criterion to compare several vector representations for sentence-aligned bilingual corpora. Choosing an appropriate vector space model is crucial to many language technology applications, for instance, machine translation and cross-language information retrieval. The indirect evaluation is typically done in the application setting which is time consuming and also restricted to given application. In this thesis, we propose a direct measure based on CCA to evaluate vector space models for bilingual corpora.

As a main contribution of this thesis, a novel concept of multi-view learning in a non-standard setting is introduced. The novel concept generalizes the multi-view learning methods to a situation where the co-occurrence of samples between the views is not known or only partially known. In the absence of such co-occurrence, it is difficult to jointly analyze multiple views. The task in the proposed approach is to infer the matching of samples using the associated data values. A novel matching algorithm to infer the one-to-one correspondence of samples between two views is proposed in this thesis. The proposed matching algorithm solves the matching problem in a general setting where the samples in two views do not have the same feature representation. The underlying assumption is that the correct matching of samples will reflect the statistical dependencies between the views. Hence, the idea is to find a matching of samples by maximizing the statistical dependencies between the views.

An iterative two-step approach to the matching algorithm based on statistical dependency between views is proposed and three variants of the generic algorithm are introduced:

- Matching of samples using canonical correlation analysis (CCA) by maximizing the linear dependency between views. The first step assumes a matching of samples, and uses CCA to find maximally dependent subspaces for the views. Given CCA subspaces, the second step computes the new matching of samples by again maximizing the dependency between views.

- A kernelized extension of the matching algorithm using kernel CCA to capture non-linear dependencies for the matching task. The idea is to use the kernel trick, that is, to project the data onto a kernel space and then apply the same two-step iterative approach to infer the matching.

- Semi-supervised matching algorithm that utilizes the given seed pairs of matched samples in the two views as a soft constraint to infer the matching of remaining samples. Semi-supervised approach is implemented in both the linear and the kernel variant of the matching algorithm.

The three variants of the matching algorithm are demonstrated on a few test cases in different applications. In Publication 4, gene probes between two versions of Affymetrix platforms are matched, and in Publication 5, sentences in Finnish-English parallel corpora are matched.

The next contribution of this thesis is a *generalized matching problem* which extends the matching problem to a more realistic setting where each

view has multiple realizations. A novel concept of the *consensus matching* is introduced in Publication 6. The consensus matching is inferred by combining all possible individual matching solutions based on any two realizations, one from each view.

Future research directions

The problem of learning from multiple views with unpaired samples is a new and emerging problem setting in many application domains, for instance, bioinformatics and natural language processing. The first obvious research direction is to apply the proposed matching algorithm also in other application areas.

The current matching algorithm infers the one-to-one matching of samples in two views. Another possible future research direction is to extend the matching algorithm for the one-to-many correspondences of samples which might be more practical in many applications. There are some other practical questions in the current matching algorithm that could be taken further for future research. The first one is the initialization strategy. The matching algorithm in this thesis, and also in (Haghighi et al., 2008; Quadrianto et al., 2009) depend on the initialization. In addition to the random initialization, prior information-based initialization in this thesis and KPCA-based initialization in (Quadrianto et al., 2009) are also proposed. A possible research direction is to develop a more generalized and robust initialization strategy. Another practical issue is the complexity of the Hungarian algorithm implementations which limits the use of matching algorithms in this thesis and in (Haghighi et al., 2008; Quadrianto et al., 2009) to really large scale data sets. In order to improve the scalability of the matching algorithms, faster and lighter implementations can be developed to solve the bipartite matching.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Archambeau, C., Delannay, N., and Verleysen, M. (2006). Robust probabilistic projections. In Cohen, W. and Moore, A., editors, *Proceedings of the 23rd International Conference on Machine Learning*, pages 33–40. ACM.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley.

Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the smo algorithm. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 6, New York, NY, USA. ACM.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105.

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining*.

Bie, T. D. and Moor, B. D. (2003). On the regularization of canonical correlation analysis. In *ICA '03: Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation*.

Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–250, New York, NY, USA. ACM.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100. ACM, New York, NY, USA.

Borga, M. (2001). Canonical correlation - a tutorial. http://people.imt.liu.se/∼magnus/cca/.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, pages 169–176, Morristown, NJ, USA. Association for Computational Linguistics.

Burkard, R., Dell'Amico, M., and Martello, S. (2009). *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia.

Cesa-Bianchi, N., Hardoon, D. R., and Leen, G. (2010). Special issue on learning from multiple sources. *Machine Learning*, 79(1–2).

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Damian, D., Orešič, M., Verheij, E., Meulman, J., Friedman, J., Adourian, A., Morel, N., Smilde, A., and van der Greef, J. (2007). Applications of a new subspace clustering algorithm (COSA) in medical systems biology. *Metabolomics*, 3:69–77.

Dasgupta, S., Littman, M., and McAllester, D. (2001). PAC generalization bounds for co-training. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 375–382, Cambridge, MA. MIT Press.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Dell'amico, M. and Toth, P. (2000). Algorithms and codes for dense assignment problems: the state of the art. *Discrete Applied Mathematics*, 100(1-2):17–48.

Dudoit, R., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139.

Duff, I. S. and Koster, J. (2001). On algorithms for permuting large entries to the diagonal of a sparse matrix. *SIAM Journal on Matrix Analysis and Applications*, 22(4):973–996.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman&Hall, New York.

Farquhar, J. D. R., Hardoon, D. R., Meng, H., Shawe-Taylor, J., and Szedmak, S. (2006). Two view learning: SVM-2K, theory and practice. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 355–362, Cambridge, MA. MIT Press.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 489–496. MIT Press, Cambridge, MA.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, pages 177–184, Morristown, NJ, USA. Association for Computational Linguistics.

Girolami, M. and Zhong, M. (2007). Data integration for classification problems employing gaussian process priors. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 465–472. MIT Press, Cambridge, MA.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings Algorithmic Learning Theory*, pages 63–77. Springer-Verlag.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129.

Haghighi, A., Liang, P., Berh-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Hardoon, D. R., Leen, G., Peltonen, J., Rogers, S., Caputo, B., Orabona, F., and Cesa-Bianchi, N., editors (2009). *Proceedings of the NIPS Workshop on Learning from Multiple Sources with Applications to Robotics*.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hopcroft, J. E. and Karp, R. M. (1973). An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, 14(12).

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hwang, K.-B., Kong, S., Greenberg, S., and Park, P. (2004). Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, 5(1):159.

Jebara, T. (2004). Kernelizing sorting, permutation and alignment for minimum volume pca. In *Conference on Learning Theory*.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Kennedy, Jr., R. J. (1995). *Solving unweighted and weighted bipartite matching problems in theory and practice*. PhD thesis, Stanford University, Stanford, CA, USA.

Kettenring, J. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Klami, A. and Kaski, S. (2006). Generative models that discover dependencies between data sets. In McLoone, S., Adali, T., Larsen, J., Hulle, M. V., Rogers, A., and Douglas, S., editors, *Machine Learning for Signal Processing XVI*, pages 123–128. IEEE.

Klami, A. and Kaski, S. (2007). Local dependent components. In Ghahramani, Z., editor, *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pages 425–432. Omnipress.

Klami, A. and Kaski, S. (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1-3):39–46.

Klami, A., Virtanen, S., and Kaski, S. (2010). Bayesian exponential family projections for coupled data sources. In Grunwald, P. and Spirtes, P., editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, pages 286–293, Corvallis, Oregon. AUAI Press.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Kuss, M. and Graepel, T. (2003). The geometry of kernel canonical correlation analysis. Technical Report 108, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Lanckriet, G. R., Bie, T. D., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lavelli, A., Sebastiani, F., and Zanoli, R. (2004). Distributional term representations: an experimental comparison. In *CIKM '04: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pages 615–624, New York, NY, USA. ACM.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):725–740.

Li, Y. and Shawe-Taylor, J. (2006). Using kcca for japanese—english cross-language information retrieval and document classification. *Journal of Intelligent Information Systems*, 27(2):117–133.

Mecham, B. H., Klus, G. T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D. Z., Mariani, T. J., Kohane, I. S., and Szallasi, Z. (2004a). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32(9):e74–.

Mecham, B. H., Wetmore, D. Z., Szallasi, Z., Sadovsky, Y., Kohane, I., and Mariani, T. J. (2004b). Increased measurement accuracy for sequence-verified microarray probes. *Physiological Genomics*, 18(3):308–315.

Melamed, D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Mitchell, T. M. (2006). The discipline of machine learning. Technical Report CMU-ML-06-108, Carnegie Mellon University, Pittsburgh, PA 15213.

Nakov, P., Popova, A., and Mateev, P. (2001). Weight functions impact on lsa performance. In *RANLP'2001: Proceedings of the EuroConference Recent Advances in Natural Language Processing*, pages 187–193.

Nikkilä, J., Roos, C., Savia, E., and Kaski, S. (2005). Explorative modeling of yeast stress response and its regulation with gCCA and associative clustering. *International Journal of Neural Systems*, 15(4):237–246.

Nikkilä, J., Sysi-Aho, M., Ermolov, A., Seppänen-Laakso, T., Simell, O., Kaski, S., and Orešič, M. (2008). Gender dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4:197.

Orešič, M., Simell, S., Sysi-Aho, M., Nanto-Salonen, K., Seppänen-Laakso, T., Parikka, V., Katajamaa, M., Hekkala, A., Mattila, I., Keskinen, P., Yetukuri, L., Reinikainen, A., Lähde, J., Suortti, T., Hakalax, J., Simell, T., Hyöty, H., Veijola, R., Ilonen, J., Lahesmaa, R., Knip, M., and Simell, O. (2008). Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *Journal of Experimental Medicine*, 205(13):2975–2984.

Papageorgiou, H., Cranias, L., and Piperidis, S. (1994). Automatic alignment in parallel corpora. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 334–336, Morristown, NJ, USA. Association for Computational Linguistics.

Pasupa, K., Saunders, C., Szedmak, S., Klami, A., Kaski, S., and Gunn, S. (2009). Learning to rank images from eye movements. In *HCI '09: IEEE International Workshop on Human-Computer Interaction, October 4, 2009, Kyoto, Japan*, pages 2009–2016.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III: Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*, A187:253–318.

Peng, Y., Zhang, D., and Zhang, J. (2010). A new canonical correlation analysis algorithm with local discrimination. *Neural Processing Letters*, 31:1–15. 10.1007/s11063-009-9123-3.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306 – 1326.

Quadrianto, N., Song, L., and Smola, A. (2009). Kernelized sorting. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1289–1296.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521.

Rényi, A. (1959). On measures of dependence. *Acta mathematica Academiae scientiarum Hungaricae*, 10:441–451.

Ross, M. E., Zhou, X., Song, G., Shurtleff, S., Girtman, K., Williams, W., Liu, H.-C., Mahfouz, R., Raimondi, S., Lenny, N., Patel, A., and Downing, J. (2003). Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959.

Rüping, S. and Scheffer, T., editors (2005). *Proceedings Of The ICML Workshop on Learning with Multiple Views*.

Sahlgren, M. (2006). Towards pertinent evaluation methodologies for word-space models. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):620.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(3):1299–1319.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86.

Viinikanoja, J., Klami, A., and Kaski, S. (2010). Variational bayesian mixture of robust cca models. In Balcázar, J., Bonchi, F., Gionis, A., and Sebag, M., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 370–385. Springer Berlin / Heidelberg.

Vinokourov, A., Christianini, N., and Shawe-Taylor, J. (2003a). Inferring a semantic representation of text via cross-language correlation analysis. In S. Becker, S. T. and Obermayer, K., editors, *Advances in Neural Processing Systems 15*, pages 1473–1480. MIT Press, Cambridge, MA.

Vinokourov, A., Hardoon, D. R., and Shawe-taylor, J. (2003b). Learning the semantics of multimedia content with application to web image retrieval and classification. In *Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation*.

Wang, C. (2007). Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18:905–910.

Wang, C. and Mahadevan, S. (2009). Manifold alignment without correspondence. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1273–1278, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, pages 723–732. Oxford University Press.

Williamson, S. and Ghahramani, Z. (2008). Probabilistic models for data combination in recommender systems. In *NIPS 2008 Workshop: Learning from Multiple Sources Workshop*.

Wu, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. In *ACL-94: 32nd Annual Meeting of the Assoc. for Computational Linguistics*, pages 80–87. Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Morristown, NJ, USA. Association for Computational Linguistics.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143.