TKK Dissertations in Information and Computer Science
Espoo 2010 TKK-ICS-D17

MUTUAL DEPENDENCY-BASED MODELING OF RELEVANCE IN CO-OCCURRENCE DATA

Eerika Savia

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium T1 at the Aalto University (Espoo, Finland) on the 18th of June, 2010, at 12 noon.

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
PO Box 15400
FI-00076 Aalto
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 470 23272
Fax +358 9 470 23277
Email: series@ics.tkk.fi

ISBN 978-952-60-3209-2 (Print)
ISBN 978-952-60-3210-8 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2010/isbn9789526032108/

AALTO ICS
Espoo 2010

ABSTRACT

Savia, E. (2010): **Mutual Dependency-Based Modeling of Relevance in Co-Occurrence Data**. Doctoral thesis, Aalto University School of Science and Technology, Dissertations in Information and Computer Science, TKK-ICS-D17, Espoo, Finland.

Keywords: Canonical Correlation Analysis, collaborative filtering, co-occurrence data, dependency modeling, eye movements, fMRI, gene regulation, latent topic models, natural stimulation, two-way grouping.

In the analysis of large data sets it is increasingly important to distinguish the relevant information from the irrelevant. This thesis outlines how to find what is relevant in so-called co-occurrence data, where there are two or more representations for each data sample.

The modeling task sets the limits to what we are interested in, and in its part defines the relevance. In this work, the problem of finding what is relevant in data is formalized via dependence, that is, the variation that is found in both (or all) co-occurring data sets was deemed to be more relevant than variation that is present in only one (or some) of the data sets. In other words, relevance is defined through dependencies between the data sets.

The method development contributions of this thesis are related to latent topic models and methods of dependency exploration. The dependency-seeking models were extended to nonparametric models, and computational algorithms were developed for the models. The methods are applicable to mutual dependency modeling and co-occurrence data in general, without restriction to the applications presented in the publications of this work. The application areas of the publications included modeling of user interest, relevance prediction of text based on eye movements, analysis of brain imaging with fMRI and modeling of gene regulation in bioinformatics. Additionally, frameworks for different application areas were suggested.

Until recently it has been a prevalent convention to assume the data to be normally distributed when modeling dependencies between different data sets. Here, a distribution-free nonparametric extension of Canonical Correlation Analysis (CCA) was suggested, together with a computationally more efficient semi-parametric variant. Furthermore, an alternative view to CCA was derived which allows a new kind of interpretation of the results and using CCA in feature selection that regards dependency as the criterion of relevance.

Traditionally, latent topic models are one-way clustering models, that is, one of the variables is clustered by the latent variable. We proposed a latent topic model that generalizes in two ways and showed that when only a small amount of data has been gathered, two-way generalization becomes necessary.

In the field of brain imaging, natural stimuli in fMRI studies imitate real-life situations and challenge the analysis methods used. A novel two-step framework was proposed for analyzing brain imaging measurements from fMRI. This framework seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available.

TIIVISTELMÄ

Savia, E. (2010): **Keskinäisiin riippuvuuksiin perustuva relevanssin mallinnus yhteisesiintymädatassa.** Väitöskirja, Aalto-yliopiston teknillinen korkeakoulu, Dissertations in Information and Computer Science, TKK-ICS-D17, Espoo, Suomi.

Avainsanat: fMRI, geenien säätely, kaksisuuntainen ryhmittely, kanoninen korrelaatioanalyysi, kollaboratiivinen suodatus, luonnolliset ärsykkeet, piilomuuttujamallit, riippuvuuden mallinnus, silmänliikkeet, yhteisesiintymädata.

Laajojen tietoaineistojen analysoinnissa on yhä tärkeämpää erottaa olennainen tieto epäolennaisesta. Tässä työssä jäsennettiin tapoja tutkia, mikä on relevanttia niin sanotussa yhteisesiintymäaineistossa, jossa jokaista näytettä vastaa kaksi tai useampia esityksiä. Kulloinkin kyseessä oleva mallinnustehtävä asettaa rajat sille, mikä on kiinnostavaa tietoa ja siten omalta osaltaan määrittelee, mikä on relevanttia. Olennaisen tiedon löytämisen ongelma on tässä muotoiltu riippuvuuden avulla; eli sellaisen variaation, joka esiintyy molemmissa (tai kaikissa) yhteisesiintyvissä datajoukoissa katsottiin olevan merkityksellisempää kuin sellaisen variaation, joka esiintyy vain yhdessä (tai joissakin) datajoukoista. Toisin sanoen, relevanssi määriteltiin datajoukkojen välisten riippuvuuksien avulla.

Menetelmänkehityksen kontribuutiot liittyvät eräisiin piilomuuttujamalleihin (topic models) sekä aineistojen keskinäisiä riippuvuuksia mallintaviin menetelmiin. Riippuvuuden etsimiseen kehitettyjä malleja laajennettiin epäparametrisiin malleihin, ja niille kehitettiin laskennallisia algoritmeja. Kehitettyjen menetelmien soveltuvuus ei rajoitu vain näissä julkaisuissa esitettyihin sovelluksiin, vaan ne ovat yleisesti käyttökelpoisia yhteisesiintymäaineistoihin sekä niiden keskinäisten riippuvuuksien mallintamiseen. Julkaisujen sovellusalueita olivat käyttäjän kiinnostuksen mallinnus, tekstin relevanssin ennustaminen silmänliikkeiden perusteella, aivojen fMRI-kuvantamisen tulosten analysointi ja geenien säätelyn mallintaminen bioinformatiikassa. Lisäksi eräisiin sovelluksiin esitettiin menetelmäkehyksiä.

Viime aikoihin asti datajoukkojen välisten riippuvuuksien mallintamisessa on ollut vallitseva käytäntö olettaa, että aineistot ovat normaalijakautuneita. Tässä työssä kanoniselle korrelaatioanalyysille (CCA) esitettiin jakaumasta riippumaton epäparametrinen laajennus sekä tätä laskennallisesti tehokkaampi semiparametrinen versio. Sen lisäksi kanoniseen korrelaatioanalyysiin johdettiin vaihtoehtoinen näkökulma, joka mahdollistaa tulosten uudenlaisen tulkinnan sekä CCA:n käyttämisen piirrevalintaan, jossa riippuvuutta pidetään relevanssin kriteerinä.

Perinteisesti nämä piilomuuttujamallit klusteroivat yhteen suuntaan, eli yksi muuttujista klusteroidaan latentin muuttujan avulla. Tässä työssä ehdotettiin piilomuuttujamallia, joka yleistää kahteen suuntaan, ja osoitettiin, että jos vain pieni määrä dataa on saatavilla, niin kahteen suuntaan yleistäminen tulee välttämättömäksi.

Aivokuvantamisen alueella luonnolliset ärsykkeet imitoivat tosielämän tilanteita ja haastavat käytössä olevat analyysimenetelmät. Uusi kaksivaiheinen kehys esitettiin fMRI-kuvantamisen mittausten analysointiin. Tämä kehys vaikuttaa lupaavalta käytettäväksi aivosignaalien analysointiin heti, kun tällaisia luonnollisten ärsykkeiden mittauksia on laajemmin saatavilla.

Contents

Preface

This thesis work has been carried out in the Adaptive Informatics Research Centre of the Department of Information and Computer Science at the Helsinki University of Technology, that is called since 2010 the Aalto University School of Science and Technology. I also have had the pleasure of being a part of the PASCAL Network of Excellence and the Helsinki Institute for Information Technology HIIT. The work has been supported by the Academy of Finland through the PROACT programme and by the European Union under the PASCAL Network of Excellence, IST-2002-506778. Additionally, I have received funding from the Graduate School in Computational Methods of Information Technology (ComMIT).

I wish to thank my supervisor Professor Sami (Samuel) Kaski for providing the quiet environment and the possibility to full-time research that enabled concentration to thinking, and for giving me the necessary amount of much-needed guidance. Without his support especially in the last years of my chronic illness this work would not have been possible. I would also like to thank Doctor Kai Puolamäki for the many worthwhile mathematical dicussions, and for his ability to maintain certain sense of humor and fun even in the most stressful of times before the paper deadlines.

Likewise, I am most thankful to my other co-authors. Writing the papers with you has been a pleasure. I wish to thank the co-authors from the fields of neuroscience and genetics: Riitta Hari, Sanna Malinen and Christophe Roos, who have greatly elevated the value of the articles by their expertise. I also want to thank the co-authors with interdisciplinary expert knowledge; Jarkko Ylipaavalniemi, Ricardo Vigário, Jarkko Salojärvi, Janne Nikkilä, Jaana Simola, Ville Tuulos and Petri Myllymäki, who have made it possible for me to experiment in different application areas. With co-authors Arto Klami and Janne Sinkkonen I have had many enjoyable and enlightening discussions concerning machine learning and the mathematics involved. I would also like to thank all the other members of our MI research group, former and present (specifically Jaakko Peltonen, Merja Oja and Jarkko Venna) for valuable discussions on both science and on everyday matters like how to use Personec Travel.

I would like to express my gratitude to the pre-examiners of this thesis, Professor Tapio Salakoski and Doctor David R. Hardoon, for their valuable feedback. I very much appreciate that they managed to review the work so promptly.

For my mental health I am greatly indebted to the horses in my life, especially Rarissime and Primadonna, and all the horsey friends, particularly my after-riding therapists Pia and Susanna.

I am grateful to my parents for the support and stubbornness in getting me to complete the thesis. But I still disagree on whether it was a wise career move. Most of all, I wish to thank my spouse Timo, and our dear little jack russell Ada, for absolutely everything.

Vihti, May 16th, 2010

Eerika Savia

List of Publications

The thesis consists of an introduction and the following publications:

1. Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola and Samuel Kaski. Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval. In Gary Marchionini, Alistair Moffat, John Tait, Ricardo Baeza-Yates and Novio Ziviani, editors, *Proceedings of SIGIR 2005, Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153. ACM, New York, NY, 2005. DOI 10.1145/1076034.1076062.

2. Eerika Savia, Samuel Kaski, Ville Tuulos and Petri Myllymäki. On Text-Based Estimation of Document Relevance. In *Proceedings of IJCNN'04, International Joint Conference on Neural Networks*, pages 3275–3280. IEEE, Piscataway, NJ, 2004. DOI 10.1109/IJCNN.2004.1381204.

3. Eerika Savia, Kai Puolamäki and Janne Sinkkonen and Samuel Kaski. Two-Way Latent Grouping Model for User Preference Prediction. In Fahiem Bachus and Tommi Jaakkola, editors, *Proceedings of UAI 2005, Uncertainty in Artificial Intelligence*, pages 518–525. AUAI Press, Corvallis, OH, 2005.

4. Eerika Savia, Kai Puolamäki and Samuel Kaski. Latent Grouping Models for User Preference Prediction. In *Machine Learning*, 74(1), pages 75–109, 2009. DOI 10.1007/s10994-008-5081-7.

5. Eerika Savia, Kai Puolamäki and Samuel Kaski. Two-Way Grouping by One-Way Topic Models. In Niall M. Adams, Céline Robardet, Arno Siebes, Jean-François Boulicaut, editors, *Proceedings of the 8th International Symposium on Intelligent Data Analysis, IDA 2009*, pages 178–189. Springer, Berlin/Heidelberg, 2009. DOI 10.1007/978-3-642-03915-7_16.

6. Janne Nikkilä, Christophe Roos, Eerika Savia and Samuel Kaski. Explorative Modeling of Yeast Stress Response and its Regulation with gCCA and Associative Clustering. In *International Journal of Neural Systems*, 15(4), pages 237–246, 2005. DOI 10.1142/S0129065705000220.

7. J. Ylipaavalniemi and E. Savia and S. Malinen and R. Hari and R. Vigário and S. Kaski. Dependencies Between Stimuli and Spatially Independent fMRI Sources: Towards Brain Correlates of Natural Stimuli. In *NeuroImage*, 48(1), pages 176–185, 2009. DOI 10.1016/j.neuroimage.2009.03.056.

8. Jarkko Ylipaavalniemi and Eerika Savia and Ricardo Vigário and Samuel Kaski. Functional Elements and Networks in fMRI. In Wei Zhang and Ilya Shmulevich, editors, *Proceedings of ESANN 2007, the 15th European Symposium on Artificial Neural Networks*, pages 561–566. D-side Publications, Bruxelles, Belgium, 2007.

9. Eerika Savia, Arto Klami and Samuel Kaski. Fast Dependent Components for fMRI Analysis. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1737–1740. IEEE, Piscataway, NJ, 2009. DOI 10.1109/ICASSP.2009.4959939.

Author's Contribution

Publication 1 is a feasibility study to test if the relevance predictions from collaborative filtering can be improved by implicit feedback measured from eye movements. The author was mainly responsible of the choices made in the collaborative filtering and made the experiments of that part of the work. The eye movement measurement setup was planned and the article was written together with the co-authors.

It was shown in Publication 2 that content-based filtering of textual movie descriptions can be improved by learning their connection to genre-information and using the learned model for new texts missing the movie genre. The author was responsible of the implementation of the experiments, which were planned as joint work with the co-authors. The preprocessing of the text data and the experiments with baseline methods were taken care of by the co-authors. The article was written in collaboration.

Publication 3 introduced a new two-way grouping latent topic model for collaborative filtering. The model was shown to be useful when only little information has been gathered of the documents. The author carried out the experiments and had the main responsibility in the writing of the article. The original idea of the model was developed jointly with the co-authors.

In Publication 4, the Two-Way Model and its closest variants were thoroughly analyzed. It was shown that generative modeling of both users and documents is advisable in normal circumstances and that two-way generalization is needed when both users and documents can be new. The author had the main responsibility of carrying out the work and writing of the article.

Publication 5 introduced a way to approximate the Two-Way Model with two one-way grouping models. The proposed method reduces the computational load significantly and produces good experimental results. The author had the main responsibility of both carrying out the work and of the original idea. The article was also mostly written by the author.

In Publication 6 the author had only a side role of analyzing how the generalized CCA could be used in the study, where multiple data sources were fused to get a more reliable estimate for which transcription factors regulate each gene of the yeast *Saccharomyces cerevisiae*. The actual method and the biological interpretations were the responsibility of the co-authors.

Publication 8 introduces a new two-step framework for studying brain activity in functional magnetic resonance imaging (fMRI) experiments. Publication 7 studies the same framework with better-justified measurements. In Publications 7 and 8 the author was responsible of the dependency analysis of the second step in the framework. The use of ICA and preprocessing of the fMRI data were responsibilities of the co-authors, as well as the neuroscientific analysis of the results.

Finally, in Publication 9, a faster semi-parametric version of the Nonparametric DeCA method is introduced and used in the fMRI framework of the earlier articles. The author had the main responsibility of carrying out the work. The original idea of the model was largely developed by the co-authors and the article was written in collaboration.

List of Abbreviations

AC	Associative Clustering
BF	Bayes Factor
BOLD	Blood Oxygenation Level Dependent (signal changes in fMRI)
BSS	Blind Source Separation (problem)
CCA	Canonical Correlation Analysis
ChIP	Chromatin Immunoprecipitation (microarray)
DDMM	Discriminant Dirichlet Mixture Model
DeCA	Dependent Component Analysis, general term for CCA and its extensions
dHMM	Discriminative Hidden Markov Model
DNA	Deoxyribonucleic acid (in genetics)
Doc URP-GEN	Document-based generative URP model
EM	Expectation Maximization (algorithm)
ESR	Common Environmental Stress Response (genes)
FMM	Flexible Mixture Model
fMRI	Functional Magnetic Resonance Imaging
gCCA	Generalized CCA
GLM	Generalized linear model
Gibbs URP	Gibbs-sampled variant of URP model
Gibbs URP-GEN	Gibbs-sampled generative variant of URP model
ICA	Independent Component Analysis
IR	Information Retrieval
KL	Kullback-Leibler (divergence)
LDA	Linear Discriminant Analysis
MAP	Maximum A Posteriori (estimate)
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MoG	Mixture of Gaussians
mPCA	Multinomial Principal Component Analysis
mRNA	Messenger-RNA
NP-DeCA	Nonparametric Dependent Component Analysis
PCA	Principal Component Analysis
pLSA	Probabilistic Latent Semantic Analysis
PoE	Product of Experts (model)
RNA	Ribonucleic acid (in genetics)
SP-DeCA	Semi-Parametric Dependent Component Analysis
TF	Transcription factor (in genetics)
TF-IDF	Term frequency – inverse document frequency weighting (in document modeling)
URP	User Rating Profile Model
User URP-GEN	User-based generative URP model

List of Symbols

N	number of data samples
d_m	dimensionality of a data set
K	number of components in dimensionality reduction
$\mathbf{x}$, $\mathbf{y}$	(vectorial) data samples ($\in \mathbb{R}^d$)
$\mathbf{x}_i$	the i'th (vectorial) data sample
$\mathbf{X}$, $\mathbf{Y}$	data matrices ($\in \mathbb{R}^{d\times N}$)
X, Y	random variables
$\mathcal{D}$	observed data
$\|D\|,\ \|v\|$	size of a collection D or absolute value of a scalar v
$\boldsymbol{\mu}_X$	mean of observations in data matrix $\mathbf{X}$
$\boldsymbol{\varphi}$	set of all model parameters
Matrices:	
$\mathbf{x}^T,\ \mathbf{X}^T$	transpose of a vector or a matrix
$\mathbf{I}$	identity matrix
$\mathbf{y}\perp\mathbf{w}$	orthogonality of two vectors in L_2 inner product
$\\| \cdot \\|$	L_2-norm of a vector or a matrix
$\\|\mathbf{C}\\|_{\mathcal{F}}$	Frobenius-norm of matrix $\mathbf{C}$
$\det \mathbf{\Sigma}$	determinant of matrix $\mathbf{\Sigma}$
λ	eigenvalue of a matrix
$\mathbf{\Lambda}$	diagonal matrix of eigenvalues in decreasing order
Probability:	
$p(\mathbf{x})$, $q(\mathbf{x})$	probability densities of (vectorial) random variable $\mathbf{x}$
$P(A \mid B)$	conditional probability of A given B
$\mathbb{E}[\,\cdot\,]$	expected value of a random variable
σ_x^2	variance of a scalar variable x
$\mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi})$	likelihood of data $\mathcal{D}$ given model parameters $\boldsymbol{\varphi}$
$\mathcal{L}(C \mid \mathcal{D}, \boldsymbol{\varphi})$	conditional likelihood of variable C, given data $\mathcal{D}$ and parameters $\boldsymbol{\varphi}$
$KL(p\\|q)$	Kullback-Leibler divergence
Dependency:	
$\rho(X, Y)$	correlation between X and Y
$H(X)$	entropy of random variable X
$I(X;Y),\ I(X_1;\ldots;X_m)$	mutual information or multi-information between random variables
$c(u, v)$	copula between cumulative distributions u and v
τ	Kendall's tau coefficient

Distributions:

$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{Bern}(\theta)$	Bernoulli distribution with parameter θ
$\mathrm{Mult}(\boldsymbol{\theta})$	Multinomial distribution with parameters $\boldsymbol{\theta}$
$\mathrm{Dir}(\boldsymbol{\alpha})$	Dirichlet distribution with parameters $\boldsymbol{\alpha}$
$\mathrm{IW}(\mathbf{S}, \nu)$	Inverse Wishart distribution with parameters $\mathbf{S}$ and ν
$\mathrm{IG}(\alpha, \beta)$	Inverse Gamma distribution with parameters α and β

Math:

$\arg\max$, $\arg\min$	argument that maximizes/minimizes the subsequent expression
δ_{ij}	Kronecker delta
$\Gamma(\,\cdot\,)$	the gamma function
$\mathcal{O}(\,\cdot\,)$	the big O notation for algorithmic complexity

Topic models:

u, d	user or document index
r	binary relevance value or rating
u^*, d^*	user group or document cluster index
N_U, N_D	number of users or documents
K_U, K_D	number of user groups or document clusters
$\boldsymbol{\theta}_U$, $\boldsymbol{\theta}_D$	multinomial parameters either for u^* (or u), and for d^* (or d)
θ_R	Bernoulli parameter for ratings r
$\boldsymbol{\alpha}_U$, $\boldsymbol{\alpha}_D$, $\boldsymbol{\alpha}_R$	Dirichlet prior parameters for $\mathrm{Mult}(\boldsymbol{\theta}_U)$, $\mathrm{Mult}(\boldsymbol{\theta}_D)$ and $\mathrm{Bern}(\theta_R)$
$\boldsymbol{\beta}_U$, $\boldsymbol{\beta}_D$	multinomial parameters for u given u^*, or d given d^*

CCA:

$\bar{\mathbf{X}}$, $\bar{\mathbf{Y}}$	deflated data sets in PCA or CCA
$\mathbf{C}$	full covariance matrix in PCA and covariance matrix of concatenated data in CCA
$\mathbf{C}_m$	covariance matrix of the m'th data set in CCA
$\mathbf{C}_{mq}$	cross-covariance matrix between m'th and q'th data sets in CCA
$\mathbf{D}$	block diagonal containing the covariance matrices of individual data sets in CCA
$\boldsymbol{\xi}$	eigenvector in CCA formulation
$\mathbf{D}^{-1}$	inverse matrix of matrix $\mathbf{D}$
V	the generalized variance
$\mathbf{X}'$	linearly transformed data set in CCA derivation

Probabilistic CCA:

$\mathbf{\Psi}_x,\ \mathbf{\Psi}_y$	covariance matrices (in probabilistic CCA)
$\mathbf{U}_x,\ \mathbf{U}_y$	matrices containing the probabilistic CCA projection vectors as columns
$\mathbf{Q}_x,,\ \mathbf{Q}_y$	matrices defining the rotation (in probabilistic CCA)
$\mathbf{P}$	diagonal matrix of canonical correlations (in probabilistic CCA extensions)
$\mathbf{w}_x^k,\ \mathbf{w}_y^k$	linear projections defining the components (in probabilistic CCA extensions)
β_k	component-wise common variance for both $\mathbf{w}_x$ and $\mathbf{w}_y$

Document modeling:

tf_{td}	term frequency of term t in document d (in document modeling)
idf_t	inverse document frequency of term t (in document modeling)

Markov models:

s_t	state of a Markov model at step t
P_{ij}	transition probability from i to j in a Markov model
y_t	output y_t at step t in a Markov model
q_{ij}	distribution of different outputs for transition $i \rightarrow j$
$\mathbf{Q}_k$	output probability matrix of state s_k
$\boldsymbol{\pi}$	state distribution (or state frequency) of a Markov model

fMRI:

V	number of voxels
p	number of reliable ICA components
q	number of stimulus features
$\mathbf{A}$	mixing matrix in ICA
$\mathbf{S}$	matrix of independent sources in ICA

Clustering:

π_k	mixture coefficients in mixture models
C_k	a cluster with index k in K-means
$\boldsymbol{\mu}_k$	mean of a mixture component or centroid of a cluster in K-means
J	K-means cost

1 Introduction

A certain perspective to data is taken in this thesis, by trying to answer the question "Which aspects of the data are relevant to the task at hand?" The ways that relevance in data is seen in this thesis are described in Section 1.1, naturally noticing that there are many other ways to consider relevance in data than the ones considered here.

More specifically, we restrict ourselves to so-called *co-occurrence data*, where there are two or more representations for each data sample. The concept of co-occurrence data is explained in Section 1.2 together with descriptions of the types of co-occurrence that were present in the publications of this thesis.

1.1 Different Views on Relevance in Data

We consider two types of relevance. The first is the relevance as seen by the user subjectively. The subjective relevance can be given, e.g., by ratings in a 1–5 stars scale or even gathered implicitly from user's actions. The second type considered is a more abstract perspective to relevance, where we define the relevant variation in several data sets to be the shared part of the variation. In other words, relevance is defined through dependencies between the data sets.

1.1.1 Relevance Seen Subjectively by User

The type of relevance commonly discussed in information retrieval literature, for instance, is the user's subjective relevance; what does the user consider relevant in the text he or she is browsing. Within our eye movement study, Publication 1, we targeted the subjective relevance. Generally, every user has his own interests or motivations in an information retrieval task, which we assumed would somehow show in the gaze pattern, while the user is looking at the interesting parts of the text. The users defined the ground truth themselves by giving relevance judgements to the document titles. The modeling task was to learn a model based on such features of the eye movements that would reveal hints of the user's relevance judgement.

On the other hand, in Publications 3, 4, and 5 the user's interests were modeled by groupings of similar users or items. The concept of relevance was, still, the user's subjective relevance for the items.

Furthermore, in Publication 2, the user's subjective ratings for different movies were modeled by such textual features that were discriminative with respect to the movie genre.

1.1.2 Relevance via Dependency

All observations incorporate some noise. The noise can arise due to errors or disturbances in the measurement process, or it can be due to some of the millions of phenomena that are not under investigation at the moment and thus, have been left unmodeled. In modeling, all variation in the data that has been left out of the model are typically considered as noise, regardless of their origin. This leads us to the view of considering *relevance* being that part of variation that we are interested in, within the current modeling task. In this view, all other variation is considered to be noise, even if it might be the focus of the modeling task in some other case.

In Publications 6, 7, 8, and 9 we defined the relevant variation to be the shared part of the variation between the two or more coupled data sets. In this view of relevance, one is looking for dependencies between several data sets, and hence, the within-dataset-variation implicitly becomes defined as noise.[1]

1.1.3 Relevance via Dependency in Brain Imaging

In Publications 7, 8, and 9 we defined the relevant parts of the brain activity and the relevant parts of the stimuli to be those that show dependency between the two data sets. The rationale behind this approach is described as follows.

Even in the most strictly controlled experimental setups there will always be a lot of meaningful activity present in the brain, other than the functional behavior that is targeted by the stimulation. By dependency analysis we can distinguish the brain activity exclusive to the hypotheses of the current experiment from all other activity.

Also, the stimulus sequences contain a lot of physical information, that is, there are many measurable features available about the stimulus sequences, out of which dependency analysis can reveal those that are the most relevant to the current study. As an example, in Publication 7, we examined the relevance of various spectral properties of auditory stimuli to the study.

1.2 Co-Occurrence Data

In co-occurrence data, there are two or more representations for each sample i (Hofmann et al., 1999; Meeds et al., 2007). In all the publications in this thesis some co-occurrence data was studied, and they will be used as examples of co-occurrence in this section.

The observations of co-occurrence data can, for instance, be vectors $\mathbf{x}_i$ and $\mathbf{y}_i$, or a vector $\mathbf{x}_i$ and a scalar class c_i. A number of different gene expression measurements $(\mathbf{x}_i, \mathbf{y}_i, \dots)$ were coupled by the genes i in Publication 6. Each data set covered a different condition of the cells, for which the gene activity was measured. In Publications 7, 8 and 9 the time steps of the brain imaging measurement paired the two data vectors $\mathbf{x}_i$ and $\mathbf{y}_i$, the vectorial representation of the stimulus and the measured brain activity, respectively.

Our application to user modeling contained a more complex example of co-occurrence. We predicted users' relevance evaluations *rel* to documents *doc*. We assumed we had observed triplet samples $(user_i, doc_i, rel_i)$, which can also be seen as two tuples $(user_i, rel_i)$ and (doc_i, rel_i) that are paired by the sample identifier i and hence, build up to a co-occurrence data set. This kind of co-occurrence data was present in Publications 2, 3, 4 and 5. There was even co-occurrence on top of co-occurrence in Publication 1. These co-occurrences will be described in more detail in Section 4.3.1.

[1]One could also use the dependencies in the opposite way, by defining the shared variation as uninteresting and by focusing on the dataset-specific variation to study source-specific phenomena. In this thesis, however, the focus is on the shared variation of the data sets.

1.3 Contributions and Organization of the Thesis

The thesis is organized as follows: in Chapters 2 and 3 the novel method development of the thesis is presented and the related methodological background is given. The scientific questions from the application point of view are discussed in the later chapters. The application areas of the publications include modeling of user interest (Ch. 4), relevance prediction based on eye movements (Ch. 5), analysis of brain imaging with fMRI (Ch. 6) and modeling of gene regulation in bioinformatics (Ch. 7).

The presentation ordering of the publications reflects the research themes as follows.
– Publication 1 is about proactive information retrieval using implicit feedback from eye movements to enhance collaborative filtering. The suggested combined method outperformed predictions of either single source of information.
– In Publication 2 the application area was also information retrieval but with content-based filtering methods.
– Publication 3 introduced a two-way grouping latent topic model motivated by collaborative filtering. The model will be discussed in Section 2.3.1.
– In Publication 4 the benefit of two-way grouping was shown when only small amount of data is available. This issue will be discussed in Section 2.3.
– Publication 5 proposed an efficient approximation to the two-way grouping model discussed in Section 2.4.
– In Publication 6 our new alternative view to CCA was applied to modeling of gene regulation in bioinformatics. This alternative view to CCA, derived in Section 3.4.4, allowed a new kind of interpretation of the results and using CCA in feature selection that regards dependency as the criterion of relevance.

The last three publications introduced our novel two-step framework for analyzing brain imaging measurements. In natural stimulation relevant combinations of stimulus features could be behind the more complex brain activation patterns. In our two-step framework dimensionality reduction by ICA produces meaningful brain activity patterns and it is followed by a dependency-seeking method between the brain patterns and the stimuli. The framework will be discussed in Section 6.6.
– In Publication 7 classical CCA was used for the dependency seeking step.
– In Publication 8 a nonparametric extension of CCA was applied. Until recently it has been a prevalent convention to assume the data to be normally distributed when modeling dependencies between data sets. We suggested using a distribution-free nonparametric extension of CCA as the dependency-seeking method in Publication 8, discussed in Section 3.6.2.
– In Publication 9 a faster semi-parametric variant of the nonparametric model was developed. This method will be discussed in Section 3.6.3

2 Latent Topic Models

In document modeling where the concept of latent topic model was first introduced, the data consists of word occurrences in text, and mutually related words are clustered into latent *topics*. The topics can then be used to categorize documents or find documents related to each other.

In user interest modeling, where the task is to predict users' relevance evaluations to documents, it is possible to generalize from already seen data either by grouping the users or by grouping the documents, which are two possible one-way groupings. If both users and documents were grouped at the same time to their respective groupings, it would constitute a two-way grouping.

In Publications 3 and 4 we have introduced a two-way grouping latent topic model and analyzed it together with its closest related models. In Publication 5 we have presented a way to approximate our Two-Way Model using two one-way grouping latent topic models. In this chapter these latent topic models and their background are discussed.

2.1 Background in Probabilistic Modeling

In this section some basic concepts of probabilistic modeling commonly used in machine learning are presented, as they are the prerequisites of the rest of this chapter.

2.1.1 Some Basics

A *probability density* function expresses the probabilities of all the possible values of the target random variables as a function of the parameters and sums up to unity,

$$\sum_j p(X = j \mid \boldsymbol{\varphi}) = 1 \quad \text{and} \quad \int_x p(x \mid \boldsymbol{\varphi})\, dx = 1 \;, \tag{1}$$

in the discrete and continuous cases, respectively.

When a sample of independent observations , $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, has been observed, the probability of the observations can be expressed as the *likelihood function* $\mathcal{L}$, which is a function of the parameters $\boldsymbol{\varphi}$:

$$\mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\varphi}) \;. \tag{2}$$

After seeing some data $\mathcal{D}$, a representative summary can be obtained by taking the parameter setting that maximizes the likelihood function, to get the so-called *Maximum Likelihood* estimate, $\boldsymbol{\varphi}_{ML}$,

$$\boldsymbol{\varphi}_{ML} = \arg\max_{\boldsymbol{\varphi}} \mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi}) = \arg\max_{\boldsymbol{\varphi}} \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\varphi}) \;. \tag{3}$$

In the case of multivariate distributions, e.g., $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$, the *marginal distribution* with respect to one of the variables is defined by summing out, or *marginalizing* over all the other variables:

$$p(\mathbf{x}) = \sum_{\mathbf{y}} \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \;. \tag{4}$$

When the distributions are continuous, the summations are replaced with integrations over the domain.

A commonly used similarity measure between two distributions, $p(\mathbf{x})$ and $q(\mathbf{x})$, is the Kullback-Leibler or KL-divergence (Kullback, 1959)

$$KL(p\|q) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} , \tag{5}$$

which can be interpreted as the average inefficiency of assuming that the distribution is $q(\mathbf{x})$ when the true distribution is $p(\mathbf{x})$, measured in bits (Cover and Thomas, 1991).

2.1.2 EM-Algorithm

The so-called expectation maximization (EM) algorithm (Dempster et al., 1977) is commonly used to find maximum likelihood solutions for probabilistic models with joint distribution

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) , \tag{6}$$

where $\mathbf{X}$ denotes all the observed variables, $\mathbf{Z}$ denotes all the latent variables of the model, and $\boldsymbol{\varphi}$ denotes the set of all the model parameters. The presentation follows the description of EM-algorithm by Bishop (2006).[2]

The goal is to maximize the likelihood function given by

$$p(\mathbf{X} \mid \boldsymbol{\varphi}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) . \tag{7}$$

By observing that for any choice of probability distribution $q(\mathbf{Z})$ we can write

$$\ln p(\mathbf{X} \mid \boldsymbol{\varphi}) = \ln\, p(\mathbf{X} \mid \boldsymbol{\varphi}) \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z})}_{=1} = \sum_{\mathbf{Z}} \left[q(\mathbf{Z}) \ln p(\mathbf{X} \mid \boldsymbol{\varphi}) \right] , \tag{8}$$

we get the following decomposition for the log-likelihood

$$\begin{aligned}
\ln p(\mathbf{X} \mid \boldsymbol{\varphi}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\ln p(\mathbf{X} \mid \boldsymbol{\varphi}) + \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) - \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\varphi})} \\
&= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) - \ln q(\mathbf{Z}) \right]}_{\mathcal{L}(q, \boldsymbol{\varphi})} + \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \left[\ln q(\mathbf{Z}) - \ln p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\varphi}) \right]}_{KL(q\|p)} \\
&= \mathcal{L}(q, \boldsymbol{\varphi}) + KL(q\|p) ,
\end{aligned} \tag{9}$$

where $KL(q\|p)$ denotes the Kullback-Leibler or KL-divergence between distributions $q(\mathbf{Z})$ and $p(\mathbf{Z} \mid \mathbf{X})$. Since the KL-divergence is guaranteed to be non-negative, the functional $\mathcal{L}(q, \boldsymbol{\varphi})$ is a lower bound for the target log-likelihood (7).

[2] Here we assume that the latent variables $\mathbf{Z}$ are discrete but an analogous derivation holds for continuous variables with summations replaced with integrations over the domain of $\mathbf{Z}$.

The EM-algorithm is a two-step iterative optimization technique based on the decomposition (9).

- **E-step (Expectation step)**
  Evaluate $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\varphi}_{old})$. In the E-step the lower bound $\mathcal{L}(q, \boldsymbol{\varphi}_{old})$ is maximized with respect to $q(\mathbf{Z})$ while holding the parameters $\boldsymbol{\varphi}_{old}$ fixed.

- **M-step (Maximization step)**
  Find the maximum

  $$\boldsymbol{\varphi}_{new} = \arg\max_{\boldsymbol{\varphi}} \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\varphi}_{old}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\varphi}) \ .$$

  In the M-step the distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\varphi})$ is maximized with respect to the parameters $\boldsymbol{\varphi}$.

It is useful to note that although the EM-algorithm does not decrease the observed data likelihood function, there is still no guarantee that it converges to a maximum likelihood estimator. For example, in the case of multimodal distributions, it may converge to a local maximum (or saddle point) of the observed data likelihood function, depending on the starting point.

The EM-algorithm is particularly useful when the likelihood belongs to the family of exponential distributions; then both the E-step and the M-step attain convenient forms. In this thesis EM-algorithm was used in the classical case of a mixture of Gaussians (Sect. 6.7.1), where it is possible to derive closed-form updates for each step.

2.1.3 Bayesian Modeling

In *Bayesian modeling* (see Bernardo and Smith (2000); Efron (2005a,b)) it is assumed that it is somehow possible to define a prior probability density for the different models (or parameter values). The prior could be given by an expert of the field under study, be based on common sense or be as non-informative as possible. In any case, the idea is that the events that have not occurred in the data so far do not necessarily have zero probability, but there is a prior belief of what can happen and how probably. It is also thought that every bit of new data updates the beliefs from what they were prior to seeing the data.

The probabilities of different models (parameter values $\boldsymbol{\varphi}$) *after* seeing some data $\mathcal{D}$, is defined as the *posterior distribution*

$$p(\boldsymbol{\varphi} \mid \mathcal{D}) = \frac{\mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi})\, p(\boldsymbol{\varphi})}{p(\mathcal{D})} = \frac{\mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi})\, p(\boldsymbol{\varphi})}{\sum_j \mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi}_j)\, p(\boldsymbol{\varphi}_j)} \quad , \tag{10}$$

which is derived by using the definition of conditional probability and by the Bayes formula

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)\, P(A)}{\sum_j P(B \mid A = j)\, P(A = j)} \ , \tag{11}$$

where j goes over all possible values of random variable A. Analogously, the formula can be written for continuous distributions as

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{\int\limits_A p(B \mid A)\, p(A)\, dA} \,. \tag{12}$$

In Bayesian modeling one is typically interested in deriving an estimate for the posterior distribution over different models rather than one fixed model, as in the case of point estimates. If a point estimate is needed, however, it can be taken as the highest peak of the posterior distribution. It is called the *Maximum A Posteriori* estimate, or *MAP-estimate* $\boldsymbol{\varphi}_{MAP}$,

$$\boldsymbol{\varphi}_{MAP} = \arg\max_{\boldsymbol{\varphi}} \mathcal{L}(\mathcal{D} \mid \boldsymbol{\varphi})\; p(\boldsymbol{\varphi}) = \arg\max_{\boldsymbol{\varphi}} \prod_{i=1}^{N} \left[\, p(\mathbf{x}_i \mid \boldsymbol{\varphi})\; p(\boldsymbol{\varphi}) \,\right] \,. \tag{13}$$

Nevertheless, the posterior distribution can have many peaks or nodes and taking just the highest node as a point estimate might be misleading. Since the evaluation of the posterior distribution typically involves nontrivial integrations, it is rare that a closed-form solution would be available. Therefore, the usual way to evaluate Bayesian models is either by Markov chain Monte Carlo sampling (MCMC; for textbook reference, see MacKay (2003)) or by variational approximation (Sect. 2.1.4).

The Bayesian models were evaluated by Gibbs sampling in Publications 1, 3, 4 and 5, described in more detail in Section 4.5.3. Gibbs sampling is one of the MCMC sampling methods where the model parameters are sampled one at a time, each from a conditional distribution where all other parameters are assumed to take their current sampled values (Casella and George, 1992; Geman and Geman, 1984). Furthermore, some of the alternative methods we compared against were evaluated by variational approximation. The principle of variational approximation is illuminated in the next subsection.

2.1.4 Variational Approximation

Variational approximation is a commonly used practice to approximate Bayesian models (see, e.g., Bishop (2006) for textbook reference). Variational methods have their origins in the 18th century work on the calculus of variations (Sagan, 1969). A *functional* is a mapping that takes a function as input and returns a scalar value as output. The calculus of variations deals with infinitesimal changes in the functions and their reflections on the value of the functional. Many problems can be considered as optimization problems where the solution is obtained by exploring all possible input functions. If the range of possible functions is suitably restricted, a more easily computable approximation can be obtained. Usually, in the application to probabilistic inference the restriction is made by assuming the distribution to be factorizable in a certain way (Jordan et al., 1999). The factorized form of variational inference corresponds to the approximation framework called *mean field theory* in physics.

If our model is specified by the joint distribution $p(\mathbf{X}, \mathbf{Z})$ and we wish to compute the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$, the decomposition of the EM-algorithm, Eq. (9), can be used:

$$\begin{aligned}\ln p(\mathbf{X}) &= \int\limits_{\mathbf{Z}} q(\mathbf{Z}) \left[\ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right] d\mathbf{Z} + \int\limits_{\mathbf{Z}} q(\mathbf{Z}) \left[\ln \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X})} \right] d\mathbf{Z} \\ &= \mathcal{L}(q) + KL(q \| p) \end{aligned} \tag{14}$$

and, instead of allowing any probability distribution $q(\mathbf{Z})$, a factorizable distribution

$$q(\mathbf{Z}) = \prod_{j=1}^{M} q_j(\mathbf{Z}_j) \tag{15}$$

is substituted into the optimization. In effect, a lower bound for the log-likelihood is maximized by choosing the optimal $q(\mathbf{Z})$ within the model family. If all possible distributions $q(\mathbf{Z})$ were available, the optimum would become the true posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$. In many cases this factorizable approximation makes an intractable posterior estimation tractable.

Variationally computed models were used in Publications 4 and 5.

2.2 One-Way Grouping Latent Topic Models

Latent topic models are a class of models where one or many latent variables are assumed to generate the observed variables (Steyvers and Griffiths, 2005). In document modeling it is assumed that one or many topics determine the probabilities of certain words to occur in a document.

The topic models that are discussed here have their roots specifically in the following two topic models: probabilistic Latent Semantic Analysis (pLSA, Hofmann (2001, 2004)) and Latent Dirichlet Allocation (Pritchard et al. (2000), see Blei et al. (2003)), also known as Multinomial Principal Component Analysis (mPCA, Buntine (2002)).

In one-way grouping models there is one latent variable, for example user group u^*, that is assumed to be responsible for generating groups of similar-minded users and their ratings. Another example of a one-way grouping model would be a topic model with latent variable Z expressing a topic of a text document, that is assumed to be responsible for generating the words of text documents from different topics.

Throughout this chapter the notation of Table 1 will be used in the model descriptions. We use the term “document” to refer to any items that users could give ratings or relevance evaluations for.

2.2.1 User Rating Profile Model

User Rating Profile model (URP, Marlin (2004a)) is a one-way grouping model that predicts user preferences on documents or other items. The model generalizes over users u, which belong to user groups $u*$ probabilistically. Therefore, each user can belong to many groups with varying degrees. Once the “attitude” or user group $u*$ and the document d have been fixed, the rating r depends on the corresponding multinomial $\boldsymbol{\theta}_R$ over the different rating values. In our studies the ratings were restricted to be binary, which reduced the multinomial θ_R to be a Bernoulli distribution. All the multinomials of the URP model have conjugate prior distributions (i.e., Dirichlet $\boldsymbol{\alpha}_U$ and $\boldsymbol{\alpha}_R$). See Figure 1 for a graphical model representation of the URP model.

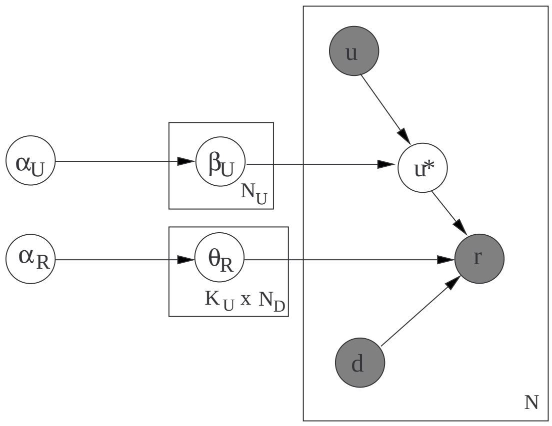


Figure 1: Graphical model representation of the original User Rating Profile model (URP). The grey circles indicate observed values. The boxes are "plates" representing replicates and the value at a corner of each plate indicates the number of replicates. The rightmost plate is repeated for each given (u, d) pair (altogether N pairs). The upper left plate represents the multinomial models of different users. The lower left plate represents the multinomial relevance probabilities of the different (user group, document) pairs. Both priors $\boldsymbol{\alpha}_U$ and $\boldsymbol{\alpha}_R$ follow Dirichlet distribution.

Table 1: Notation

SYMBOL	**DESCRIPTION**
u	user index
d	document index
r	binary relevance (relevant $=1$, irrelevant $=0$)
u^*	user group index (attitude in URP)
d^*	document cluster index
N_U	number of users
N_D	number of documents
N	number of triplets (u, d, r)
K_U	number of user groups
K_D	number of document clusters
$\mathcal{D}$	observed data

Originally, the URP model was suggested to be estimated by variational approximation (variational URP, Marlin (2004a,b)), but we have introduced also two Gibbs-sampled variants of the model in Publications 3 and 4 (Gibbs URP and Gibbs URP-GEN, sketched in Section 2.4, Fig. 3).

2.3 Two-Way Grouping Latent Topic Models

When making user preference predictions, it is possible to generalize from the observed data either by grouping the users or by grouping the documents. We have shown that grouping the users is needed when the documents are new, that is, have only few ratings available. On the other hand, grouping the documents is necessary when the users are new, that is, they have given only few ratings. In two-way grouping models both users and documents are grouped at the same time. Two-way grouping has been shown to be beneficial when there is not enough data about either users or documents in order to learn a more detailed model.

2.3.1 Two-Way Model

In Publication 3 we go one step further from the URP model which has a latent structure for the users, and introduce a similar latent structure for the documents as well[3]. The effects of the two-way grouping and generation of the marginals have been analyzed in Publication 4. The graphical model representation of the Two-Way Model is shown in Figure 2.

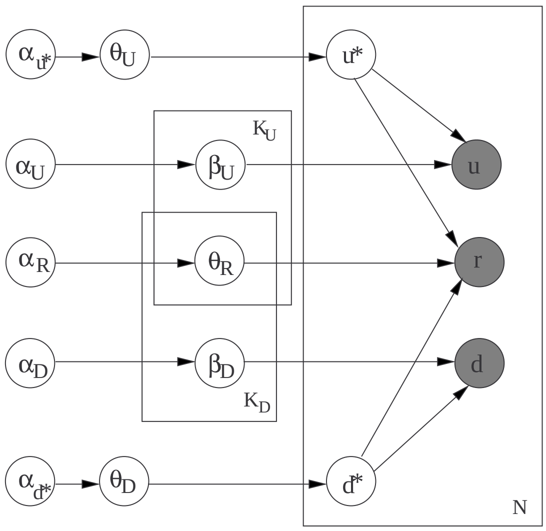


Figure 2: Graphical model representation of the Two-Way Model. The rightmost plate represents the repeated choice of N (user, document, rating) triplets. The plate labeled with K_U represents the different user groups u^*, and $\boldsymbol{\beta}_U$ denotes the vector of multinomial parameters for each user group. The plate labeled with K_D represents the different document clusters d^*, and $\boldsymbol{\beta}_D$ denotes the vector of multinomial parameters for each document cluster. In the intersection of these plates there is a Bernoulli-model with parameter θ_R for each of the $K_U \times K_D$ combinations of user group and document cluster. All the priors $\boldsymbol{\alpha}$ follow Dirichlet distributions.

[3]A similar two-way structure has been suggested by Si and Jin (2003) with some technical differences that will be discussed in Section 2.3.2.

2.3.2 Comparison to Other Two-Way Models

In this section comparisons are made between our Two-Way Model and other two-way grouping models. Additionally, the main differences and similarities with URP model are brought up.

In the Flexible Mixture Model (FMM, Si and Jin (2003)), as well as in our Two-Way Model, both *users* and *documents* can belong to many latent groups, in much the same way as users do in the URP model. In both the Two-Way model and in the FMM model the relevance is assumed to depend only on the latent groups, that is, there is a probability distribution of different ratings, $\mathrm{Mult}(\boldsymbol{\theta}_R)$, for each (user group, document cluster) pair.

In addition to being two-way, our model and FMM differ from URP in that the users u and documents d are explicitly generated. In contrast, the original URP model contains no generative process for the users or documents.

The main difference between our Two-Way model and Flexible Mixture Model is that our model is fully Bayesian and optimized by sampling the posterior distribution. FMM simply finds the maximum likelihood solution with the EM-algorithm.

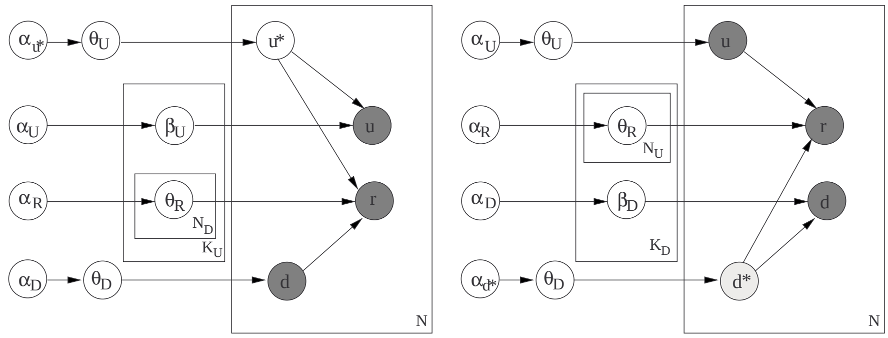
Other related models include a graphical model for gene expression model by Segal et al. (2003), where genes and measurement conditions correspond to users and documents in our model, but they are taken as given covariates of the model. The expression level corresponds to ratings in our Two-Way model, and it is assumed to be normally distributed, with mean and variance depending on the group memberships. However, in this model the genes and conditions are not clustered in parallel, but belong with varying extent to the one and only set of clusters of the model, called processes. It is a maximum-likelihood model that is optimized with EM-algorithm.

2.3.3 Generation of Marginals

Besides the difference of being one-way or two-way, there is an additional difference between URP and the Two-Way Model in whether the users and documents are assumed to be generated by the model or treated as covariates of the model. In Publication 4, it was found that unless the data marginal densities $p(u)$ and $p(d)$ are especially misleading about the structure of the full data density $p(r, u, d)$, it is always useful to design the model to be fully generative, in contrast to seeing users and documents as given covariates of the model.[4] For that reason we have introduced two fully generative variants of the URP model in Publications 4 and 5, user-grouping generative URP (User Gibbs URP-GEN, see Figure 3(a)) and document-grouping generative URP (Doc Gibbs URP-GEN, see Figure 3(b)).

Furthermore, by combining these two one-way grouping models it is possible to approximate the Two-Way Model with reduced computational complexity, as discussed in the next section.

[4]In such misleading data sets, most of the observations lie in the area that is non-informative about the relevance. Then, it makes a difference whether the model generates the users and documents from their marginal distributions, because ignoring the generation essentially equals to assuming that all (u, d) pairs carry equal amount of information about the relevance r.



(a) User Gibbs URP-GEN groups only users and assumes that the relevance depends solely on the user group and the document.

(b) Doc Gibbs URP-GEN groups only documents and assumes that the relevance depends solely on the document cluster and the user.

Figure 3: Graphical model representations of the generative Gibbs URP models with user grouping (User Gibbs URP-GEN) and with document grouping (Doc Gibbs URP-GEN). The grey circles indicate observed values. The boxes are "plates" representing replicates; the value in a corner of each plate is the number of replicates. The rightmost plate represents the repeated choice of N (user, document, rating) triplets. The plate labeled with K_U (or K_D) represents the different user groups (or document clusters), and $\boldsymbol{\beta}_U$ (or $\boldsymbol{\beta}_D$) denotes the vector of multinomial parameters for each user group (or document cluster). The plate labeled with N_D (or N_U) represents the documents (or users). In the intersection of these plates there is a Bernoulli-model for each of the $K_U \times N_D$ (or $K_D \times N_U$) combinations of user group and document (or document cluster and user). Since $\boldsymbol{\alpha}_D$ and $\boldsymbol{\theta}_D$ (or $\boldsymbol{\alpha}_U$ and $\boldsymbol{\theta}_U$) are conditionally independent of all other parameters given document d (or user u), they have no effect on the predictions of relevance $P(r \mid u, d)$ in these models. They only describe how documents d (or users u) are assumed to be generated.

2.4 Approximate Two-Way Grouping by One-Way Topic Models

The task of *biclustering* is to simultaneously cluster the rows and columns of a data matrix in such a way that the submatrices spanned by pairs of row and column clusters are as uniform as possible (Madeira and Oliveira, 2004; Tanay et al., 2006). Different definitions have been suggested, some of which allow overlapping of the biclusters (*soft* biclustering) while others require mutually exclusive biclusters (*hard* biclustering).

It has been shown for hard biclustering of binary data matrices that clustering the marginals independently to produce a checkerboard-like biclustering is guaranteed to achieve relatively good results compared to the NP-hard optimal solution. An approximation ratio for the crossing of two one-way clusterings has been proven (Anagnostopoulos et al., 2008; Puolamäki et al., 2008).

Inspired by the above-mentioned theoretical guarantee, we suggested in Publication 5 approximating the Two-Way Model with two URP models: one that groups users and one that groups documents. The combination of the two Gibbs-sampled probabilistic predictions was made using a product of experts model (Hinton (2002), see Section 4.5.4 for details).

2.5 Combining Many Probabilistic Predictions

The above-mentioned combination of one-way models is a practical and efficient way to combine two probabilistic predictions, but the combination can be done in a more principled way, as well. In Publication 1, a well-justified means for a similar combination task was needed, with the additional challenge that one of the sources was very noisy.

In situations where one needs to combine two or more models using different sources of information about the same topic of study, the simplest way to combine the models is to train the models independently and combine the predicted probabilities to produce the final prediction. This approach has the advantage of being modular and easily extensible. In Publication 1, a generative model called the *Discriminative Dirichlet-Mixture Model* for combining probabilistic predictions was introduced.[5] The goal of the model is to find an expression for $P(r|P_A, P_B, \boldsymbol{\varphi})$, where $\boldsymbol{\varphi}$ denotes all parameters of the model and the (noisy) predictions of the two separate models are denoted by P_A and P_B. The model is described in figure 4.

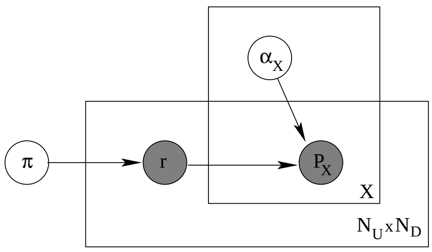


Figure 4: A graphical model representation of the Discriminative Dirichlet-Mixture Model. X is the index of the model that predicts relevance, thus in our case $X \in \{A, B\}$. The grey circles indicate observed values. For each (user, document) pair, a binary relevance r is drawn from Bernoulli(π). For each $X \in \{A, B\}$, a Bernoulli-parameter P_X is drawn from Dirichlet($\boldsymbol{\alpha}_X^r$). Observations are triplets (r, P_A, P_B) for each user-document pair. The model is optimized by maximizing the conditional log-likelihood of the relevances.

2.6 Summary

In Publications 1, 3, 4 and 5 latent topic models have been developed. To give the needed amount of background knowledge some basics of probabilistic modeling were recapitulated in the beginning of this chapter.

Topic models generally assume the data samples to be a set of discrete observations, like the set of words occurring in a document in the bag-of-words model. Traditionally, the topic models are one-way clustering models, that is, one of the variables is clustered by the latent variable. Motivated by the application of collaborative filtering we can see that it would be useful to be able to generalize both over the users and over the items of interest. In Publication 3 we proposed a generative latent topic model that groups two ways, coined the Two-Way Model. We have shown in Publication 4 that when only small amount of data has been

[5]Besides giving predictions of relevance, the Dirichlet-mixture reveals how useful the different sources of relevance information are relative to each other. Some of the feedback channels may produce smaller prediction probabilities P_X than others for the observed relevances r. Some of the relevance feedback channels may additionally be noisy, that is, the prediction probabilities P_X for a given relevance r have a large variance.

gathered, two-way generalization becomes necessary. Naturally, the two-way generalization is not restricted to the application of user interest modeling but can be applied to any co-occurrence data.

In Publication 5 we introduced a new efficient approximation of the Two-Way model that achieves the prediction performance of the original Two-Way Model but with the computational complexity of the one-way grouping model.

Finally, as a more principled means of combining many probabilistic predictions, e.g., predictions from two one-way topic models, the Discriminative Dirichlet-Mixture Model was introduced in Publication 1.

3 Modeling Dependence between Data Sets

3.1 Motivation

The aim of mutual dependency modeling in this thesis is to find maximally dependent representations for two data sets with co-occurring samples in the setting of exploratory data analysis. Our argumentation for searching for dependencies between data sets is to use dependency as a definition for what is relevant in the data. In particular, this way we target such information that could only by chance be unveiled by examining any individual source alone. This has been a main research topic for our research group; for a recent view to earlier work, see Klami (2008).

It is important to note that this approach is not equivalent to just concatenating all observations and analyzing the resulting data set with unsupervised methods. In general, unsupervised model, like PCA, cannot find the dependency structure from the concatenated data set. Since different sources can have their own characteristic ways of producing noise and irrelevant information to the data, the distinction between the sources bears meaning, and taking it into account makes a difference.

Data fusion or *data integration* is a field within data analysis where several data sources are joined into the analysis, aiming to improve the performance. When used in supervised learning data fusion is conceptually straightforward. Given the target criterion, e.g., classification accuracy, in a supervised task one can just utilize all the aspects of the data sources that improve the criterion. On the other hand, in an unsupervised setting, the selection of the criteria defining relevant aspects in data is in a key role, since the performance of the final result cannot be evaluated in such a straightforward manner. The focus of this thesis is on unsupervised learning.

A general term *multi-view learning* has been recently used for various methods that take into account many different views or data sources while modeling the same objects of interest. In practice this means searching for commonalities between many data sets about the same objects, in order to learn better-generalizing models than from individual data sources. The perspective of this thesis is related to multi-view learning, in the sense that relevance in co-occurrence data is here defined by statistical dependencies between the paired data sets. However, here we make an important additional assumption compared to much of the multi-view learning, that any model learned based on only one of the sources might be insufficient or even misleading.

3.2 Measures of Statistical Dependency

Search for commonalities between the data sources requires a measure of dependency. There are several alternative measures and the choice is reflected in the kind of dependencies that one is able to find. In this section some measures of dependency central to this thesis are introduced. Deza and Deza (2009) give a comprehensive overview of distances and similarity measures, including distances between probability distributions.

3.2.1 Correlation

The Pearson correlation coefficient ρ measures linear dependence between two random variables X and Y,

$$\rho(X,Y) = \frac{\mathbb{E}\left[X - \boldsymbol{\mu}_X\right]^T \mathbb{E}\left[Y - \boldsymbol{\mu}_Y\right]}{\mathbb{E}\left[\|X - \boldsymbol{\mu}_X\|\right] \ \mathbb{E}\left[\|Y - \boldsymbol{\mu}_Y\|\right]} \ . \tag{16}$$

Here $\mathbb{E}$ denotes expectation, $\boldsymbol{\mu}_X$ the mean of X and $\boldsymbol{\mu}_Y$ the mean of Y. This measure is used in Canonical Correlation Analysis, discussed in Section 3.4. Other correlation-related measures are found, e.g., in (Deza and Deza, 2009; Rényi, 1959).

3.2.2 Information-Theoretic Measures

The amount of information that is possible to convey by a random variable X can be measured using a quantity called *(Shannon) entropy* (Cover and Thomas, 1991). In other words, entropy is the average amount of uncertainty or randomness there is in the value of an unobserved random variable. Entropy over a discrete random variable X is defined[6] as

$$H(X) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}), \tag{17}$$

and entropy over a continuous random variable analogously as

$$H(X) = -\int_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \, d\mathbf{x} \ . \tag{18}$$

Mutual information measures the amount of information shared by two random variables X and Y (Cover and Thomas, 1991), and its definition for discrete random variables is

$$I(X;Y) = \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}, \tag{19}$$

where the summations go over all possible values of X and Y, and in the continuous case the definition is

$$I(X;Y) = \int_{\mathbf{x}} \int_{\mathbf{y}} p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \, d\mathbf{x} \, d\mathbf{y} \ . \tag{20}$$

It cannot, however, be easily estimated in practice (Bromiley et al., 2004; Hutter and Zaffalon, 2004). *Multi-information* is a generalization of mutual information to more than two data sets. In the discrete form it is

$$I(X_1; X_2; \ldots; X_m) = \sum_{\mathbf{x}_1} \cdots \sum_{\mathbf{x}_m} p(\mathbf{x}_1, \ldots, \mathbf{x}_m) \log \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_m)}{p(\mathbf{x}_1) \cdots p(\mathbf{x}_m)} \quad . \tag{21}$$

We used it in Publication 6 as a measure of dependency between 16 data sets (See Section 7.4 for more details).

A useful connection is that a commonly used divergence measure between two distributions, the Kullback-Leibler divergence Eq. (5), allows mutual information to be stated as a divergence between the joint distribution $p(\mathbf{x},\mathbf{y})$ and the product of the marginal distributions $p(\mathbf{x})$ and $p(\mathbf{y})$.

[6] The choice of base of the logarithm determines the unit of entropy, e.g., $\log_2$ corresponds to bits.

3.2.3 Bayes Factor

The so-called *Bayes Factor* from statistical analysis can also be used as a measure of dependence (Kass and Raftery, 1995). Assume that we have to choose between two models H and $\bar{H}$ based on the observed data $\mathcal{D}$; this is generally referred to as the model selection problem. In Bayesian modeling the superiority of one model over the other can be evaluated by the Bayes factor (BF) given by

$$\frac{P(\mathcal{D} \mid \bar{H})}{P(\mathcal{D} \mid H)} = \frac{P(\bar{H} \mid \mathcal{D})}{P(H \mid \mathcal{D})} \cdot \frac{P(H)}{P(\bar{H})} \quad . \tag{22}$$

If it is used as a measure of dependence, the null hypothesis H assumes independent marginal distributions and $\bar{H}$ is the interesting hypothesis that shows dependency between the marginals. It is used as a measure of dependence, e.g., in *Associative Clustering* (see Sect. 3.6.1, Kaski et al. (2005b)).

3.2.4 Other Measures of Dependence

The multivariate normal distribution and linear correlation are the basis of most models used to model dependency, since linear correlation is a natural measure in the context of normal distributions. However, empirical linear correlation underestimates the amplitude of the actual correlation in the case of non-Gaussian marginals (Calsaverini and Vicente, 2009). In particular, the linear correlation is dependent on the marginal distributions, for instance, it is not invariant under strictly increasing transformations of the marginals, and can hence be misleading as a measure of dependence (Lindskog, 2000).

Copulas

In statistics, a copula is used to formulate a multivariate distribution in such a way that each marginal variable is transformed to have a uniform distribution. In a joint distribution, dependence and the marginal behavior can be separated, and the copula can be considered to be the part describing the dependence structure (Lindskog, 2000).

A copula function $C(u, v)$ can be regarded as the joint cumulative distribution function of two uniformly distributed variables u and v, both in the interval $[0, 1]$, where u and v denote the cumulative distributions, $u = P_x(x)$ and $v = P_y(y)$, of the marginal distributions $p_x(x)$ and $p_y(y)$ (Calsaverini and Vicente, 2009).

The correlation ρ in Eq. (16) can be rewritten in terms of copula densities as:

$$\rho(X, Y) = \int\limits_{[0,1]^2} c(u, v)\, P_x^{-1}(u)\, P_y^{-1}(v)\, du\, dv \ . \tag{23}$$

If X and Y are statistically independent, $c(u, v) = 1$ and $\rho(X, Y) = 0$. However, it is possible that $\rho(X, Y) = 0$, but $c(u, v) \neq 1$, in the case where there is other than linear dependence between X and Y. Moreover, there exist a connection between mutual information and copulas in the case of two continuous random variables

$$I(X; Y) = \int\limits_{[0,1]^2} c(u, v)\ \log c(u, v)\, du\, dv = -h(c) \ , \tag{24}$$

where $h(c)$ is the differential entropy associated with the distribution $c(u, v)$ (Calsaverini and Vicente, 2009).

Dependence modeling with copula functions has been widely used in applications of financial risk assessment, for example in the pricing of collateralized debt obligations (CDOs, Galiani (2003)). As any other measure of dependence, copula too, has the risk of being misused in deceitful hands (Salmon, 2009).

Scale-Invariant Measures

Scale-invariant measures of association such as Kendall's tau (Kendall, 1938) and Spearman's rank correlation (Spearman, 1904) only depend on the copula and are therefore invariant under all strictly increasing transformations of the marginals. Kendall tau coefficient is defined as

$$\tau = (N_c - N_d) / \binom{N}{2} \quad , \tag{25}$$

where N_c is the number of concordant pairs, and N_d is the number of discordant pairs in the data set. Kendall's tau can be expressed in terms of copulas (Calsaverini and Vicente, 2009) as:

$$\tau = 4 \int_{[0,1]^2} C(u, v)\, dC(u, v) \; - 1 \quad . \tag{26}$$

Spearman's rank correlation ρ_S can be seen as computing Pearson's correlation between the ranks of the observations on the two variables (Spearman, 1904), and it can be expressed in terms of copulas (Calsaverini and Vicente, 2009) as:

$$\rho_S = 12 \int_{[0,1]^2} c(u, v)\, uv\, du\, dv \; - 3 \, . \tag{27}$$

3.3 Principal Component Analysis

In order to study mutual dependencies between random variables with a classical linear technique called *Canonical Correlation Analysis* (See Sect. 3.4) it is useful to first understand how another linear technique, *Principal Component Analysis* (PCA), works. In this section, the derivation of PCA is reviewed somewhat in detail in order to alleviate the following of the derivations of CCA properties in Section 3.4.

Principal component analysis is a widely used technique for dimensionality reduction[7](for textbook references, see Bishop (2006); Jolliffe (1986)). It has two equivalent formulations; one is based on orthogonal projections onto a lower dimensional subspace, so that the variance of the projected data is maximized (Hotelling, 1933). The other formulation defines it as the projection that minimizes the mean squared distance between the data points and their projections (Pearson, 1901). The following derivation takes the maximum variance perspective.

We start with an N-sized sample of M-dimensional random vectors $\mathbf{x}_i$. Let us assume the data has been centered as a preprocessing step to simplify the formulas

[7]Also known as the Karhunen-Loève transform.

($\mathbb{E}[\mathbf{x}] = \mathbf{0}$). We start by looking for a projection vector $\mathbf{w}$ that maximizes the variance of the projected scalar variable y,

$$y = \mathbf{w}^T \mathbf{x} \ , \tag{28}$$

where the norm of the projection vector $\mathbf{w}$ is constrained to be $\|\mathbf{w}\| = 1$, to rule out the possibility of increasing the length of the projection vector to infinity in order to maximize the inner product. The variance of y is, by definition,

$$\mathrm{var}[y] = \mathbb{E}[(y - \mu_y)^2] = \mathbb{E}[\mathbf{w}^T \mathbf{x}\mathbf{x}^T \mathbf{w}] = \mathbf{w}^T \mathbf{C}\mathbf{w} \ , \tag{29}$$

where the covariance matrix of $\mathbf{x}$ and its estimate (sample mean) are

$$\mathbf{C} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T \ . \tag{30}$$

By maximizing Equation (29) with constraint $\|\mathbf{w}\| = 1$ using a Lagrange multiplier λ, we get

$$\mathbf{C}\mathbf{w} = \lambda \mathbf{w} \ , \tag{31}$$

which is the eigenvalue equation for matrix $\mathbf{C}$. Therefore, the largest variance is attained when the largest eigenvalue and its corresponding eigenvector are chosen as the (first) principal component, and the variance of the projection becomes equal to the largest eigenvalue

$$\sigma_y^2 = \mathbf{w}^T \mathbf{C}\mathbf{w} = \lambda \mathbf{w}^T \mathbf{w} = \lambda \ . \tag{32}$$

From this formulation it is easy to see that the data variance can be fully described with a decomposition of M orthogonal principal components by using the eigenvalue decomposition of the covariance matrix $\mathbf{C}$,

$$\mathbf{C}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda} \ , \tag{33}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues in the decreasing order, and $\mathbf{W}$ is the corresponding matrix of orthonormal eigenvectors, $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

In general, the problem of PCA dimensionality reduction relates to the approximation problem of finding a matrix $\mathbf{C}_K$ of rank $K < M$ such that the error in Frobenius-norm is minimized,

$$\mathbf{C}_K = \arg\min_{\mathbf{C}_K} \|\mathbf{C} - \mathbf{C}_K\|_{\mathcal{F}} \ , \tag{34}$$

given that $\mathbf{C}_K$ and $\mathbf{C}$ are both $M \times M$ matrices. The Frobenius-norm of a matrix is defined by

$$\|\mathbf{C}\|_{\mathcal{F}} = \sqrt{\sum_i \sum_j |c_{ij}|^2} \ . \tag{35}$$

The solution to this problem is to take the first K eigenvectors of the eigenvalue decomposition of $\mathbf{C}$, where the eigenvalues λ_m are ordered so that $\lambda_m \geq \lambda_{m+1}$ (Eckart and Young, 1936). Thus,

$$\mathbf{C}_K = \mathbf{W}_K \boldsymbol{\Lambda}_K \mathbf{W}_K^T \ , \tag{36}$$

where $\mathbf{W}_K$ is a $M \times K$ matrix containing the first K (column) eigenvectors and $\mathbf{\Lambda}_K$ is the diagonal matrix containing the first K rows and columns of the eigenvalue matrix $\mathbf{\Lambda}$. Additionally, it is known that the approximation error made in Frobenius-norm is exactly the sum of the left-out eigenvalues (Eckart and Young, 1936),[8]

$$\|\mathbf{C} - \mathbf{C}_K\|_{\mathcal{F}} = \sum_{m=K+1}^{M} |\lambda_m| \ . \tag{37}$$

If we now wanted to reduce the dimensionality of $\mathbf{x}$ to K while preserving the variance maximally, we would need to find a $M \times K$-sized projection matrix $\mathbf{W}_K$,

$$\mathbf{y} = \mathbf{W}_K^T \mathbf{x} \ , \tag{38}$$

that maximizes

$$\mathrm{var}[\mathbf{y}] = \mathrm{trace}\left[\mathbf{C}_{\mathbf{y}}\right] = \mathrm{trace}\left[\mathbb{E}[(\mathbf{W}_K^T\mathbf{x})(\mathbf{W}_K^T\mathbf{x})^T]\right] = \mathrm{trace}\left[\mathbf{W}_K^T\mathbf{C}\mathbf{W}_K\right] \ . \tag{39}$$

The solution to this maximization problem is exactly the same as the rank-K-approximation with respect to the Frobenius-norm in Eq. (36): the K eigenvectors corresponding to the K largest eigenvalues of $\mathbf{C}$. Again, the variance included in the projected data and the left-out variance can be expressed in terms of the eigenvalues, as follows:

$$\mathrm{var}[\mathbf{y}] = \mathrm{trace}\left[\mathbf{W}_K^T\mathbf{C}\mathbf{W}_K\right] = \mathrm{trace}\left[\mathbf{\Lambda}_K\right] = \sum_{m=1}^{K} \lambda_m \ . \tag{40}$$

Notice that the projection to principal components without loss of information is $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ or $\mathbf{x} = \mathbf{W}\mathbf{y} = \sum_{m=1}^{M}(\mathbf{w}_m^T\mathbf{y})\mathbf{w}_m$. Taking the rank-$K$-approximation with K principal components yields $\hat{\mathbf{x}} = \sum_{m=1}^{K}(\mathbf{w}_m^T\mathbf{y})\mathbf{w}_m$. Thus, the left-out variance is

$$\begin{aligned}\mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}\|^2] &= \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{m=K+1}^{M}(\mathbf{w}_m^T\mathbf{x}_i)\mathbf{w}_m\right\|^2 = \sum_{m=K+1}^{M}\mathbf{w}_m^T\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{w}_m \\ &= \sum_{m=K+1}^{M}\mathbf{w}_m^T\mathbf{C}\mathbf{w}_m = \sum_{m=K+1}^{M}\lambda_m \ . \end{aligned} \tag{41}$$

These properties and connections of PCA come into play in Section 3.4.4, when generalized CCA is discussed, together with the deflation methods.

[8] The eigenvalues are always real and positive in the case of real symmetric matrices like covariance matrices.

Deflation in PCA

Since the PCA problem reduces to solving an eigenvalue equation, it is possible to look for the components either all at once or one at a time, by deflation. Similar deflation will be needed in different variants of CCA (Sect. 3.4 and 3.4.6) and in its nonparametric generalization NP-DeCA (Sect. 3.6.2).

In PCA the successive components are defined to always be orthonormal to the earlier ones, $\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta. In PCA we choose the direction of each new component to that which maximizes the projected variance amongst all possible directions orthogonal to those already selected. Deflation is the procedure where the already considered variance is removed from the data, so that the next component can be sought by maximizing the projected variance. After extracting the first component $\mathbf{w}_1$, the data is deflated by

$$\begin{aligned}\bar{\mathbf{X}} &= \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_1}\mathbf{X}^T\,\mathbf{w}_1\,\mathbf{w}_1^T\,\mathbf{X}\right)\\ &= \left(\mathbf{I} - \mathbf{w}_1\,\mathbf{w}_1^T\right)\mathbf{X}\,, \qquad (42)\end{aligned}$$

and the next component is searched for by applying the algorithm to the deflated data $\bar{\mathbf{X}}$ instead of the original data. When another component has been found the data $\bar{\mathbf{X}}$ is deflated again, and the process continues until the desired number of components has been reached.

Note that the first line in Equation (42) takes care that the projected variables are orthogonal ($\mathbf{w}_i^T\,\mathbf{X} \perp \mathbf{w}_j^T\,\mathbf{X}$), while the second line states the orthogonality of the projection vectors ($\mathbf{w}_i \perp \mathbf{w}_j$). In the PCA case these two orthogonalities are equivalent. In Canonical Correlation Analysis the two ways to define orthogonality are not, however, equivalent and there only the projected variables are orthogonal ($\mathbf{w}_i^T\,\mathbf{X} \perp \mathbf{w}_j^T\,\mathbf{X}$). See Appendix 8 for more detailed explanation.

3.4 Canonical Correlation Analysis

A classical approach to searching for dependencies between two data sets is to project them both onto a lower-dimensional subspace, in which it is easier to estimate dependencies than in the original high-dimensional spaces. When the projection is chosen in such a way that it maximizes dependency between the two sets of variables, it discards variation that is not present in both sets, while keeping the interesting shared variation. When the projections are restricted to be linear and the dependency is measured by Pearson correlation, the method is the well-known Canonical Correlation Analysis (CCA, Hotelling (1936), see Hardoon et al. (2004); Timm (2002)).

The formulation given for CCA in this section allows straightforward generalization to more than two data sets (*generalized CCA*, Sect. 3.4.2) and, furthermore, avails in showing the connection between mutual information and CCA in Section 3.4.3. A generalization of mutual information to more than two data sets, multi-information, was used in Publication 6 as measure of dependency, and its connection to generalized CCA is also discussed in Section 3.4.3.

3.4.1 Formulation of CCA

Let us denote the original data spaces X_1 and X_2 (with dimensionalities d_1 and d_2, respectively). We assume the means have been removed so that $\boldsymbol{\mu}_i = \mathbb{E}(X_i) = \mathbf{0}$ for both data sets, $i = 1$ and 2. In CCA we look for such linear projections with projection vectors $\mathbf{a}$ and $\mathbf{b}$ that the Pearson correlation between the projections is maximized, i.e.,

$$\max_{\mathbf{a},\mathbf{b}} \rho = \max_{\mathbf{a},\mathbf{b}} \mathbf{a}^T \mathbf{X}_1 \mathbf{X}_2^T \mathbf{b} \tag{43}$$

with normalization constraints $\frac{1}{N}\|\mathbf{a}^T\mathbf{X}_1\|^2 = \frac{1}{N}\|\mathbf{b}^T\mathbf{X}_2\|^2 = 1$.

To obtain the desired formulation of CCA, the two data vectors can be concatenated into one d-dimensional data vector $\mathbf{z}$, whose covariance matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_2 \end{pmatrix} . \tag{44}$$

The blocks on the diagonal are the covariance matrices of the individual data sets and the off-diagonal blocks are the cross-covariances between the data sets[9]. While PCA (Sect. 3.3) works with a single random vector and maximizes the variance of projections of the data, CCA works with a pair of random vectors (or in generalized case with a set of m random vectors) and maximizes the correlation between sets of projections. While PCA leads to an eigenvalue problem, CCA leads to a so-called *generalized eigenvalue problem* (Timm, 2002). CCA reduces to the following generalized eigenvalue problem

$$\begin{pmatrix} 0 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} = \rho \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & \mathbf{C}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} , \tag{45}$$

where ρ is the canonical correlation to be maximized and the diagonal blocks are of size $d_i \times d_i$, and the sizes of the off-diagonal blocks are $d_i \times d_j$[10]. For computational reasons equation (45) is written as (where $\lambda = 1 + \rho$)

$$\begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & \mathbf{C}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} . \tag{46}$$

If we denote the block diagonal of $\mathbf{C}$ by $\mathbf{D}$,

$$\mathbf{D} = \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & \mathbf{C}_2 \end{pmatrix} , \tag{47}$$

we can write Eq. (46) in short as

$$\mathbf{C}\boldsymbol{\xi} = \lambda \mathbf{D}\boldsymbol{\xi} . \tag{48}$$

From this formulation of CCA it is easy to generalize to more than two data sets.

[9] Using the covariance matrices the problem can be rewritten as

$$\max_{\mathbf{a},\mathbf{b}} \rho = \max_{\mathbf{a},\mathbf{b}} \mathbf{a}^T C_{12} \mathbf{b}$$

with normalization constraints $\mathbf{a}^T\mathbf{C}_1\mathbf{a} = \mathbf{b}^T\mathbf{C}_2\mathbf{b} = 1$.

[10] With zero-mean variables $\mathbf{C}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$ and $\mathbf{C}_{xy} = \mathbb{E}[\mathbf{x}\mathbf{y}^T]$, which are in practice replaced with their sample estimates.

3.4.2 Generalizing CCA to Multiple Data Sets

There are several ways to generalize correlation to more than two sets of variables (Kettenring, 1971), leading to several possibilities to generalize CCA to multiple data sets. Here the one chosen by Bach and Jordan (2002) is presented and it is referred to as *generalized CCA* or *gCCA*.

Let us assume that the original data consists of m coupled data sets $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, having dimensionalities $d_1, \ldots, d_m$ respectively. The m coupled data vectors can be concatenated into one d-dimensional data vector $\mathbf{z}$, whose covariance matrix is $\mathbf{C}$, and the corresponding CCA generalization is

$$\begin{pmatrix} \mathbf{C}_1 & \ldots & \mathbf{C}_{1m} \\ \mathbf{C}_{21} & \ldots & \mathbf{C}_{2m} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{m1} & \ldots & \mathbf{C}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_m \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_1 & \ldots & 0 \\ 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \mathbf{C}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_m \end{pmatrix} . \tag{49}$$

This can be written in short as

$$\mathbf{C}\boldsymbol{\xi} = \lambda \mathbf{D} \boldsymbol{\xi} . \tag{50}$$

A procedure for computationally solving this generalized eigenvalue problem with symmetric $\mathbf{C}$ and $\mathbf{D}$ and with positive definite $\mathbf{D}$ can be found, e.g., in (Golub and van Loan, 1996). It is based on the so-called Cholesky decomposition and the symmetric QR-algorithm.

3.4.3 Connection between CCA and Mutual Information

For two vectorial Gaussian variables $\mathbf{x}_1$ and $\mathbf{x}_2$ there is a simple relationship between canonical correlation analysis and the mutual information (Kullback, 1959), as follows:

$$I(\mathbf{x}_1; \mathbf{x}_2) = -\frac{1}{2} \log \left(\frac{\det \mathbf{C}}{\det \mathbf{C}_1 \det \mathbf{C}_2} \right) = -\frac{1}{2} \log V , \tag{51}$$

where V is known as the *generalized variance*. The connection can be derived from the following decomposition of mutual information

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) . \tag{52}$$

We next derive the equation for multi-information between Gaussian data sets, which was used in Publication 6 as means of interpreting the gCCA results. Multi-information is a generalization of mutual information (52) to more than two random variables[11], and has the following decomposition

$$I(X_1; \ldots ; X_m) = \sum_{i=1}^{m} H(X_i) - H(X_1, \ldots , X_m). \tag{53}$$

The individual Gaussian data sets' entropies are

$$\begin{aligned}
H(X_i) &= -\int\limits_{\mathbf{x}_i} p(\mathbf{x}_i) \ln p(\mathbf{x}_i)\, d\mathbf{x}_i \\
&= -\int\limits_{\mathbf{x}_i} p(\mathbf{x}_i) \left(-\frac{d_i}{2} \ln(2\pi) - \frac{1}{2} \ln \det \mathbf{C}_i - \frac{1}{2} \mathbf{x}_i^T \mathbf{C}_i^{-1} \mathbf{x}_i \right) d\mathbf{x}_i \\
&= \frac{1}{2} [\, d_i \ln(2\pi) + \ln \det \mathbf{C}_i \,] \int\limits_{\mathbf{x}_i} p(\mathbf{x}_i)\, d\mathbf{x}_i + \frac{1}{2} \mathbb{E} \left[\mathbf{x}_i^T \mathbf{C}_i^{-1} \mathbf{x}_i \right] \\
&= \frac{d_i}{2} [\, \ln(2\pi) + 1 \,] + \frac{1}{2} \ln \det \mathbf{C}_i \qquad \forall i,
\end{aligned} \tag{54}$$

where d_i denotes the dimensionality of the i'th data set. On the other hand, the joint entropy is

$$\begin{aligned}
H(X_1, \ldots , X_m) &= -\int\limits_{\mathbf{z}} p(\mathbf{z}) \ln p(\mathbf{z})\, d\mathbf{z} \\
&= \frac{d}{2} [\, \ln(2\pi) + 1 \,] + \frac{1}{2} \ln \det \mathbf{C},
\end{aligned} \tag{55}$$

where $\mathbf{C}$ is the joint covariance matrix and d denotes the dimensionality of the joint data set. So, substituting (54) and (55) into equation (52) yields

$$\begin{aligned}
I(X_1; \ldots ; X_m) &= \frac{1}{2} [\ln(2\pi) + 1] \left(\sum_{i=1}^{m} d_i - d \right) + \frac{1}{2} \left(\sum_{i=1}^{m} \ln \det \mathbf{C}_i - \ln \det \mathbf{C} \right) \\
&= -\frac{1}{2} \ln \frac{\det \mathbf{C}}{\det \mathbf{C}_1 \cdots \det \mathbf{C}_m} = -\frac{1}{2} \ln V.
\end{aligned} \tag{56}$$

[11] In order to have a high value, multi-information requires some pairs of the m variables to be dependent in terms of mutual information, but not necessarily all of them.

3.4.4 Dimensionality Reduction by Generalized CCA

In dimension reduction by gCCA the mutual information or multi-information is maximized, i.e., the generalized variance

$$V = \frac{\det \mathbf{C}}{\det \mathbf{C}_1 \cdots \det \mathbf{C}_m} \tag{57}$$

is minimized, while the dimensionality is reduced.

In this section we show that if we choose to look for such a linear transformation that the following four framed properties below are met, and then choose the maximum-variance-projection in the dimensionality reduction, we end up with the CCA generalization of Publication 6 and Tripathi et al. (2008). There have been many suggestions for generalization of CCA to more than two data sets (see for example Bach and Jordan (2002)).

We look for a linear transformation that fulfils these properties:

i) removes the correlations between the within-dataset-variables

ii) normalizes all the variables to have equal variances

iii) normalizes the entropies of individual data sets to constants depending only on the dimensionality of the data set

iv) preserves the multi-information between the data sets (equivalent to preserving generalized variance V)

It will become evident that a procedure of whitening the within-dataset-covariances and maximization of the variance[12] of a linear projection is equivalent to searching for the first generalized canonical correlation and the canonical correlates. This intuition of generalized CCA gives an alternative view to the interpretation of the gCCA results, and was utilized in Publication 6 to study different co-occurrence data sets about yeast stress response. The next section shows how this equivalence can be seen.

Derivation of an Alternative View to CCA

Let us make the following linear transformation to the concatenated variable $\mathbf{z}$ ($\mathbf{z} = [\mathbf{x}_1^T \mathbf{x}_2^T \ldots \mathbf{x}_m^T]$)

$$\mathbf{z}' = \mathbf{D}^{-1/2}\mathbf{z} \quad , \tag{58}$$

where

$$\mathbf{D}^{-1/2} = \begin{pmatrix} \mathbf{C}_1^{-1/2} & 0 & \ldots & 0 \\ 0 & \mathbf{C}_2^{-1/2} & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{C}_m^{-1/2} \end{pmatrix} . \tag{59}$$

[12] or equivalently entropy in the case of Gaussian data

We can first check that the Conditions i) and ii) are met. As a result of the transformation, the covariance matrix of the (transformed) data $\mathbf{C}'$ is

$$\mathbf{C}' = \mathbb{E}[\mathbf{z}'\mathbf{z}'^T] = \mathbf{D}^{-1/2}\mathbb{E}[\mathbf{z}\mathbf{z}^T]\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{C}\mathbf{D}^{-1/2} \quad . \tag{60}$$

In effect, we have now whitened the individual original data sets $\mathbf{X}_1, \ldots, \mathbf{X}_m$ to get $\mathbf{X}'_1, \ldots, \mathbf{X}'_m$ (Condition i) is met). Additionally, the variances of all the variables are now equal, i.e., Condition ii) is met, and what is more, they are all equal to one[13].

In order to check the Conditions iii) and iv), we need to write down the entropies of both the individual data sets and the concatenated data after the transformation. If the data is Gaussian, after the transformation the entropy of an individual data set is

$$\begin{aligned} H(X'_i) &= \frac{d_i}{2}\left[\ln(2\pi) + 1\right] + \frac{1}{2}\ln\det\mathbf{I} \\ &= \frac{d_i}{2}\left[\ln(2\pi) + 1\right] \qquad\qquad \forall i. \end{aligned} \tag{61}$$

Hence, the entropy of an individual data set is dependent only on the dimensionality of the data set, d_i, as required in Condition iii). On the other hand, the joint entropy of the transformed (concatenated) Gaussian data is

$$\begin{aligned} H(X'_1, \ldots, X'_m) &= \frac{d}{2}\left[\ln(2\pi) + 1\right] + \frac{1}{2}\ln\det\mathbf{C}' \\ &= \frac{d}{2}\left[\ln(2\pi) + 1\right] + \frac{1}{2}\ln\frac{\det\mathbf{C}}{\det\mathbf{C}_1\cdots\det\mathbf{C}_m}, \end{aligned} \tag{62}$$

because

$$\begin{aligned} \det\mathbf{C}' &= \det(\mathbf{D}^{-1/2})\det(\mathbf{C})\det(\mathbf{D}^{-1/2}) = \det(\mathbf{C})\det(\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}) \\ &= \det(\mathbf{C})\det(\mathbf{D}^{-1}) \\ &= \frac{\det\mathbf{C}}{\det\mathbf{C}_1\cdots\det\mathbf{C}_m}. \end{aligned} \tag{63}$$

We can now proceed to checking that the multi-information (or mutual information) is preserved in the transformation (58), required in Condition iv), as follows:

$$\begin{aligned} I(X'_1; \ldots; X'_m) &= \sum_{i=1}^{m} H(X'_i) - H(X'_1, \ldots, X'_m) \\ &= -\frac{1}{2}\ln\det\mathbf{C}' \\ &= -\frac{1}{2}\ln\frac{\det\mathbf{C}}{\det\mathbf{C}_1\cdots\det\mathbf{C}_m} = -\frac{1}{2}\ln V \\ &= \sum_{i=1}^{m} H(X_i) - H(X_1, \ldots, X_m) \\ &= I(X_1; \ldots; X_m). \end{aligned} \tag{64}$$

[13]Note that it is possible to construct other related methods by choosing another whitening transformation than the one in Eq. (58). One could equally well choose to weight each variable differently or just remove the between-dataset-correlations, but preserve the variances of the within-dataset-variables. Then, the block diagonal of the covariance matrix $\mathbf{C}'$ would consist of diagonal matrices instead of identity matrices. However, our choice is the one that coincides with gCCA.

In order to perform dimensionality reduction, we look for the maximum-variance-direction after having removed the within-dataset-correlations. This should emphasize the dependencies *between* the data sets in contrast to dependencies within the data sets. Therefore, in the transformed feature space we look for the one-dimensional projection $\mathbf{z}'' = \mathbf{a}^T\mathbf{z}'$ (with normalization $\mathbf{a}^T\mathbf{a} = 1$) that maximizes the variance of the projected variable $\mathbb{E}[\mathbf{z}''\mathbf{z}''^T]$. The (co)variance of the projected variable $\mathbf{z}''$ is

$$\mathbf{C}'' = \mathbb{E}[\mathbf{a}^T\mathbf{z}'(\mathbf{z}')^T\mathbf{a}] = \mathbf{a}^T\mathbf{C}'\mathbf{a}\,, \tag{65}$$

and since $\mathbf{a}^T\mathbf{C}'\mathbf{a}$ is the variance of $\mathbf{z}'$ in the direction of $\mathbf{a}$, it attains its maximum when $\mathbf{a}$ is the first principal component of $\mathbf{C}'$.[14]

So, actually, this procedure of whitening the within-dataset-covariances and maximization of the variance of a linear projection is equivalent to searching for the first generalized canonical correlation and the canonical correlates, i.e., solving the following generalized eigenproblem

$$\mathbf{C}\boldsymbol{\xi} = \lambda\mathbf{D}\boldsymbol{\xi}\,,$$

since, if we denote $\mathbf{C} - \lambda\mathbf{D}$ by $\mathbf{A}$:

$$\begin{aligned}
\mathbf{C} - \lambda\mathbf{D} &= \mathbf{A} && (66)\\
\Leftrightarrow \mathbf{D}^{-1/2}\,\mathbf{C} - \lambda\,\mathbf{D}^{1/2} &= \mathbf{D}^{-1/2}\,\mathbf{A} &&\\
\Leftrightarrow \mathbf{D}^{-1/2}\,\mathbf{C}\,\mathbf{D}^{-1/2} - \lambda\,\mathbf{I} &= \mathbf{D}^{-1/2}\,\mathbf{A}\,\mathbf{D}^{-1/2} &&\\
\Leftrightarrow \mathbf{C}' - \lambda\,\mathbf{I} &= \mathbf{D}^{-1/2}\,\mathbf{A}\,\mathbf{D}^{-1/2}\,, && (67)
\end{aligned}$$

and take determinants on both sides of the first and the last equation, we get

$$\det(\mathbf{C} - \lambda\mathbf{D}) = 0 \Leftrightarrow \det\mathbf{A} = 0 \Leftrightarrow \det(\mathbf{C}' - \lambda I) = 0\,. \tag{68}$$

Hence, the solution of the generalized eigenproblem in CCA is equal to the solution of the eigenvalue problem $\mathbf{C}'\boldsymbol{\xi} = \lambda\boldsymbol{\xi}$, i.e., the PCA solution for the within-dataset-whitened variable $\mathbf{z}'$.

3.4.5 Deflation in CCA

The CCA components can also be computed either all at once, or one at a time iteratively, by deflation like in PCA (See Section 3.3). The normalization constraints of CCA are $\mathbf{w}_x^T\,\mathbf{C}_x\mathbf{w}_x = \mathbf{w}_y^T\,\mathbf{C}_y\mathbf{w}_y = 1$, and the orthogonality constraints hold for the projected variables $(\mathbf{w}_x^i)^T\,\mathbf{X} \perp (\mathbf{w}_x^j)^T\mathbf{X}$, but this is no longer equivalent to the orthogonality of the projection vectors $\mathbf{w}_x^i$ and $\mathbf{w}_x^j$, as was the case in PCA, in Eq. (42). (See Appendix 8 for details.)

After extracting the first component $\mathbf{w}_x$, the data is transformed by

$$\begin{aligned}
\bar{\mathbf{X}} &= \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_x^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_x}\mathbf{X}^T\,\mathbf{w}_x\,\mathbf{w}_x^T\,\mathbf{X}\right)\\
&= \mathbf{X}\left(\mathbf{I} - \frac{1}{N}\mathbf{X}^T\,\mathbf{w}_x\,\mathbf{w}_x^T\,\mathbf{X}\right)\,. && (69)
\end{aligned}$$

[14]For Gaussian data this is equivalent with maximization of the entropy of the projected variable $\mathbf{z}''$

$$H(Z'') = \frac{1}{2}\left[\ln(2\pi) + 1\right] + \frac{1}{2}\ln\det\mathbf{C}'' = \frac{1}{2}\ln\mathbf{a}^T\mathbf{C}'\mathbf{a} + \text{const.}$$

The deflation is performed analogously for $\mathbf{Y}$. The next component is sought by applying the algorithm to $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ instead of the original data. This procedure can be continued up to the minimum of the data dimensionalities, or until there are no significant dependencies left between $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$.

3.4.6 Sparse and Non-Negative Variants of CCA

It is well-known that CCA for one projection vector in the case of two data sets can also be formulated as an equivalent norm minimization problem:

$$(\mathbf{w}_x, \mathbf{w}_y) = \underset{\mathbf{w}_x, \mathbf{w}_y}{\arg\min} \|\mathbf{w}_x^T \mathbf{X} - \mathbf{w}_y^T \mathbf{Y}\|^2 \quad , \tag{70}$$

with normalization $\frac{1}{N}\|\mathbf{w}_x^T \mathbf{X}\|^2 = \frac{1}{N}\|\mathbf{w}_y^T \mathbf{Y}\|^2 = 1$ and orthogonality constraints for the successive components $\mathbf{w}_x^{(i)T}\mathbf{X} \perp \mathbf{w}_x^{(j)T}\mathbf{X}$ (and analogously for $\mathbf{w}_y$). In this formulation successive components have to be sought iteratively by deflation (Vía et al., 2005). This formulation opens up new ways to constrain the optimization, e.g., by requiring the components to be non-negative or sparse, or both at the same time (Sigg et al., 2007). Sparsity of the components can be achieved using LASSO-like $L1$-norm-based sparsity constraints (Tibshirani, 1996). So-called non-negative matrix factorization has also been introduced as a way to restrict the projection weights to be non-negative (Lee and Seung, 1999, 2001). Another variant of CCA is the so-called functional CCA, which has also been used in fMRI brain imaging (He et al., 2003).

Kernel-CCA has been suggested to be used when the connection between the two data sets could be nonlinear, where the nonlinearity can be embedded in the kernel (Fyfe and Lai, 2000; Shawe-Taylor and Cristianini, 2004). However, it requires quite cautious regularization in order not to overfit (Fukumizu et al., 2007).

3.5 Probabilistic Extensions of CCA

Since classical CCA implicitly assumes the two (or more) sets of random variables to be normally distributed, in which case the shared variance coincides with mutual information, there is clearly a need for extensions that allow other than normally distributed data. In Publications 6, 7, 8, and 9 we used information-theoretic extensions of CCA (see Sect. 3.6). To give a more complete picture of the closest related methods let us first look at the probabilistic extensions of CCA.

The probabilistic extensions of CCA are built on the basis of the probabilistic extensions of PCA. In the next section the underlying PCA extension is presented, and in Sections 3.5.2 and 3.5.3 both a maximum likelihood probabilistic extension of CCA (Bach and Jordan, 2005) and a fully Bayesian CCA (Klami and Kaski, 2007) are presented. Recently, other probabilistic extensions have been suggested, including probabilistic sparse PCA and sparse CCA (Archambeau and Bach, 2009).

3.5.1 Probabilistic PCA

A generative model for PCA has been introduced by Tipping and Bishop (1999), and its graphical model representation can be seen in Figure 5. This presentation follows the textbook reference (Bishop, 2006).

In this section the dimensionality of the observed random variable $\mathbf{x}$ is denoted by M_x and the dimensionality of the paired observable $\mathbf{y}$ by M_y. The data samples are indexed by i, $(i = 1, \ldots, N)$, and the components by k with $k = 1, \ldots, K$. As before, $\boldsymbol{\varphi}$ denotes the set of all model parameters. First, a $K \times 1$ latent variable $\mathbf{z}$ is drawn from the standard normal distribution

$$\mathbf{z} \mid \boldsymbol{\varphi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ . \tag{71}$$

Given the latent variable $\mathbf{z}$, the observed variable $\mathbf{x}$ follows the normal distribution of Eq. (72)

$$\mathbf{x} \mid \mathbf{z}, \boldsymbol{\varphi} \sim \mathcal{N}\left(\mathbf{W}\,\mathbf{z} + \boldsymbol{\mu}, \sigma^2\,\mathbf{I}\right) \ , \tag{72}$$

where the matrix $\mathbf{W}$ contains the K projections that define the principal components. The $M_x \times 1$ mean vector of $\mathbf{x}$ is denoted by $\boldsymbol{\mu}$, and σ^2 is the common variance of each element of $\mathbf{x}$. For optimization of the model there are, e.g., EM-algorithms introduced by Roweis (1998).

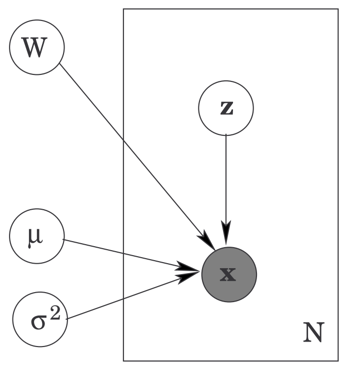


Figure 5: Graphical model representation of the probabilistic PCA. The plate represents the repeated choice of N data samples $\mathbf{x}$. The matrix $\mathbf{W}$ contains the K projection vectors that define the principal components. The mean of $\mathbf{x}$ is denoted by $\boldsymbol{\mu}$, and there is a common variance σ^2 for of each element of $\mathbf{x}$.

3.5.2 Probabilistic CCA

On the basis of probabilistic PCA, a probabilistic CCA has been introduced (Bach and Jordan, 2005). The model is represented as a graphical model in Figure 6. As in generative PCA, first, a $K \times 1$ latent variable $\mathbf{z}$ is drawn from the standard normal distribution

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ . \tag{73}$$

Given the latent component $\mathbf{z}$, the observed variables $\mathbf{x}$ and $\mathbf{y}$ follow the normal distributions

$$\begin{aligned} \mathbf{x} \mid \mathbf{z}, \boldsymbol{\varphi} &\sim \mathcal{N}(\mathbf{W}_x \mathbf{z} + \boldsymbol{\mu}_x, \boldsymbol{\Psi}_x) \\ \mathbf{y} \mid \mathbf{z}, \boldsymbol{\varphi} &\sim \mathcal{N}(\mathbf{W}_y \mathbf{z} + \boldsymbol{\mu}_y, \boldsymbol{\Psi}_y) \quad , \end{aligned} \tag{74}$$

where $\mathbf{W}_x$ and $\mathbf{W}_y$ are matrices containing the K projection vectors $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$ as their columns. These projections define the K canonical components. The $M_x \times 1$ mean vector of $\mathbf{x}$ is denoted by $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ is the $M_y \times 1$ mean vector of $\mathbf{y}$. $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively.

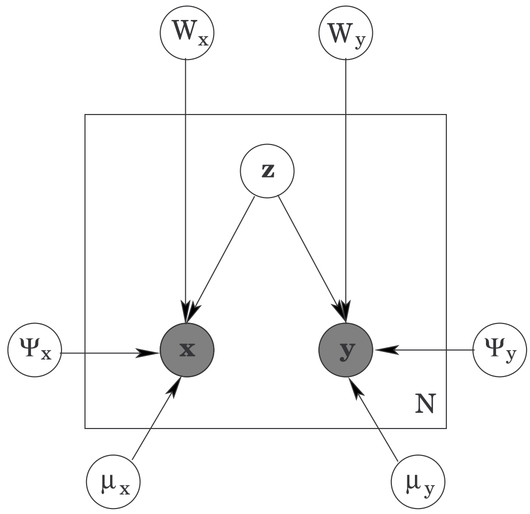


Figure 6: Graphical model representation of the probabilistic CCA. The plate represents the repeated choice of N paired data samples $\mathbf{x}_i$ and $\mathbf{y}_i$. The matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ contain as their columns the K projections that define the canonical components. The mean vectors of $\mathbf{x}$ and $\mathbf{y}$ are denoted by $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$. $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively.

The maximum likelihood solution of this model has been shown (Bach and Jordan, 2005) to relate to the classical CCA via the following equations

$$\left\{ \begin{aligned} \mathbf{W}_x &= \mathbf{C}_x \mathbf{U}_x \mathbf{Q}_x \\ \mathbf{W}_y &= \mathbf{C}_y \mathbf{U}_y \mathbf{Q}_y \end{aligned} \right. \quad , \tag{75}$$

where $\mathbf{C}_x$ and $\mathbf{C}_y$ are the empirical covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{U}_x$ is the $M_x \times K$ matrix containing the CCA projection vectors as columns ($\mathbf{U}_y$ analogously for $\mathbf{y}$). One should note, however, that the ML-solution does not define the rotation of the K-dimensional subspace, but allows $\mathbf{Q}_x$ and $\mathbf{Q}_y$ to define the rotation. $\mathbf{Q}_x$ and $\mathbf{Q}_y$ are random matrices of size $K \times K$, with spectral norms smaller than one, and satisfying $\mathbf{Q}_x\,\mathbf{Q}_y^T = \mathbf{P}$, where $\mathbf{P}$ is a $K \times K$ diagonal matrix containing the corresponding canonical correlations. Archambeau et al. (2006) have presented a method for solving the rotational ambiguity caused by $\mathbf{Q}_x$ and $\mathbf{Q}_y$. This makes it possible to find the projections of the classical CCA, instead of just the subspace.

3.5.3 Bayesian CCA

The probabilistic CCA opens interesting possibilities to interpret the model but a fully Bayesian treatment is needed in order to estimate posterior distributions instead of just maximum likelihood point estimate – which is anyway typically more efficient to reach just by solving the original eigenvalue problem. In this thesis, the model by Klami and Kaski (2007) is presented, but there are also other possible ways to define the model distributions.[15]

As before, a $K \times 1$ latent variable $\mathbf{z}$ is drawn from the standard normal distribution

$$\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)\ . \tag{76}$$

Again, given the latent variable $\mathbf{z}$, the observed variables $\mathbf{x}$ and $\mathbf{y}$ follow the normal distributions

$$\begin{aligned}\mathbf{x} \mid \mathbf{z}, \boldsymbol{\varphi} &\sim \mathcal{N}\left(\mathbf{W}_x\,\mathbf{z} + \boldsymbol{\mu}_x, \boldsymbol{\Psi}_x\right)\\ \mathbf{y} \mid \mathbf{z}, \boldsymbol{\varphi} &\sim \mathcal{N}\left(\mathbf{W}_y\,\mathbf{z} + \boldsymbol{\mu}_y, \boldsymbol{\Psi}_y\right)\end{aligned}\quad , \tag{77}$$

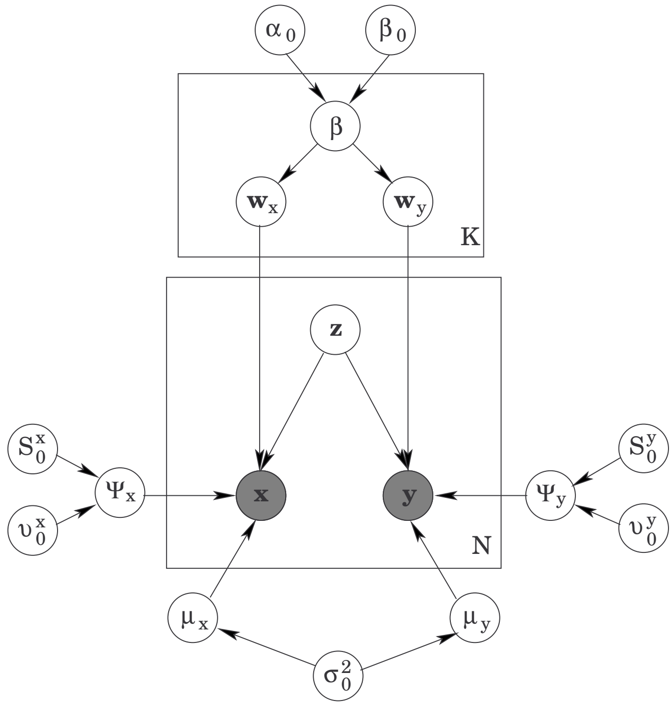


Figure 7: Graphical model representation of the Bayesian CCA. The lower plate represents the repeated choice of N paired data samples $\mathbf{x}_i$ and $\mathbf{y}_i$. The upper plate labeled with K represents the different canonical components, defined by K pairs of projection vectors.

[15] A similar Bayesian extension to probabilistic PCA has been introduced by Bishop (1999).

where $\mathbf{W}_x$ and $\mathbf{W}_y$ are matrices containing the K projection vectors $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$ as their columns. These projections define the K canonical components. The mean vectors of $\mathbf{x}$ and $\mathbf{y}$ are denoted by $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, and $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are the corresponding covariance matrices. But, in contrast to the above probabilistic CCA, all these model parameters are drawn from their prior distributions. For computational convenience, conditionally conjugate prior distributions are used for all parameters. Thus, the means and covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ are drawn as follows

$$\begin{aligned} \boldsymbol{\mu}_x &\sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}\right) \quad \text{and} \quad \boldsymbol{\mu}_y \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}\right) \\ \boldsymbol{\Psi}_x &\sim \mathrm{IW}\left(\mathbf{S}_0^x, \nu_0^x\right) \quad \text{and} \quad \boldsymbol{\Psi}_y \sim \mathrm{IW}\left(\mathbf{S}_0^y, \nu_0^y\right) , \end{aligned} \tag{78}$$

where $\mathrm{IW}(\mathbf{S}, \nu)$ is the Inverse Wishart distribution with parameters $\mathbf{S}$ and ν. For each component k, a common variance for both $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$ is drawn from

$$\beta_k \sim \mathrm{IG}\left(\alpha_0, \beta_0\right), \tag{79}$$

where $\mathrm{IG}(\alpha, \beta)$ is the Inverse Gamma distribution with parameters α and β. Finally, the $M_x \times 1$ linear projection $\mathbf{w}_x^k$ for $\mathbf{x}$ and $M_y \times 1$ linear projection $\mathbf{w}_y^k$ for $\mathbf{y}$ for component k can be drawn as follows

$$\mathbf{w}_x^k \mid \beta_k \sim \mathcal{N}\left(\mathbf{0}, \beta_k \mathbf{I}\right) \quad \text{and} \quad \mathbf{w}_y^k \mid \beta_k \sim \mathcal{N}\left(\mathbf{0}, \beta_k \mathbf{I}\right) . \tag{80}$$

They are gathered as columns into matrices $\mathbf{W}_x$ and $\mathbf{W}_y$.

In (Klami and Kaski, 2007) this model was evaluated by Gibbs sampling. A variational Bayesian approach to CCA has been presented by Wang (2007), with the subtle difference that the β_k are not shared between $\mathbf{w}_x^k$ and $\mathbf{w}_y^k$.

3.6 Information-Theoretic Extensions of CCA

Replacing correlation with mutual information makes the discovery of more general types of dependency possible, extending the range of applications. However, mutual information cannot be computed as easily as correlation, and approximations are therefore needed. In this section we describe such approximations used in the publications of this thesis; Associative Clustering (Publication 6) and Nonparametric Dependent Component Analysis (Publication 8 and extended in Publication 9).

3.6.1 Associative Clustering

Associative Clustering (AC, Kaski et al. (2005b)) is a method that clusters two continuous-valued multi-dimensional variable spaces X and Y, by maximizing their dependencies. The dependencies are modeled based on observed paired data without prior knowledge about the structure of the data sets. The method effectively looks for compact areas of data points in both spaces with the property that the pairs of the clustered data points are also clustered in the other data space. The compactness assures a touch of internal homogeneity within the clusters, which makes them more interpretable.

The dependencies between the partitionings of the spaces are presented with a contingency table, where the slots result from the Cartesian product of the one-way partitions of X and Y, respectively, and the count of each slot reflects the data density in that area. Traditionally, the so-called Bayes-factor (81) has been used as a measure of dependency between the marginals of a given contingency table but in AC the Bayes factor is used to select the partitions so as to maximize the Bayes factor

$$
\begin{aligned}
\frac{P(\mathcal{D} \mid \bar{H})}{P(\mathcal{D} \mid H)} &= \frac{P(\bar{H} \mid \mathcal{D})}{P(H \mid \mathcal{D})} \cdot \frac{P(H)}{P(\bar{H})} \\
&= \frac{P(\bar{H} \mid \{n_{ij}\})}{P(H \mid \{n_{ij}\})} \propto \frac{\prod_{ij} \Gamma(n_{ij}+1)}{\prod_i \Gamma(n_{i.}+1) \prod_j \Gamma(n_{.j}+1)} \quad ,
\end{aligned}
\tag{81}
$$

where we have assumed the priors of the hypotheses to be constant (for details see Kaski et al. (2005b)). The hypothesis H is the null hypothesis that assumes independent marginal distributions and $\bar{H}$ is the interesting hypothesis that shows dependency. We denote by n_{ij} the count of samples in the (i, j)'th bin of the contingency table, and by $\{n_{ij}\}$ the set of all counts in the table. Γ denotes the gamma function, which has the property $\Gamma(n) = (n-1)!$ for positive integers n.

The larger the Bayes-factor is, the more dependent the marginal clusters are. The technical details of the optimization of the AC model can be found in Sinkkonen et al. (2005).

3.6.2 Nonparametric Dependent Component Analysis

A straightforward option to estimating mutual information between data sets is to empirically estimate the probability density in the projection space, and estimate mutual information based on the density estimate. Fisher III and Darrell (2004); Klami and Kaski (2005) and Yin (2004) have introduced methods based on nonparametric Parzen-kernel estimates. We call these methods Nonparametric Dependent Component Analysis (NP-DeCA), since they search for general statistical dependencies. By Dependent Component Analysis (DeCA) we refer to a broader range of dependence seeking algorithms, including also CCA and its different variants.

The fundamental task in NP-DeCA (Klami and Kaski, 2005) is to find linear projections of two data sets[16], $\mathbf{X}$ and $\mathbf{Y}$, so that the mutual information between the projections $\mathbf{s}_x = \mathbf{w}_x^T \mathbf{X}$ and $\mathbf{s}_y = \mathbf{w}_y^T \mathbf{Y}$ is maximized. The objective is thus to maximize

$$
I(s_x, s_y) = \int\int p(s_x, s_y) \log \frac{p(s_x, s_y)}{p(s_x)\, p(s_y)}\, ds_x ds_y \tag{82}
$$

with respect to linear transformations $\mathbf{w}_x$ and $\mathbf{w}_y$. Here s_x and s_y are the random variables $s_x = \mathbf{w}_x^T \mathbf{x}$ and $s_y = \mathbf{w}_y^T \mathbf{y}$, and $\mathbf{x}$ and $\mathbf{y}$ are random vectors corresponding to data sets $\mathbf{X}$ and $\mathbf{Y}$.

[16] The method can be easily extended to more than two data sets, as was done for Nonparametric DeCA in (Klami and Kaski, 2005), using multi-information instead of mutual information as the objective function.

In practice, the integral is estimated with a sum

$$\hat{I}(s_x, s_y) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{p}(s_x^i, s_y^i)}{\hat{p}(s_x^i)\,\hat{p}(s_y^i)} \,, \tag{83}$$

where N is the number of observations and $\hat{p}(s_x, s_y)$ is a Parzen-estimate, which is nonparametric. Hence, optimizing the cost of Eq. (82) is straightforward; it only requires deriving the gradient of the cost and using any standard optimization technique to find a local optimum.

NP-DeCA can be used instead of CCA when we have reason to believe that the data is not normally distributed. In Publication 8 it was applied to brain imaging analysis, where dependencies between stimulus time series and spatially independent brain activity patterns were sought using NP-DeCA. The application is discussed in Section 6.6.

3.6.3 Fast Semi-Parametric Extension of NP-DeCA

In the original NP-DeCA algorithm (Klami and Kaski, 2005) the density estimation of the data sets was done in a nonparametric fashion by Parzen-estimates. The Parzen-estimates are consistent and accurate, but unfortunately computationally demanding for large data sets. The computation can be speeded up by using only a subset of data points as kernels, as suggested by Klami and Kaski (2005), but replacing the nonparametric density estimates with semi-parametric estimates should give more accurate results.

In Publication 9, we introduced a faster semi-parametric variant of the NP-DeCA-algorithm, which uses a mixture of Gaussians to estimate the density in the projection space, and where the integral in (82) is estimated as an average over the observed data points. An analogous method has earlier been shown to improve efficiency in a related modeling task of finding supervised linear projections, that is, projections informative of co-occurring categorical variables (Peltonen et al., 2007).

We applied the novel Semi-Parametric DeCA (SP-DeCA) to the task of finding dependencies between measured brain activity and multi-sensory stimuli in Publication 9. The application is discussed in Section 6.6.

The main advantage of the method proposed in Publication 9, compared to earlier nonparametric DeCA methods, is in the computational speed. The Parzen-kernel estimate used in the earlier works has a computational complexity of $\mathcal{O}(N^2)$, where N is the number of training data points, and each iteration of a gradient-based optimization algorithm requires evaluating the densities for a new set of projected values. A mixture model with K mixture densities, however, has only a complexity of $\mathcal{O}(NK)$ for evaluating the density or the gradient with respect to the projections. The disadvantage of the parametric estimate is that we need to learn the parameters of the mixture model as well, but in practice the number of iterations required for convergence is small, and good performance can be achievable already with a very small K. The resulting method is considerably faster already in applications with some hundreds of data points.

Cost and Optimization of Semi-Parametric DeCA

In this semi-parametric case we consider parametric estimates of the form

$$\hat{p}(s_x, s_y \mid \boldsymbol{\varphi}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}([s_x; s_y] \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ , \tag{84}$$

where $\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$. The π_k's represent the probabilities of the mixture components, and are therefore non-negative values that sum up to unity. This estimate has a set of parameters $\boldsymbol{\varphi} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, which need to be learned. Hence, straightforward optimization of the objective function

$$\hat{I}(s_x, s_y) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{p}(s_x^i, s_y^i)}{\hat{p}(s_x^i)\, \hat{p}(s_y^i)} \tag{85}$$

with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ is not possible.

In Publication 9, we have proposed an alternating algorithm following the work in Peltonen et al. (2007). Starting with some initial projections, a mixture of Gaussians in the projection space is learned using the expectation maximization (EM) algorithm (Sect. 2.1.2). After learning the density estimate, the projections $\mathbf{w}_x$ and $\mathbf{w}_y$ are optimized. The algorithm then proceeds by alternating these two steps until convergence.

Given a fixed density estimate, the objective function Eq. (85) can easily be differentiated, and gradient-based methods can be used to learn the projections. In Publication 9 we used a conjugate-gradient method, with the number of iterations equal to the dimensionality of the parameter space. The density estimate was always optimized until convergence of the EM algorithm.[17]

Deflation in Semi-Parametric DeCA

The components are optimized one at a time. This is done because density estimation in high-dimensional spaces is very difficult. By restricting to one-dimensional projections the joint density $p(s_x, s_y)$ will be estimated in a two-dimensional space, which can be done accurately enough already with reasonably small data sets.

After finding the first component, the next maximally dependent component is sought with the following constraint: The projections on the consecutive components are required to be independent of the projections on the previous components, in both the X- and Y-spaces.

In practice, searching for a component that maximizes dependency with the other data set while minimizing dependency with the earlier component(s) is difficult. It could be done by adding a separate term in the cost function, but in practice there is a computationally more efficient approximation available.

[17] In practice it only took a few steps, since we started from the previous estimate, which was typically very close to a local optimum.

Instead of full independence, we require the components to be uncorrelated with the earlier projections, analogously to how successive CCA components are defined. This can be satisfied with a simple deflation procedure, as follows.[18]

After extracting the first component $\mathbf{w}_x$, the data is transformed by

$$\bar{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_x^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_x}\mathbf{X}^T\,\mathbf{w}_x\,\mathbf{w}_x^T\,\mathbf{X}\right) \tag{86}$$

(and analogously for $\mathbf{Y}$), and the second component is searched for by applying the algorithm to $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ instead of the original data. This procedure can be continued up to the minimum of the X- and Y-space dimensionalities, or until there are no significant dependencies left between $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$.

3.7 Summary

This chapter first motivated dependency modeling between data sets as a way to define what is relevant in the data. A lot of background was given about the classical methods PCA and CCA to remind the reader about certain properties of the methods as a prerequisite for the extended work. Also probabilistic extensions of PCA and CCA were briefly discussed to get a full picture of the closest related methods.

An alternative view to CCA was derived, as whitening of the within-dataset-covariances followed by PCA to search for those features that have maximal variance and at the same time are dependent on both (or all the) datasets. This alternative interpretation led us to a new way of using CCA in feature selection in Publication 6.

As classical CCA has the restriction of implicitly assuming the data to be normally distributed, we wanted to find an unsupervised dependency-seeking method without the assumption of normally distributed data. We have developed an extension of CCA to non-normally distributed data (NP-DeCA) and it was applied to searching dependencies between two co-occurring data sets in Publication 8. However, the Parzen-estimate-based Nonparametric DeCA is not efficient enough for large data sets, and therefore, we pursued to find a faster algorithm for the task. In Publication 9 we introduced a new faster variant of the method, Semi-Parametric DeCA (SP-DeCA) where the density estimation is based on Mixture of Gaussians.

[18] If maximal variance direction would determine the direction of the next component, this would exactly match the orthogonality constraint of CCA. In our case the deflation does not strictly guarantee uncorrelatedness of the successive components, but takes care that correlated successive components are highly improbable.

4 Modeling of User Interest

User modeling is generally understood as a broad field of studying and developing systems that acquire information about a user (or group of users) so as to be able to adapt the system's behavior to that user or user group. Numerous applications of such systems exist, e.g., in the area of natural language understanding and dialogue systems, in computer-based educational systems and online learning environments, and in recommender systems for e-commerce, news and entertainment. This thesis focuses on the specific sub-genre of recommender systems.

4.1 Feedback in Recommender Systems

Our application areas have been related to the so-called *recommender systems*, where the task is to predict the subjective relevance of new items to the user based on his or her learned user profile. In Publication 2 we used content-based filtering (Sect. 4.2), whereas and in Publications 3, 4 and 5 collaborative filtering (Sect. 4.3) was used.

In recommender systems users give feedback about the recommended items – e.g., music pieces or albums (Shardanand and Maes, 1995), movies, books, news (Jokela et al., 2001), restaurants, wines (e.g. Viinitupa, a recommender system prototype for wines[19]) etc. – and the system gathers a user interest profile for each user. The system matches the user interest profile against a database of items in order to find more items that match the interest profile. The most reliable way to gather user profiles is based on so-called *explicit feedback* where users give ratings to the items on a given scale (e.g., 0–5 stars). Because explicit feedback is laborious to the users and typically at least tens of ratings are needed in order to make reasonable recommendations, there has been various attempts to find out about the user interests without bothering the user – using so-called *implicit feedback*. If a user buys a book, bookmarks a product page or follows a link, it can be inferred to convey information about his interest profile.

Explicit and implicit feedback can be also used in an information retrieval setup. The task of an information retrieval (IR) system is to identify documents that best match the query given by users, based on the contents of the documents. The systems may additionally collect explicit relevance feedback from the user, by asking which of the retrieved documents were relevant. Traditionally, implicit feedback in information retrieval has been derived from document reading time, or by monitoring user behavior: saving, printing, and selecting of documents (Kelly and Teevan, 2003).

In Publication 1, we inferred user interest implicitly from eye movements using probabilistic models that predict whether a user finds a text relevant, given her eye movement trajectory while reading the text. The key assumption motivating the use of eye movements is that attention patterns correlate with relevance, and that attention patterns are reflected in eye movements (see Porta (2007); Salojärvi et al. (2003)). The application is discussed further in Chapter 5.

[19] http://www.soberit.hut.fi/∼jti/winemag.htm

4.2 Content-Based Filtering

In so-called *content-based filtering*, documents or other items are filtered for the user based on his or her interest profile. But in contrast to so-called *collaborative filtering* (Sect. 4.3), the opinions of other users do not affect the recommendations, only the content descriptions of the items matter. The system maintains user profiles and matches the profiles against a database of items in order to find more items that match each interest profile. The profile can be defined in advance or adapted based on gradually gathered explicit or implicit feedback about the items of interest.

In case of text documents, the actual words in the vectorial form of bag-of-words might be the content (see Sect. 4.4.3). In the case of music recommendations some audio signal analysis of the sound could serve as a content-based description (Eck et al., 2008; Lamere and Eck, 2007; West et al., 2006). In case of news reporting, metadata that categorizes news articles in various ways is being typically defined (Jokela, 2001; Savia, 1999; Tintarev and Masthoff, 2006; Turpeinen, 2000).

In Publication 2, we looked for such textual features from movie synopses that best separated between the 10 available movie genres, and used them for content-based filtering for individual users. Finding such textual features would help in relevance prediction for future movies that would not have the same genre classification available.

4.3 Collaborative Filtering

Collaborative filtering is another way to make recommendations based on user profiles. It is suitable for situations where one might expect the opinions of other similar-minded people to be relevant, e.g., in matters of taste. It is assumed that if two users have liked a group of same items – let us say books – they are likely to share this liking also for other books in the future. The goal of collaborative filtering is to predict the relevance of an item to a given user, based on a database of explicit or implicit relevance ratings from a large population of users.

Traditionally, collaborative filtering has been performed by so-called memory-based techniques, in which one first identifies users similar to a given user and then gives predictions based on interests of those similar users (see, e.g., GroupLens (Konstan et al., 1997), or Ringo (Shardanand and Maes, 1995)). However, the time and memory requirements of the memory-based techniques do not generally scale well as the number of users and documents increases, which has led to development of model-based approaches (Bohnert et al., 2009; Hofmann, 2004; Jin and Si, 2004; Wettig et al., 2003; Zitnick and Kanade, 2004). However, specific scalable memory-based techniques have also been introduced (Bell and Koren, 2007).

An interesting family of models are the latent topic models, which have been successfully used in document modeling but also in collaborative filtering (Blei et al., 2003; Blei and Jordan, 2003; Erosheva et al., 2004; Hofmann, 2004; Keller and Bengio, 2004; Marlin, 2004a; Marlin and Zemel, 2004; McCallum et al., 2004; Popescul et al., 2001; Pritchard et al., 2000; Rosen-Zvi et al., 2004; Salakhutdinov and Mnih, 2008; Si and Jin, 2003; Yu et al., 2005a,b). When applying these models to collaborative filtering, each user is assumed to belong to one or many latent user groups that explain her preferences.

In Publication 4 we applied the collaborative filtering idea to British Parliament votings (Votings of the British Parliament in 1997–2001). Later there has been related work with same kind of senate voting data (Heller et al., 2008).

4.3.1 Connection to Co-Occurrence

In the Introduction there was discussion about two kinds of co-occurrence; in the first kind the observations are simply coupled by the sample identifier i, such as in vectors $\mathbf{x}_i$ and $\mathbf{y}_i$. If we denote users by u, documents by d and relevance by r, in the latter kind of co-occurrence we assume we have observed triplets (u_i, d_i, r_i), and view them as two tuples (u_i, r_i) and (d_i, r_i) that are paired by the sample identifier i.

In Publications 3, 4 and 5, the data sets consisted of the latter kind of co-occurrence, where (u_i, d_i, r_i) triplets were seen as co-occurring samples of pairs (u_i, r_i) and (d_i, r_i).

In Publication 2, the documents were represented as feature vectors $\mathbf{d}_i$ instead of scalars, but otherwise the co-occurrence was seen in the same manner, as two tuples (u_i, r_i) and $(\mathbf{d}_i, r_i)$ that are paired by the sample identifier i.

As mentioned in the Introduction, there was even co-occurrence on top of co-occurrence in Publication 1. First (u_i, d_i, r_i) triplets were seen as co-occurring samples of pairs (u_i, r_i) and (d_i, r_i) in a collaborative filtering model. In parallel, another model for co-occurrence data, based on eye movements was trained. In the eye movement model (u_i, d_i)-pairs were seen as individual samples, denoted here by ud_i for short, and the data therefore consisted of triplets $(ud_i, \mathbf{e}_i, r_i)$, where $\mathbf{e}_i$ stands for the feature vector from the eye movement path. This constitutes the latter kind of co-occurrence, when viewing it as paired samples $(ud_i, \mathbf{e}_i)$ and (ud_i, r_i). We predicted the relevance for each ud_i both with a model that was based on eye movements (resulting in prediction denoted by r_eye_i), and with a collaborative filtering model (denoted by r_cf_i). Eventually, the second layer of co-occurrence was exhibited when the predictions based on both models (ud_i, r_cf_i) and (ud_i, r_eye_i) were combined to produce the final relevance prediction (ud_i, r_i). This second layer can be considered to be co-occurrence of the first kind.

4.3.2 Combination of Eye Movements and Collaborative Filtering

In Publication 1, we complemented the rich but noisy eye-movement-based relevance feedback with collaborative filtering, using a probabilistic latent variable model. We proposed a model-based approach, which uses the User Rating Profile model (URP, Marlin (2004a)). The URP was optimized by using Markov Chain Monte Carlo (MCMC) integration (see Section 4.5.3) instead of the variational approximation used in Marlin (2004a,b). The two sources of relevance information were combined to one less noisy prediction using the Discriminative Dirichlet-Mixture Model, presented in Section 2.5. The combination of using implicit feedback from eye movements and relevance prediction from a collaborative filtering model is new. Collaborative filtering and content-based filtering have been combined earlier (e.g., Basilico and Hofmann (2004); Popescul et al. (2001)); so, combining all three would be a natural extension.

4.4 Related Issues

4.4.1 Cold-Start Problem

The problem of making preference predictions for unseen or barely seen users and documents is generally referred to as the *cold-start problem* in recommender system literature (see, for instance, Lam et al. (2008); Lashkari et al. (1994)). All systems that build profiles of their users have to rely on the past experiences of the users, and are therefore sensitive to the cold-start problem. In content-based approaches the problem concerns only new users, whose profile does not yet contain enough information. Once a new item is introduced it has its content description available for matching, so the cold-start problem does not have an effect on the item-side.

In contrast, a collaborative filtering system would have problems when assessing new documents that have not yet been seen by most of the users, because it only relies on the opinions of other users on the document. Making the collaborative filtering scheme item-based, that is, grouping items or documents instead of users, would in turn imply the problem where new users who have only few ratings will get poor predictions. In Publication 3, we proposed the Two-Way Model to tackle this problem of either new users or new documents, and the issue of new documents (or users) in collaborative filtering has been assessed in Publications 3, 4 and 5.

4.4.2 Sparsity and Missing Data

In general, user interest modeling deals with very sparse data. There can be thousands of movies or millions of other products of interest, from which a typical user has rated maybe some tens of items, leaving the rest of the ratings "unknown". Even when implicit feedback is gathered, e.g., based on selecting, clicking etc., there must be some actions on the user's behalf to indicate certain items to be more relevant than others, so the scale of sparsity cannot be tremendously altered by making the feedback mechanism implicit. From possibly millions of items, e.g., books, an individual user has typically given any kind of feedback to at most hundreds of items.

In practice, most models make the *missing at random* assumption for the missing data (Rubin, 1976). In statistical analysis it means that given the observed data, the mechanism of missingness does not depend on the unobserved data, i.e., $P(\, r \,|\, y_{obs} \,, y_{miss}) = P(\, r \,|\, y_{obs} \,)$ (for textbook reference, see Little and Rubin (2002)). The assumption is problematic since a rating could be missing because a certain user only rates items he or she liked, or because this particular item is unseen to her. In either case, the rating is not necessarily missing at random[20].

Considering both predictions based on user-similarity and on item-similarity, models can be made more robust to sparsity in rating data. Our approach is a generative two-way grouping model, while Wang et al. (2006) combines two memory-based collaborative filtering models to cope with the sparsity.

[20]In collaborative filtering context the missing at random assumption has been studied by Marlin et al. (2005, 2007), and it has been found that when users give ratings to music pieces the missing ratings have a different distribution than the ratings that were actually given.

4.4.3 Document Modeling

Document modeling is most often done for information retrieval or text categorization, but also content-based filtering in recommender systems can utilize various document models, like TF-IDF or topic models.

Documents are most generally represented with a bag-of-words model, where the frequencies of different words define the representation of the document and the order of the words is disregarded.[21] The frequencies of different words are weighted; the most popular weighting is the TF-IDF representation (see, Manning and Schütze (1999)), where each term is weighted according to its term-frequency and its inverse document frequency, giving the following weight for term t in document d:

$$tfidf_{td} = tf_{td} \times idf_t \quad . \tag{87}$$

Here tf_{td} is the count of term t in document d divided by its counts in the whole collection D, and $idf_t = \log \frac{|D|}{|\{d|t \in d\}|}$ models the inverse of the fraction of documents the term t occurs in within collection D. Although originally justified heuristically, the weighting has outperformed many improvements suggested later, and there might be information-theoretic connections to explain the good performance (Aizawa, 2003; Elkan, 2005)

Words tend to appear in documents in "bursts", that is, in such a way that if a word has already appeared in the document, it is more likely to appear again. This so-called *word burstiness* has been taken into account by Elkan (2005); Madsen et al. (2005) and also in the topic models by Doyle and Elkan (2009).

In Publication 2, we used a TF-IDF model with binary weights that only indicate the presence of the terms in the document, disregarding their relative frequencies.

4.5 Other Used Machine Learning Tools

This section briefly introduces the various machine learning tools used in the context of user interest modeling in this thesis.

4.5.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA, for a textbook reference see, Timm (2002)) searches for such a linear separation between two normally distributed classes that the expected Bayesian 0/1-loss is minimized. We denote by π_k the prior probability of class k, usually estimated by empirical frequencies in the training set. The class-conditional density of X in class k is multivariate Gaussian, with density

$$p(\mathbf{x} \mid k) = \left[(2\pi)^d \det \mathbf{\Sigma} \right]^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \quad , \tag{88}$$

where the covariance matrix $\mathbf{\Sigma}$ is assumed to be same for both classes. The class means $\boldsymbol{\mu}_k$ and the common covariance matrix $\mathbf{\Sigma}$ are estimated from the training set. So, the most probable class for observation $\mathbf{x}$ according to the posterior distribution is the MAP-estimate

$$\hat{k}(\mathbf{x}) = \arg\max_k \left[p(\mathbf{x} \mid k)\, \pi_k \right] \quad . \tag{89}$$

[21] Additionally the very common words, or *stop words*, are ignored. The same applies for too infrequent words, which only occur in one of the documents.

Finding Relevant Textual Features with LDA

It was shown in Publication 2 that content-based filtering of textual movie descriptions can be improved by learning their connection to genre-information and using the learned LDA model for new texts missing the movie genre. Finding such textual features would help in relevance prediction for future movies that would not have the same genre classification available. Using genre-information together with collaborative filtering to find pleasing movies for users has been suggested by Lee et al. (2007).

We simply sought one feature for each genre, to discriminate between movies belonging and not belonging to it. Therefore, we used LDA 10 times, each time trying to separate one class from all the other classes. We concluded that supervising feature selection by the genres improved performance of the subsequent prediction of relevance, giving almost the same performance as the original genre information.

4.5.2 Log-Linear Classifier

The log-linear classifier makes it possible to use a simple "linear" model with unconstrained vectorial input and still produce values in range $[0, 1]$. The input $\mathbf{x}_i$ denotes vectorial observations. The probability of an observation $\mathbf{x}_i$ to belong to a binary class ($r_i = 1$) is assumed to be Bernoulli-distributed with input-specific mean $\mu_i(\mathbf{x}_i)$, i.e.,

$$p(r_i \mid \mathbf{x}_i) = \mu(\mathbf{x}_i)_i^{r_i} \left[\, 1 - \mu(\mathbf{x}_i)_i \,\right]^{1-r_i}. \tag{90}$$

The logit function of the mean is assumed to obey a linear model with parameters $\mathbf{w}$:

$$\mathrm{logit}(\mu) := \log\left(\frac{\mu}{1-\mu}\right) = \mathrm{logit}(\,\mathbb{E}[r \mid \mathbf{x}]\,) = \mathbf{w}^T\mathbf{x}. \tag{91}$$

The parameters $\mathbf{w}$ can be sought by maximizing the likelihood of the observed data. For details of optimization see Nabney (1999). The inverse function of the logit is the logistic function

$$\mathrm{logit}^{-1}(t) = \frac{1}{1+e^{-t}}\ , \tag{92}$$

and the predicted class r_{new} of a new observation $\mathbf{x}_{new}$ is computed by taking the inverse $\mathrm{logit}^{-1}(\mathbf{w}^T\mathbf{x}_{new}) \in [0, 1]$ in this model.

Using Log-Linear Classifier for Relevance Predictions

In Publication 2 the log-linear classifier was used to model the relevances of each user. The input $\mathbf{x}_i$ denoted a vectorial representation for the document d_i, for instance a binary term vector. The probability of document d_i to be relevant ($r_i = 1$) to the user was assumed to be Bernoulli-distributed with document-specific mean and the logit of the mean was assumed to obey a linear model with user-specific parameters $\mathbf{w}$. The parameters $\mathbf{w}$ were sought by maximizing the likelihood of the observed data, i.e., ratings of the individual user. Predicted relevance of a new document $\mathbf{x}_{new}$ in this model was $\mathrm{logit}^{-1}(\mathbf{w}^T\mathbf{x}_{new}) \in [0, 1]$. In the experiments the predictions were rounded to binary predictions, $r \in \{0, 1\}$.

4.5.3 MCMC Sampling

Markov Chain Monte Carlo sampling is a class of commonly used techniques to evaluate Bayesian models. We used mainly Gibbs sampling to evaluate the latent topic models (in Publications 1, 3, 4 and 5). In Gibbs sampling the model parameters are sampled one at a time, conditional on all other parameters known (Casella and George, 1992; Geman and Geman, 1984). One iteration step consists of sampling all parameters once, in a fixed order.

We were interested in modeling the conditional distribution $\mathrm{P}(r \mid u,\, d)$ where the observations are triplets (u, d, r), meaning user, document, and the corresponding relevance. To model the conditional distribution, one can take the probability $\mathrm{P}(r \mid u, d)$ as the target function and estimate it with the mean over the M Gibbs iterations:

$$\begin{aligned} \mathrm{P}(r \mid u, d) &= \int_{\boldsymbol{\psi}} \mathrm{P}(r, \boldsymbol{\psi} \mid u, d)\, \mathrm{d}\boldsymbol{\psi} = \int_{\boldsymbol{\psi}} \mathrm{P}(r \mid u, d, \boldsymbol{\psi})\, \mathrm{P}(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\psi} \\ &= \mathbb{E}\left[\mathrm{P}(r \mid u, d, \boldsymbol{\psi})\right] \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{P}(r \mid u, d, \boldsymbol{\psi}^{(m)})\,. \end{aligned} \tag{93}$$

We always sampled three MCMC chains in parallel and monitored the convergence of predictions. First, each of the chains was run for 100 iterations of burn-in, with tempering like in Koivisto (2004) to aid the convergence. After that, the burn-in period was continued without the tempering, to burn in the actual posterior distribution. The Dirichlet priors $\boldsymbol{\alpha}_{u^*}$ and $\boldsymbol{\alpha}_{d^*}$ were sampled with the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953).

4.5.4 Product of Experts Model

In Publication 5 we introduced an approximation of the Two-Way Model with two generative URP models (see Sect. 2.4, Fig. 3); one that groups users and one that groups documents. In the approximation two Gibbs-sampled predictive Bernoulli distributions are estimated separately with user-based URP-GEN model and document-based URP-GEN model (Sect. 2.4), and their results are combined with a product of experts model (PoE, Hinton (2002)).

A Product of Experts model combines a number (K) of individual experts by taking their product and normalizing the results. Each expert is defined as a possibly unnormalized probabilistic model $q_k(\mathbf{x} \mid \boldsymbol{\varphi}_k)$ over its input space.

$$P(\mathbf{x} \mid \{\boldsymbol{\varphi}_k\}) = \frac{1}{Z} \prod_{k=1}^{K} q_k(\mathbf{x} \mid \boldsymbol{\varphi}_k) \quad , \tag{94}$$

where the normalizing coefficient is

$$Z = \int_{\mathbf{x}} \prod_{k=1}^{K} q_k(\mathbf{x} \mid \boldsymbol{\varphi}_k) d\mathbf{x}\,. \tag{95}$$

In our application, we first estimated the user's u relevance to document d with two different one-way models, the user-based URP model (denoted by $P_U(r = 1|u, d)$) and the document-based URP model (denoted by $P_D(r = 1|u, d)$). Finally, the product of these two estimates was taken, and the product distribution was normalized, as follows:

$$P_{PoE}(r = 1|u, d) = \frac{P_U(r = 1|u, d)\ P_D(r = 1|u, d)}{\sum_{r=0,1} P_U(r|u, d)\ P_D(r|u, d)} \quad . \tag{96}$$

The method has the advantage of giving better predictions than the individual one-way models with the computational complexity of the one-way model.

4.6 Summary

Modeling of user interest was the application area in Publications 1, 2, 3, 4 and 5. This chapter discusses the application point of view in these publications.

In Publication 2, we predicted the relevance of movie synopses using content-based filtering. Finding textual features that discriminate between the movie genres would help in relevance prediction for future movies that would not have the genre classification available. We concluded that supervising feature selection by the genres with linear discriminant analysis improved performance of the subsequent prediction of relevance.

To tackle the cold-start problem caused by new users and documents, in Publication 3 we proposed the Two-Way Model that groups both users and documents. In Publication 4 we showed that the two-way grouping is necessary when there can be both new users and new documents.

In Publication 5 we introduced a new efficient approximation of the Two-Way Model that achieves the prediction performance of the original Two-Way Model but with the computational complexity of a one-way grouping model.

5 Eye Movements

Research on eye movements has mainly been published in the field of psychology (Rayner, 1998). In psychological studies it is commonly assumed that in cognitively intensive tasks the attention is focused on where the eyes are fixated (eye-mind link assumption, Just and Carpenter (1976)), although the link does not always hold.

A recent overview by Salojärvi (2008) discusses applications of eye movements and introduces the background of the study in Publication 1 from a broader perspective. Here, we focus on the issues related to the joint Publication 1.

Eye-typing is one of the major applications of eye movements (Majaranta and Räihä, 2002, 2007). However, controlling a user interface solely by eye movements is laborious. Therefore, current research has also investigated the possibilities of complementing the traditional input methods with eye movements (Ajanki et al., 2009; Hardoon et al., 2007a; Hyrskykari et al., 2005; Porta, 2007; Salojärvi et al., 2003; Salojärvi, 2008; Vertegaal, 2002).

5.1 Physiological Background

The direction of the gaze contains a lot of information because the area of sharp vision is only 1–2 degrees wide (*fovea*). The eye can move either by drifting (e.g., when following a moving target) or by *saccades*, i.e., rapid jumps between more or less motionless *fixations*. When reading or browsing through written content the eye resorts to a scanpath of fixations and saccades. All the visual information is gathered during the fixations, which last approximately 200–300 ms (Ciuffreda and Tannen, 1995; Kienzle et al., 2009). The saccades, in turn, typically last 20–50 ms.

The current hypothesis is that eye movements are triggered by fairly low-level processes. Conscious control is certainly possible, but it gets burdensome over time, possibly because it requires active suppression of the standard low-level processes.

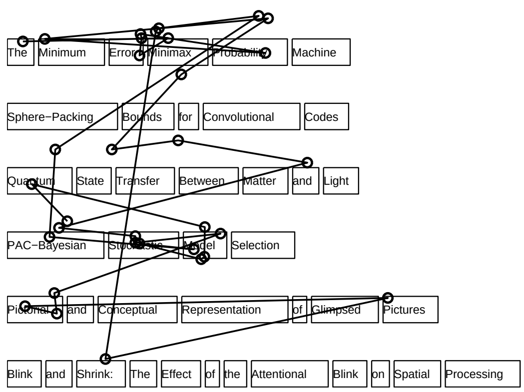


Figure 8: The fixations are marked as circles and the lines between them represent the saccades between them. The boxes surrounding the words were used to match the fixations to the word occurrences.

5.2 Measuring Eye Movements

During the last decades, eye movement measuring devices have become progressively cheaper and relatively accurate, allowing also free head movement of the user (Morimoto and Mimica, 2005).

In Publication 1 the eye movements were measured with a Tobii 1750 eye tracker with a screen resolution 1280×1024 pixels and a sampling rate of 50Hz. The equipment is shown in Figure 9, and it allows moderate head movements of the user.

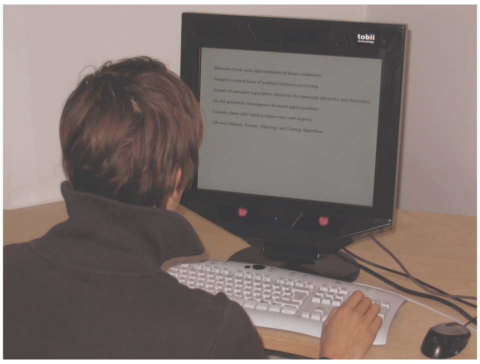


Figure 9: Tobii measurement device in use.

After some postprocessing, fixations can be identified and they are shown as circles in Figure 8, connected with solid lines, which denote the identified saccades.

There are many text-related eye movement features that have been suggested to be useful in separating those parts of text the reader considers relevant from those considered irrelevant (Calvo and Meseguer, 2002; Rayner, 1998; Salojärvi et al., 2005). Relying on an earlier feasibility study (Salojärvi et al., 2004), in Publication 1 we used the three text-related eye movement features listed in the frame below. All selected features were word-specific, resulting in data containing a feature vector for each word occurrence in the browsed text. The relevance of text lines was predicted based on the word-wise relevances using a so-called discriminative hidden Markov model (dHMM, Salojärvi et al. (2005)).

5.3 Eye Movements as Indicator of Relevance

Implicitly gathered feedback information can be used *proactively* in the background to improve the performance of the search in information retrieval (Tennenhouse, 2000). Salojärvi et al. have carried out feasibility studies on whether eye movements can be used as a source of implicit relevance feedback in information retrieval (Salojärvi et al., 2003, 2004; Salojärvi et al., 2005). For an review of the subject, see Salojärvi (2008). A possible application is to augment the traditional user interface by extracting implicit feedback from eye movements.

- One or many fixations within the word, modeled with a binomial distribution.
- Logarithm of the total fixation duration on the word, assumed to be Gaussian.
- Reading behavior (multinomial over 5 choices):
  - skip the next word
  - go back to already read words
  - read next word
  - jump to an unread line
  - the last fixation on the page.

After it had been established that relevance can be determined from eye movements to an extent, the information was exploited in Publication 1, where the relevance predictions from eye movements were combined with feedback information from collaborative filtering using Discriminant Dirichlet-Mixture Model (DDMM, Sect. 2.5). The study showed that a new type of proactive IR application is feasible, and introduced a justified way to combine relevance predictions from several information sources.

5.4 Hidden Markov Modeling Used in the Work

For the sake of completeness this section briefly introduces the methods used in modeling of the eye movement trajectories although it is not part of the contribution of this thesis.

5.4.1 Markov Chain

A Markov chain is a state model where there is a finite number of states and each state has its (multinomial) transition distribution. At each time step the chain moves from one state to another according to the transition probabilities P_{ij}. There can also be an output y_t at each step, and the output follows the output distribution q_j of the current state. Also, a starting distribution $\boldsymbol{\pi}_0$ or a starting state needs to be defined.

The Markovian property in its basic form states that the state transition probability only depends on the current state, not on how the state was reached, which can be stated as

$$P(s_{t+1} \mid s_1, \ldots, s_t) = P(s_{t+1} \mid s_t) \quad . \tag{97}$$

The states can have outputs attached to them, and the probability distribution of the current output only depends on the current state,

$$P(y_t) = P(y_t|s_t) \quad . \tag{98}$$

The transition probabilities between the states can be collected into a matrix with entries

$$P_{ij} = P(s_{t+1} = j|s_t = i) \quad . \tag{99}$$

Each row of the transition probability matrix $\mathbf{P}$ contains the multinomial transition probabilities from one state to all the other states. For example,

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.9 & 0.0 \\ 0.0 & 0.1 & 0.9 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} . \tag{100}$$

Figure 10 depicts a very simple Markov chain, with 3 states $\{s_1, s_2, s_3\}$.

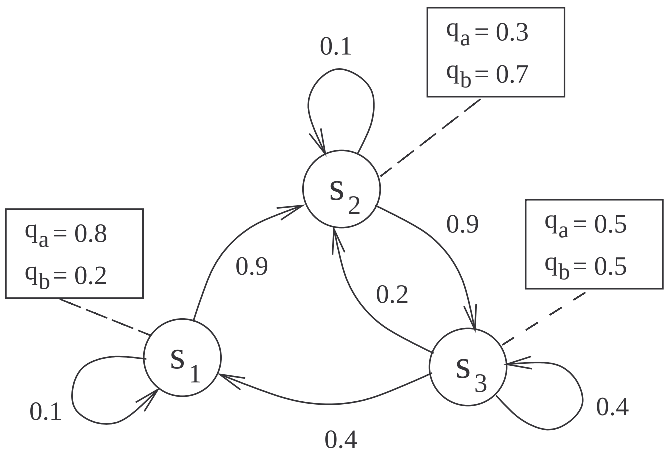


Figure 10: Illustration of a simple Markov chain. The output probabilities are shown in the boxes. (q_{jk}, where j is the state and k is the output: $k \in \{a, b\}$).

The output probabilities can also be collected into a matrix, where each row i of the output probability matrix $\mathbf{Q}$ contains the multinomial output probabilities over the outputs $\{a, b\}$ when arriving to state i. For example, in Figure 10 the corresponding output probability matrix $\mathbf{Q}$ is

$$\mathbf{Q} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix} . \tag{101}$$

5.4.2 Hidden Markov Models

Hidden Markov models are Markov chains where we assume the state sequence to consist of hidden variables. They are optimized by maximizing the likelihood of the observed output path with an EM-type algorithm (Baum-Welch or Forward-Backward algorithm; see, e.g., Bishop (2006)). It is guaranteed, as always with the EM-algorithm, that the algorithm does not decrease the likelihood,

$$P(\mathbf{y} \mid \hat{\boldsymbol{\varphi}}_{t+1}) \geq P(\mathbf{y} \mid \hat{\boldsymbol{\varphi}}_t)\ . \tag{102}$$

In Publication 1 we used a discriminative Hidden Markov Model introduced by Salojärvi et al. (2005), to predict the reader's subjective relevance for a set of documents, with the eye movements as the observations. The model is made discriminative by optimizing the conditional likelihood instead of the joint density likelihood, i.e.,

$$\mathcal{L}(C \mid \mathbf{Y}, \boldsymbol{\varphi}) = \prod_{i=1}^{N} P(c_i \mid \mathbf{y}_i, \boldsymbol{\varphi}) = \prod_{i=1}^{N} \frac{P(c_i, \mathbf{y}_i \mid \boldsymbol{\varphi})}{P(\mathbf{y}_i \mid \boldsymbol{\varphi})}\ . \tag{103}$$

5.5 Summary

In Publication 1 we developed a method for complementing the relevance predictions of collaborative filtering of text documents with implicit feedback from the eye movements. We were able to produce prediction results that outperform the predictions based on either single source of feedback. In this chapter our research on eye movements was discussed within the scope of Publication 1.

6 Brain Imaging with fMRI

By nature, brain functionality is spatially separated, that is, specific functions are often localized at specific areas in the brain. The distribution of brain activity can be identified at high spatial resolution using functional magnetic resonance imaging (fMRI). The task to be solved by modeling and data-analysis is to find the signal related to specialization of the brain functionality from the admixture of various kinds of signals.

6.1 Brief Introduction to Functional Magnetic Resonance Imaging

Basically, all magnetic resonance imaging is based on the interaction between the tissue under study, the applied magnetic fields and accurately synchronized radio frequency pulses.

Functional MRI of the brain is a non-invasive way to study brain function. The idea in fMRI is to record a sequence of images at successive time points in order to analyze the local changes in oxygenation level in different brain areas. The most widely used method is based on measuring so-called BOLD signal changes (Blood Oxygenation Level Dependent, Ogawa et al. (1992)). Detecting changes of neuronal activation in fMRI is based on the differing magnetic properties of oxygenated and deoxygenated hemoglobin molecules. Neuronal activation changes the blood flow and oxygenation levels locally, which can be measured.

The measurements can be transformed into an image of a focused slice of brain using standard signal processing techniques. The full volume image is produced by scanning several adjacent slices. Producing high resolution images can take several minutes, so such resolutions can be used only for structural imaging. Current fMRI scanners are able to produce full head volumes in a few seconds, but the spatial resolution is then only a fraction of that used in structural imaging.

The hemodynamic changes are related to the electrical activity of neurons in a complex and delayed way (see, e.g., Huettel et al. (2008)). Furthermore, the fast scanning and low signal to noise ratio of the BOLD signal make the image very noisy. In addition, the fMRI measurements are contaminated with artifacts, such as head movement and disturbances from the environment. Thus, the detection and analysis of interesting phenomena is very challenging. Therefore, the images need to be preprocessed in various ways (Worsley and Friston, 1995).

6.2 Traditional Neuroscientific Questions

Traditionally, fMRI studies use a strictly controlled stimulus, like visual patterns or audible beeps, designed to test a specific hypothesis. The standard way to do this is to repeat the stimulus a number of times, with empty periods without the stimulus in between, scan the fMRI sequence during the entire sequence, and average the scans to improve the signal-to-noise ratio. This kind of experimental setup is called *block-design*. If there are several different stimuli each "block" consists of either one of the stimuli or an empty period. In a block-design, there can be no overlap of the stimuli.

A standard procedure of analyzing an fMRI sequence is to use statistical parametric mapping (SPM2, 2002), which is based on a generalized linear model (GLM, Worsley and Friston (1995)). Current research focus is drifting towards more data-driven and adaptive methods, like Independent Component Analysis (ICA, see Section 6.5), which we also used in Publications 7, 8, and 9.

Some of the recent works in machine learning with fMRI data conceptualize the task as a prediction task: Are we able to predict the brain activity based on the past data and the given stimulus sequence? (Ghebreab et al., 2008; Kay et al., 2008; Mitchell et al., 2008; Rustandi et al., 2009) Another way to set the task has been: Are we able to predict the stimulus based on the past stimulus sequence and the given brain activity measurement sequence? (Ghebreab et al., 2007)

Both of these questions lead to more accurate modeling of the relationship between measured brain activity and the used stimuli, but there is still the flavor of regression making one of the data sets a covariate and the other the dependent variable (or response variable). If we do not trust either of the feature selections to be the absolutely relevant one, it is better to view the problem symmetrically, as in CCA; both of the data sets are seen as covariates for each other, and the dependence that is looked for is mutual rather than seeing one variable being dependent on the other.

6.3 Novel and Future Neuroscientific Questions

Natural stimuli are being increasingly used in fMRI studies to imitate real-life situations (Hari and Kujala, 2009; Malinen et al., 2007). They challenge the analysis methods used, making new kind of research questions possible. With natural stimuli it is no longer feasible to assume single features of the experimental design alone to account for the observed brain activity. Instead, relevant combinations of stimulus features could be behind the more complex activation patterns.

Canonical correlation analysis, and especially kernel CCA, have been used to meet the needs of more complex stimulations, e.g., in an unsupervised manner by Hardoon et al. (2007b,c), or in supervised or semi-supervised manner by Shelton and Bartels (2009).

Although we agree on the symmetricity of the setting by using CCA-type methods to find the relevant combinations on both sides (stimuli and fMRI-activity) we do not believe that the plain voxel data with, e.g., 10,000 voxels can be anything else than overfitted to the, e.g., 10–20 features of the stimuli with fully unsupervised methods.[22]

To reduce the complexity of the problem, it is necessary to reduce the dimensionality of the voxel data before applying an unsupervised dependency-seeking method, like CCA, to it. We have selected the dimensionality reduction method in such a way that even the results of the first step are meaningful functional elements of brain activity. In Publications 7, 8 and 9 we have suggested using ICA (Sect. 6.5) for the dimensionality reduction of the fMRI measurements and CCA-type of symmetric and unsupervised methods for finding the mutual dependencies between the stimulus features and the ICA-based functional patterns. This two-step framework will be discussed further in Section 6.6.

[22]In principle, of course, if all the *actually* involved stimuli were present in the data (the person's need to breath, blink his eyes, monitor his heart rate, thirst and hunger) there would be theoretical possibilities to model the brain activity on that level of completeness.

Other recent machine learning studies with fMRI data include work by Hutchinson et al. (2009), where a generative model is based on a set of assumed mental processes, for which parameters are learned from the data. The mental processes are, in contrast to what was assumed in our approach, given as prior knowledge. The model allows varying onset times relative to the stimulus sequence and overlapping of the mental processes. Also, NP-DeCA-type of an algorithm has been used directly on the voxel-specific fMRI measurements and the corresponding stimulus sequences by (Tsai et al., 1999).

6.4 Effects of Experimental Design

6.4.1 "Anticorrelations" within Experimental Settings

When the experimental design binds certain stimuli to always co-occur, it is not possible to distinguish the corresponding brain correlates from each other. This can be a natural consequence of true dependence between the stimuli, e.g., certain measurable auditory features that relate to female speech will always occur together with such speech. This ambiguity can also be a result of the selected level of detail in the experimental design, which does not allow us to distinguish between "naturally" correlated stimuli from those stimuli correlated only due to the experimental setup.

Somewhat unexpectedly, similar confusion may happen with stimuli that never co-occur. When the intuition has been to design uncorrelated labels by creating a sequence without any overlap, the regularity of the setup may have led to unexpected negative correlations between features. Such correlations that reflect the experimental design cannot be distinguished from true negative correlations between the stimuli. For details on how this happens, see Appendix 8. Thus one might find dependencies between stimuli and brain activity that merely represent characteristics of the experimental design, rather than of the observed brain responses. As an example, the experimental design of the data we analyzed in Publications 7 and 9 was unable to differentiate negatively correlated and uncorrelated stimulus blocks of different sensory modalities, since the occurrence of stimuli of any sensory modality was fully determined by the absence of stimuli of all the other senses. More generally, the effects of negative and other spurious correlations seem to be an emerging topic of scientific interest (see, e.g., Aguirre et al., 1998; Gretton et al., 2006; Murphy et al., 2008).

Under strict laboratory control, it is possible to some extent to design the experiment so that all relevant combinations of experimental variables are presented in a well-balanced setup. However, in natural settings we cannot rely on controlled designs but instead the data analysis has to take care of the balancing. Hence, adequate design of the experimental setups to include rich enough stimuli is important, to allow the analysis to deal with the balancing.

6.4.2 Why Rich Set of Features is Needed

The main justification for the need for expressive stimulus features is that if the actual reason for the measured brain activity cannot be represented with the stimulus features, the brain correlates can be misinterpreted in terms of the available features. However, if the set of stimulus features is expressive enough, the proposed two-step framework could find a new kind of combination of the stimuli that would give a hint of the missing feature. In principle, the CCA analysis, being invariant to linear transformations of the stimulus features, is able to compensate imperfect choice and encoding of stimulus features.

6.5 Independent Component Analysis

Independent Component Analysis (ICA) is a method for finding statistically independent components from multivariate statistical data. ICA is commonly used for separating statistically independent sources from fMRI measurements. In general, ICA is one of the most popular methods for solving the so-called *blind source separation* problem (BSS)

$$\underset{N\times T}{\mathbf{X}} = \underset{N\times K}{\mathbf{A}}\ \underset{K\times T}{\mathbf{S}} \quad , \tag{104}$$

where only the observed data $\mathbf{X}$ are known, and where N is the number of samples, K is the number of independent sources and T is the number of time points. This interpretation is valid for the so-called *temporal ICA*, where the sources are assumed to be time-dependent signals $\mathbf{s}(t)$. With fMRI measurements, however, *spatial ICA* is typically used, where the sources are assumed to be spatial instead of temporal, as follows:

$$\underset{N\times V}{\mathbf{X}} = \underset{N\times K}{\mathbf{A}}\ \underset{K\times V}{\mathbf{S}} \quad , \tag{105}$$

where V is the number of spatial features (typically voxels in fMRI measurements), and N is the number of samples (typically time points in fMRI measurements). ICA assumes only statistical independence of the sources $\mathbf{s}(v)$ and full rank of the mixing matrix $\mathbf{A}$. Many different ICA-variants have been developed since the first ICA models (see (Hyvärinen et al., 2001) for both the history of the method and for later developments).

An efficient algorithm for solving ICA is the so-called *FastICA* algorithm (Hyvärinen and Oja, 1997; FastICA, 1998), which is based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity measured by negentropy.[23]

The inherent stochasticity of the ICA algorithm leads to variability between ICA runs. The *reliable ICA* algorithm (Ylipaavalniemi and Vigário, 2004, 2008; Ylipaavalniemi and Soppela, 2009; Arabica, 2008) takes into account the algorithmic variability and the uncertainty of the sources in the data by performing multiple runs of ICA. Components that are consistent across several runs are considered reliable. In Publications 7, 8 and 9 we analyzed the fMRI measurements with the reliable ICA based on multiple runs of FastICA.

[23]Negentropy is such a measure that it is always non-negative, and zero if and only if the measured data is Gaussian. It is a desirable measure also because of its invariance to invertible linear transformations.

In fMRI analysis, independence is typically considered in the spatial domain (McKeown et al., 1998), where the corresponding mixing vectors reveal the temporal dynamics of each identified independent functional pattern. A broad overview of ICA usage in the field of brain imaging has been presented by Vigário and Oja (2008). Fig. 11 shows an example of such an independent component and its time course.

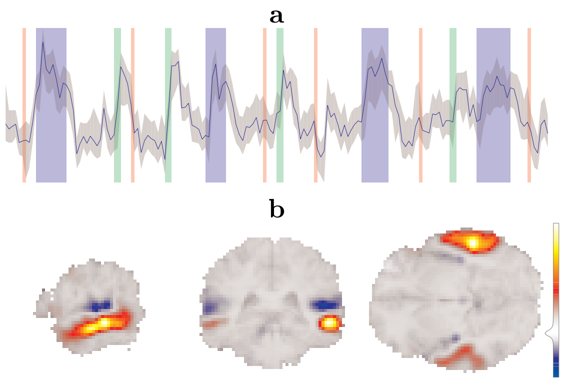


Figure 11: An illustrative example of one independent component (IC). Panel **a** shows the mean time course of the IC, averaged across the two trials of all test subjects. The mean time courses of the IC's were used as input for CCA, together with the stimulation time courses. The gray band around the trace shows the 95% confidence interval given by the reliable ICA approach. Three auditory stimulus features are shown as colored blocks behind the time course. The changes in the IC do not exactly match any of the individual features although the activity of the IC is correlated with them. Panel **b** illustrates three different slices of the average brain (sagittal, coronal and axial), always centered on the most active voxel of the IC. The bar on the right shows the used color range based on the z-score; the upper end of the scale depicts positive weights and the lower end negative weights. The left edge of the bar shows the shape of the distribution of the weight values. The more non-Gaussian the shape is, the more independent the IC is.

6.6 Symmetric Two-Step Framework with ICA Followed by DeCA

In Publications 7 and 8 we have proposed a two-step approach, where independent component analysis (ICA, Bell and Sejnowski (1995); Hyvärinen et al. (2001)) is first used to identify spatially independent brain processes, which we refer to as *functional patterns*. As the second step, temporal dependencies between stimuli and functional patterns are detected using either CCA (Publication 7) or its distribution-free variant NP-DeCA (Klami and Kaski (2005), applied in Publication 8). The latest development was to introduce a faster variant of the nonparametric NP-DeCA, based on Mixture of Gaussians density estimation (Publication 9). This Semi-Parametric or SP-DeCA is described in Section 3.6.3 in more detail.

Our two-step approach, thus, looks for combinations of stimulus features and the corresponding combinations of functional patterns. Based on the findings in Publications 7 and 8 this approach seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available. Figure 12 illustrates the framework.

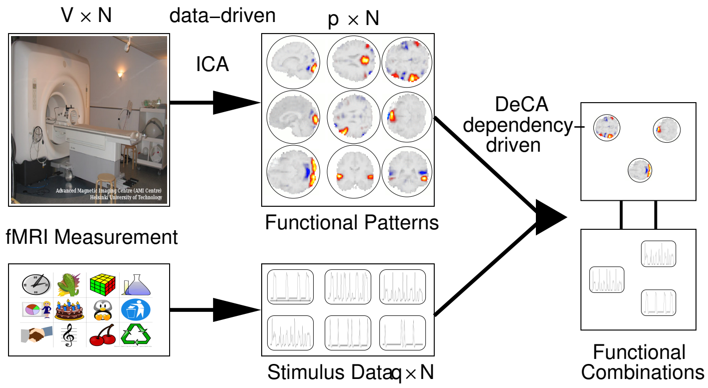


Figure 12: Sketch of the framework. Step 1: ICA is applied to the fMRI measurements ($V =$ # of voxels, $N =$ # of measurement time points, to find spatially independent patterns of brain activity $p =$ # of reliable ICA components, Step 2: DeCA is applied to identify functional combinations based on the temporal dynamics of both the stimuli and the ICA components $q =$ # of stimulus time courses).

The aim of Publications 7, 8 and 9 is to take the first step towards inferring brain correlates of natural stimuli with possibly overlapping stimuli. In this framework the statistical hypotheses are no longer self-evidently derived from the experimental setup, as conventionally. Instead, it is a goal of the analysis in itself to identify the correct hypotheses by data-driven methods. We used the two-step approach for the brain measurement analysis: First, spatially independent patters of brain activity were extracted from magnetic resonance imaging (fMRI) data with ICA. As the second step, in Publication 7, we used classical CCA (Sect. 3.4) to find dependencies between the stimuli and fMRI-based ICA-components from an existing study (Malinen et al., 2007). CCA is a fast and robust method, but for many applications correlation is too simple a measure; CCA implicitly makes an assumption about normally distributed data sets.

Replacing correlation with mutual information makes discovery of more general types of dependency possible. As discussed in Chapter 3, mutual information cannot be computed as easily as correlation, and we need to resort to approximations. In NP-DeCA, nonparametric Parzen-kernel estimates are used to model the densities in the projection space, and the mutual information between the Parzen-estimates is maximized (see Sect. 3.6.2 for detailed description of the algorithm (Klami and Kaski, 2005)). NP-DeCA was used as the method for finding mutual dependencies between the stimuli and the fMRI-based ICA-components in Publication 8.

In Publication 9, we introduced a faster semi-parametric DeCA-variant (SP-DeCA) and applied it to the task of finding dependencies between measured brain activity and multi-sensory stimuli. Semi-Parametric DeCA is discussed in Section 3.6.3. Following the earlier application of CCA to the same study in Publication 7, we used the two-step framework: First, functional patterns were extracted from fMRI data with reliable ICA.Then, the SP-DeCA algorithm was used to find mutual dependencies between the brain patterns and the stimuli.

6.7 Other Used Machine Learning Tools

This section briefly introduces the various machine learning tools used in the context of brain imaging in Publications 7, 8 and 9.

6.7.1 Mixture of Gaussians Model

The mixture of Gaussians model is a traditional method for modeling multimodal distributions with more than one "bump" or mode (McLachlan and Basford, 1988; McLachlan and Peel, 2000). The density model is simply a linear combination (or mixture) of K different Gaussian distributions,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \;, \tag{106}$$

where $\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$. The π_k's represent the prior probabilities of the mixture components (called the *mixture coefficients*), and are therefore non-negative values that sum up to unity. This estimate has a set of parameters $\boldsymbol{\varphi} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ which need to be learned. One way to determine the values of these parameters is to use the maximum (log-)likelihood

$$\ln \hat{p}(\mathbf{X} \mid \boldsymbol{\varphi}) = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \;, \tag{107}$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$. It can be maximized either by iterative numerical optimization techniques or by the expectation maximization (EM-)algorithm discussed in Section 2.1.2.

6.7.2 K-Means Clustering

K-means clustering is a popular clustering algorithm that finds compact clusters (Ball and Hall (1967), for textbook description see, e.g., Bishop (2006)). The problem is to find such a partition of given data points into K partitions that the within-cluster sum of square errors (108) is minimized

$$J = \sum_{i=1}^{N} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad . \tag{108}$$

Here, we denote the observations by $\{\mathbf{x}_i \mid i = 1 \dots N\}$, the clusters by $\{C_k \mid k = 1, \dots, K\}$ and the centroids (in this case means) of the clusters by $\{\boldsymbol{\mu}_k \mid k = 1, \dots, K\}$.

In general, the problem is NP-hard, so different heuristic algorithms have been used to solve it. The most common algorithm is presented in the frame below (K-means algorithm or Lloyd's algorithm). The number of clusters K and some initial values for the cluster centroids $\boldsymbol{\mu}_k$ need to be set before applying the algorithm.

1. *All the training set data points are assigned to the cluster whose centroid is the closest to them,*

$$C(\mathbf{x}_k) = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad . \tag{109}$$

2. *The centroids $\boldsymbol{\mu}_k$ are recalculated by*

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \quad , \tag{110}$$

i.e., they are set to the means of the data points in the cluster.

These two steps are repeated until there is no change between the iterations.

K-Means Clustering as Initialization of Semi-Parametric DeCA

We used K-means clustering to initialize the means of the Gaussians in the Semi-Parametric DeCA in Publication 9. We initialized the mixture estimate in the projection space as follows. The initial values of means $\boldsymbol{\mu}_k$ and mixture probabilities π_k were determined by running K-means clustering ($K = 5$) separately in both projection spaces, using the cluster centroids as initial means of the mixture components and the relative cluster sizes as the mixture weights. Finally, the covariance matrix $\boldsymbol{\Sigma}_k$ of each mixture component was initially set to the diagonal matrix $\begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$ containing the variances of the initial projections s_x and s_y.

6.8 Summary

Natural stimuli are being increasingly used in fMRI studies to imitate real-life situations. With more natural stimulation, questions about negative correlations and expressiveness of stimulus features become more important, together with the new kind of experimental settings.

Instead of assuming single features of the experimental design to account for the brain activity, in Publications 7, 8 and 9 we have suggested relevant combinations of stimulus features could be behind the more complex activation patterns. We have proposed a novel two-step framework where ICA is first used to identify spatially independent functional patterns. As the second step, DeCA is used for finding the temporal mutual dependencies between the stimulus features and the ICA-based functional patterns. Based on our findings this two-step framework seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available.

7 Modeling Genes and Their Regulation

In Publication 6, we studied the regulation of genes in baker's yeast (*Saccharomyces cerevisiae*) where understanding of the gene regulation of other organisms is an ultimate goal. This chapter gives an introduction to the application area of modeling genes and their regulation to the extent that is related to this thesis. The biological background of this chapter is mostly based on Nikkilä (2005) and Campbell and Reece (2001).

7.1 Basic Structure of Cells

Cells have a basic structure that is shared across the organisms, thus many inferences made on the cell level apply to a wide range of organisms. This is commonly utilized in biology and medicine, in particular when using so-called *model organisms*, such as yeast or mouse, to study human.

Organisms are divided into two main classes based on whether the cells have a membrane around their nucleus; *eukaryotes* have the membrane (e.g., human) and *prokaryotes* (e.g., bacteria) do not have it. In a eukaryotic cell the genes are encoded in the nucleus as a double-stranded DNA-molecule (Deoxyribonucleic acid)[24]. The baker's yeast *Saccharomyces cerevisiae* is one of the the best-understood eukaryotic organisms, and the focus of our study in Publication 6.

7.2 Gene Expression

A gene is a sequence of the DNA in the nucleus that contains the information the cell needs to manufacture a protein. Each DNA chain is composed of four kinds of chemical building blocks called *nucleotides*. Proteins, in their part, are involved in practically every process in the cell. A cell can get signals from its environment or it can react to its internal state (e.g., concentrations of different proteins).

When a cell starts to manufacture a certain protein it first makes a copy of the corresponding gene's DNA, coded in *messenger-RNA* or *mRNA*.[25] The process where genes are coded into mRNA for protein production is called *transcription*. The messanger-RNA is then transported outside the nucleus and after some preprocessing it is *translated* to a protein according to the instructions coded by the sequence of nucleotides.

Gene expression is measurable activity of a gene at some specific time, measured by the amount of the gene's mRNA transported outside the nucleus. With current technology, using DNA microarrays the gene expression of thousands of genes can be measured at the same time (Lockhart et al., 1996; Schena et al., 1995).

[24] DNA is the nucleic acid that carries the genetic information in the cell. The 4 different nucleotides stem from the bases adenine, thymine, cytosine and guanine. The sequence of nucleotides determines individual hereditary characteristics.

[25] Ribonucleic acid is a constituent of all living cells, consisting of a long, single-stranded chain of alternating phosphate and ribose units with the bases adenine, guanine, cytosine, and uracil bonded to the ribose. Messenger-RNA molecules are involved in the transmission of genetic information.

7.3 Gene Regulation by Transcription Factors

The most dominant regulatory mechanism for controlling the protein concentrations in the cell is *transcriptional regulation*, which means regulating which genes are transcribed to proteins and how much. A set of proteins called *transcription factors* (TF) binds to a certain DNA sequence nearby a gene (the gene's promoter region).[26] Whether the gene is going to be transcribed to a protein or not depends on the configuration of TF's on its promoter region.

Binding of transcriptional regulators can be measured genome-wide to reveal regulatory networks. The measurements are, however, noisy and expensive. We have applied the Two-Way Model (Sect. 2.3.1) to model existing binding data in order to predict binding for new factors or genes, assuming groups of genes and groups of transcription factors have similar binding patterns (Kaski et al., 2005c).

Microarray-based *chromatin immunoprecipitation* (ChIP) allows measuring the binding strength of the transcription factor (TF) proteins on any gene's promoter region (Lee et al., 2002; Ren et al., 2000). This kind of measurements were used as one of the data sets in the dependency modeling of Publication 6.

7.4 Studying Stress Response of Yeast Cells

The response of yeast cells to stress induced by drastic changes in the environment has been used as a paradigm to study gene regulation networks. Understanding yeast gene regulation will, among other things, help as a model for studies on higher organisms.

When a yeast cell is challenged by a rapid change in the surrounding conditions (e.g., temperature, osmolarity, pH or nutrient), it starts a genome stress response program. Survival of especially single-cell organisms depends on their ability to adapt to the environmental changes, and therefore stress response has received much attention. In the baker's yeast *Saccharomyces cerevisiae* several hundred genes out of about 6500 present in the genome have been found involved in a stereotyped stress response pattern (Causton et al., 2001; Gasch et al., 2000; Kaski et al., 2005a; Mager and Kruijiff, 1995; Ruis and Schuller, 1995). It has become evident that a certain group of yeast genes (so called common *environmental stress response (ESR) genes*) is always activated during various stress treatments.

In Publication 6, we searched for genes with maximal dependency between gene expression data and ChIP-data to improve the accuracy of inference of which transcription factors actually regulate each gene. The dependency maximization was shown to improve the results compared to using either data source alone.

We modeled the yeast stress reaction by extracting the shared variation between a set of stress treatments, considering all other variation irrelevant. In addition, the regulation of stress was explored by searching for maximal dependencies between the extracted stress reaction and a transcription binding data from Lee et al. (2002).

[26]Other regulatory mechanisms include alternative splicing (Campbell and Reece, 2001) and RNA interference (Dykxhoorn et al., 2003).

The gene expression data used in this analysis was combined from Causton et al. (2001) and Gasch et al. (2000) and it formed expression data in altogether 16 stress treatments for 5998 genes. The full expression data set was 104-dimensional. TF-binding data for the same genes and 113 transcription factors was obtained from Lee et al. (2002). In summary, we had 16 gene expression data sets with variable numbers of columns, paired by the common 5998 genes, and additionally, we had one TF binding data set with the binding strengths of 113 transcription factors for each gene.

Generalized CCA (see Section 3.4.4) was used to extract only the variation that was common to all 16 treatments in the gene expression data. We used cross-validation to estimate the reliability of the gCCA components, and ended up with 12 generalized canonical components. Of the 12 components 9 showed statistically significant association to the ESR genes that were previously known to be either up-regulated or down-regulated in stress.

Finally, we used Associative Clustering (see Section 3.6.1) to search for genes with maximal dependency between the gCCA-projected gene expression data (5998 genes $\times$ 12 gCCA components) and the TF-binding data (5998 genes $\times$ 113 TF's) to improve the accuracy of inference of which transcription factors actually regulate each gene.

7.5 Summary

In Publication 6 we studied a case from the field of bioinformatics, which matches especially well to the dependency modeling task. The task was to find which part of the structure in gene expression data about yeast stress response is common to all the expression data sets. The alternative interpretation of CCA (derived in Sect. 3.4.4) led us to a new way of using gCCA in feature selection, by choosing those features that are maximally dependent between the datasets.

Furthermore, by searching for the dependencies between these features selected using gCCA and another data set about binding of the gene regulators, we gained an interpretation of how the yeast genes are regulated in stress.

8 Conclusions

This thesis outlined how to find what is relevant in co-occurrence data. Two types of relevance were considered. The first was the relevance of items as seen by a user subjectively, like in the case of information retrieval. In the other view of relevance, the problem of finding what is relevant in data was formalized via dependence, that is, the variation that is found in both (or all) co-occurring data sets was deemed to be more relevant than variation that is present in only one (or some) of the data sets. Frameworks for different application areas were suggested using both existing methods and methods developed in this thesis. The dependency-seeking models were extended to nonparametric models, and computational algorithms were developed for the models.

Method Development. The method development contributions of this thesis are related to latent topic models and dependency exploration. The methods are applicable to mutual dependency modeling and co-occurrence data in general, without restriction to the applications presented in the publications of this work.

Traditionally, latent topic models are one-way clustering models, that is, one of the variables is clustered by the latent variable. Motivated by the application of collaborative filtering we proposed a generative latent topic model that generalizes in two ways, the Two-Way Model (Publication 3). We have shown that when only a small amount of data has been gathered, two-way generalization becomes necessary (Publication 4). Furthermore, we introduced a new efficient approximation of the Two-Way model that achieves the prediction performance of the original Two-Way Model but with the computational complexity of the one-way grouping model (Publication 5).

In this thesis an alternative view to CCA was derived, as whitening of the within-dataset-covariances followed by PCA to search for those features that are maximally dependent between both (or all the) datasets. This interpretation led us to a new way of using CCA in feature selection (Publication 6).

As classical CCA has the restriction of implicitly assuming the data to be normally distributed, we wanted to find an unsupervised dependency-seeking method without this constraint. We applied a distribution-free extension of CCA to searching dependency structure between two co-occurring data sets (Publication 8). However, the Parzen-estimate-based method was not efficient enough for large data sets, and therefore, we pursued to find a faster algorithm for the task. We introduced a new faster variant of the method, Semi-Parametric DeCA where the Parzen-estimates were replaced by density estimation based on Mixture of Gaussians (Publication 9).

Applications. The application areas of the publications include modeling of user interest (Chapter 4), relevance prediction based on eye movements (Chapter 5), analysis of brain imaging with fMRI (Chapter 6) and modeling of gene regulation in bioinformatics (Chapter 7). The main contribution to each of these four application areas is described below.

Modeling of User Interest. In a feasibility study, we applied content-based filtering to the prediction of the users' subjective relevance values for movies. We concluded that supervising feature selection by the genres improved performance of the subsequent prediction of relevance (Publication 2).

Modeling of user interests involves learning user interest profiles, and all recommender systems have to rely on the past experiences of the users. Therefore, such systems often have problems with new users (or new items of interest).

To tackle the cold-start problem caused by new users and documents in collaborative filtering, we proposed the Two-Way Model that groups both users and documents (Publication 3). Additionally, we introduced a new efficient approximation of the Two-Way Model by combining the predictions of two one-way grouping models. The approximation achieves the prediction performance of the original Two-Way Model but with the computational complexity of a one-way grouping model (Publication 5).

Eye Movements. The direction of the gaze can be a useful source of information in many kinds of user interfaces. In information retrieval, implicit feedback information can be used proactively in the background to improve the performance of the search. We have developed methods for enhancing the relevance predictions of collaborative filtering of text documents with implicit feedback from the eye movements. We were able to produce prediction results that outperform the predictions based on either single source of feedback (Publication 1). This work has been continued by proactive information retrieval with the aid of eye movements, and an interesting future direction would be combining eye movements with brain imaging in order to better understand to which parts of the stimuli the brain signal responses to.

Brain Imaging. Functionality of the brain is, by nature, spatially separated. This specialization of the brain sites to different tasks can be identified at high spatial resolution using functional magnetic resonance imaging. Natural stimuli are being increasingly used in fMRI studies to imitate real-life situations. Instead of assuming single features of the experimental design to account for the brain activity, we suggested that relevant combinations of stimulus features could be behind the more complex activation patterns.

We have proposed a novel two-step framework where ICA is first used to identify spatially independent patterns of brain activity. As the second step, Dependent Component Analysis is used for finding the temporal mutual dependencies between the stimulus features and the ICA-based functional patterns (Publications 7, 8 and 9). Furthermore, we introduced a new faster variant of DeCA, Semi-Parametric DeCA (Publication 9).

Based on our findings this two-step framework seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available. Future work in brain imaging will involve the ability to use more natural stimuli and eventually, use truly natural stimuli to study humans in their normal social environment.

Gene Regulation. In order to understand how the network of genes operates in an organism, the influence of the regulatory proteins needs to be investigated. They regulate the activity of genes by binding to certain areas near the genes. We used the baker's yeast as a model organism but the ultimate goal would be to understand the corresponding gene regulatory networks in human.

With current microarray technology both the gene expression and the binding strength of the regulatory proteins can be measured for thousands of genes at the same time. The task was to find which part of the structure in gene expression data about yeast stress response is common to all of the expression data sets. Using generalized CCA to multiple stress-related gene expression data sets we produced one representation of the shared variation. Our alternative view to CCA led us to a new way of using CCA in feature selection, by choosing those features that are maximally dependent between the datasets (Publication 6).

Furthermore, by searching for the dependencies between these features selected using CCA and another data set about binding of the gene regulators, we gained an interpretation of how the yeast genes are regulated under stress.

Appendix 1

Details of Deflation in PCA and CCA

Deflation in PCA

In PCA we normalize the projection vectors $\mathbf{w}_k^T\,\mathbf{w}_k = 1$ and we wish the deflation to take care that the projections are orthogonal $\mathbf{w}_i^T\mathbf{X} \perp \mathbf{w}_k^T\mathbf{X}$. The deflation considered in Section 3.3

$$\bar{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_1}\mathbf{X}^T\,\mathbf{w}_1\,\mathbf{w}_1^T\,\mathbf{X}\right) \tag{111}$$

has this property:

$$\begin{aligned}
\mathbf{w}_1^T\,\mathbf{X}\,\bar{\mathbf{X}}^T\,\mathbf{w}_k &= \mathbf{w}_1^T\,\mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_1}\mathbf{X}^T\,\mathbf{w}_1\,\mathbf{w}_1^T\,\mathbf{X}\right)\mathbf{X}^T\,\mathbf{w}_k \\
&= \mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_k - \frac{1}{\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_1}\left(\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_1\right)\mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_k \\
&= \mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_k - \mathbf{w}_1^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{w}_k = 0\ ,
\end{aligned} \tag{112}$$

regardless of how the subsequent projection vector $\mathbf{w}_k$ is chosen.[27]

When we have found the first principal component $\mathbf{w}_1$ in PCA, we can also use the fact that it is an eigenvector of the original covariance matrix $\mathbf{C}$ (or actually of its estimate $\mathbf{C} = \frac{1}{N}\,\mathbf{X}\,\mathbf{X}^T$),

$$\mathbf{C}\,\mathbf{w}_1 = \lambda_1\mathbf{w}_1\ . \tag{113}$$

Then, we can see that the same deflation formula can be written in another equivalent form:

$$\begin{aligned}
\bar{\mathbf{X}} &= \mathbf{X}\left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T\,\underbrace{\mathbf{X}\,\mathbf{X}^T}_{N\,\mathbf{C}}\,\mathbf{w}_1}\mathbf{X}^T\,\mathbf{w}_1\,\mathbf{w}_1^T\,\mathbf{X}\right) \\
&= \mathbf{X} - \frac{1}{N\mathbf{w}_1^T\,\underbrace{\mathbf{C}\,\mathbf{w}_1}_{\lambda_1\,\mathbf{w}_1}}\,N\,\underbrace{\mathbf{C}\,\mathbf{w}_1}_{\lambda_1\mathbf{w}_1}\,\mathbf{w}_1^T\,\mathbf{X} = \mathbf{X} - \frac{1}{\lambda_1\,\underbrace{\mathbf{w}_1^T\,\mathbf{w}_1}_{=1}}\,\lambda_1\,\mathbf{w}_1\,\mathbf{w}_1^T\,\mathbf{X} \\
&= \left(\mathbf{I} - \mathbf{w}_1\,\mathbf{w}_1^T\right)\mathbf{X}\ ,
\end{aligned} \tag{114}$$

which actually takes care of the orthogonality of the projection vectors, $\mathbf{w}_i \perp \mathbf{w}_k$.

[27] However, if $\mathbf{w}_1$ were chosen again, the product would not be zero because of orthogonality, but because $\mathbf{w}_1^T\,\bar{\mathbf{X}}$ is a zero vector.

The equivalence of these two orthogonalities can be seen as follows

$$
\begin{aligned}
0 &= (\mathbf{w}_1^T\,\mathbf{X})(\mathbf{w}_k^T\,\mathbf{X})^T = \mathbf{w}_1^T\,\underbrace{\mathbf{X}\,\mathbf{X}^T}_{N\,\mathbf{C}}\,\mathbf{w}_k \\
&= N\,\underbrace{\mathbf{w}_1^T\,\mathbf{C}}_{\lambda_1\,\mathbf{w}_1^T}\,\mathbf{w}_k = N\,\lambda_1\,\mathbf{w}_1^T\,\mathbf{w}_k \\
&\Leftrightarrow \mathbf{w}_1^T\,\mathbf{w}_k = 0
\end{aligned}
\tag{115}
$$

Generally, the variance in the direction of a vector $\mathbf{u}$ *before* the deflation is $\frac{1}{N}\,\mathbf{u}^T\,\mathbf{X}\,\mathbf{X}^T\,\mathbf{u}$ and *after* the PCA deflation the variance in the direction $\mathbf{u}$ is

$$
\begin{aligned}
\frac{1}{N}\,\mathbf{u}^T\,\bar{\mathbf{X}}\,\bar{\mathbf{X}}^T\,\mathbf{u} &= \frac{1}{N}\,\mathbf{u}^T\,\left(\mathbf{I}-\mathbf{w}_1\,\mathbf{w}_1^T\right)\,\mathbf{X}\,\mathbf{X}^T\,\left(\mathbf{I}-\mathbf{w}_1\,\mathbf{w}_1^T\right)\,\mathbf{u} \\
&= \mathbf{u}^T\,\left(\mathbf{I}-\mathbf{w}_1\,\mathbf{w}_1^T\right)\,\mathbf{C}\left(\mathbf{I}-\mathbf{w}_1\,\mathbf{w}_1^T\right)\,\mathbf{u} \\
&= \mathbf{u}^T\,\left(\mathbf{C}-\underbrace{\mathbf{C}\,\mathbf{w}_1}_{\lambda_1\,\mathbf{w}_1}\,\mathbf{w}_1^T-\mathbf{w}_1\,\underbrace{\mathbf{w}_1^T\,\mathbf{C}}_{\lambda_1\,\mathbf{w}_1^T}+\mathbf{w}_1\,\mathbf{w}_1^T\,\underbrace{\mathbf{C}\,\mathbf{w}_1}_{\lambda_1\,\mathbf{w}_1}\,\mathbf{w}_1^T\right)\,\mathbf{u} \\
&= \mathbf{u}^T\,\mathbf{C}\,\mathbf{u}-\lambda_1\,(\mathbf{u}^T\,\mathbf{w}_1)^2
\end{aligned}
\tag{116}
$$

In the direction of $\mathbf{w}_1$ the variance is therefore

$$
\mathbf{w}_1^T\,\underbrace{\mathbf{C}\,\mathbf{w}_1}_{\lambda_1\,\mathbf{w}_1}-\lambda_1\,(\mathbf{w}_1^T\,\mathbf{w}_1)^2 = \lambda_1\,\underbrace{\mathbf{w}_1^T\,\mathbf{w}_1}_{=1}-\lambda_1 = 0\,, \tag{117}
$$

as wanted, and in the direction of another principal component $\mathbf{w}_k$ the variance is

$$
\mathbf{w}_k^T\,\underbrace{\mathbf{C}\,\mathbf{w}_k}_{\lambda_k\,\mathbf{w}_k}-\lambda_1\,(\mathbf{w}_k^T\,\mathbf{w}_1)^2 = \lambda_k\,\underbrace{\mathbf{w}_k^T\,\mathbf{w}_k}_{=1}-0 = \lambda_k\;. \tag{118}
$$

Deflation in CCA

In the case of CCA, the normalization constraints are $\mathbf{w}_k^T \mathbf{C} \mathbf{w}_k = 1$ and the orthogonality is required between the projections, $\mathbf{w}_i^T \mathbf{X} \perp \mathbf{w}_k^T \mathbf{X}$. The deflation of Eq. (111) still holds, but it is no longer equivalent with the deflation of Eq. (114).

By straightforward substitution, it still holds that

$$\mathbf{w}_1^T \mathbf{X} \bar{\mathbf{X}}^T \mathbf{w}_k = 0 \,, \tag{119}$$

regardless of how the subsequent projection vector $\mathbf{w}_k$ is chosen.

Generally, the variance in the direction of a vector $\mathbf{u}$ *before* the deflation is $\frac{1}{N} \mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u}$, which means that for the CCA components the variance before deflation is

$$\mathbf{w}_k^T \mathbf{C} \mathbf{w}_k = 1 \,. \tag{120}$$

After the deflation the variance in the direction $\mathbf{u}$ is

$$\begin{aligned}
&\frac{1}{N} \mathbf{u}^T \bar{\mathbf{X}} \bar{\mathbf{X}}^T \mathbf{u} \\
&= \frac{1}{N} \mathbf{u}^T \mathbf{X} \left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T \mathbf{X}\mathbf{X}^T \mathbf{w}_1} \mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X} \right) \left(\mathbf{I} - \frac{1}{\mathbf{w}_1^T \mathbf{X}\mathbf{X}^T \mathbf{w}_1} \mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X} \right) \mathbf{X}^T \mathbf{u} \\
&= \frac{1}{N} \mathbf{u}^T \mathbf{X} \left(\mathbf{I} - \frac{1}{N} \mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X} \right) \left(\mathbf{I} - \frac{1}{N} \mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X} \right) \mathbf{X}^T \mathbf{u} \\
&= \frac{1}{N} \mathbf{u}^T \left(\mathbf{X}\mathbf{X}^T - \frac{2}{N} \mathbf{X}\mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X}\mathbf{X}^T + \frac{1}{N^2} \mathbf{X}\mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X}\mathbf{X}^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{X}\mathbf{X}^T \right) \mathbf{u} \\
&= \mathbf{u}^T \left(\mathbf{C} - 2 \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{C} + \mathbf{C} \mathbf{w}_1 \underbrace{\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1}_{=1} \mathbf{w}_1^T \mathbf{C} \right) \mathbf{u} \\
&= \mathbf{u}^T \left(\mathbf{C} - \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{C} \right) \mathbf{u} \\
&= \mathbf{u}^T \left(\mathbf{I} - \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \right) \mathbf{C} \mathbf{u}
\end{aligned} \tag{121}$$

In the direction of $\mathbf{w}_1$ the variance is therefore

$$\mathbf{w}_1^T \left(\mathbf{I} - \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \right) \mathbf{C} \mathbf{w}_1 = \underbrace{\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1}_{=1} - \underbrace{\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1}_{=1} \underbrace{\mathbf{w}_1^T \mathbf{C} \mathbf{w}_1}_{=1} = 0 \,, \tag{122}$$

as wanted, and in the direction of another CCA component $\mathbf{w}_k$ (for which $\mathbf{w}_k^T \mathbf{C} \mathbf{w}_k = 1$) the variance is

$$\begin{aligned}
\mathbf{w}_k^T \left(\mathbf{I} - \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \right) \mathbf{C} \mathbf{w}_k &= \underbrace{\mathbf{w}_k^T \mathbf{C} \mathbf{w}_k}_{=1} - \mathbf{w}_k^T \mathbf{C} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{C} \mathbf{w}_k \\
&= 1 - \underbrace{(\mathbf{w}_k^T \mathbf{C} \mathbf{w}_1)^2}_{=0} = 1 \,,
\end{aligned} \tag{123}$$

since the CCA projections are required to be orthogonal in the sense: $\mathbf{w}_k^T \mathbf{C} \mathbf{w}_i = 0$. However, this orthogonality constraint does not imply the orthogonality of the projection vectors ($\mathbf{w}_k^T \mathbf{w}_i \neq 0$), as was in the case of PCA.

Appendix 2

Example of Negative Correlations between Stimuli

If all the experimental stimulus features are non-overlapping, there emerge strong negative correlations between the different stimuli. Such correlations can, however, be avoided to some extent in the experimental design by delivering the stimuli in some blocks simultaneously and in some blocks separately.

In the following illustrative example the sequences consist of 8 time points, constructed so that when one of the stimuli is "on", the other one is "off". The sequence in panel **a** is $\mathbf{s}_1 = [1, 0, 1, 0, 1, 0, 1, 0]$ and the sequence in panel **b** is $\mathbf{s}_2 = [0, 1, 0, 1, 0, 1, 0, 1]$.

a

b

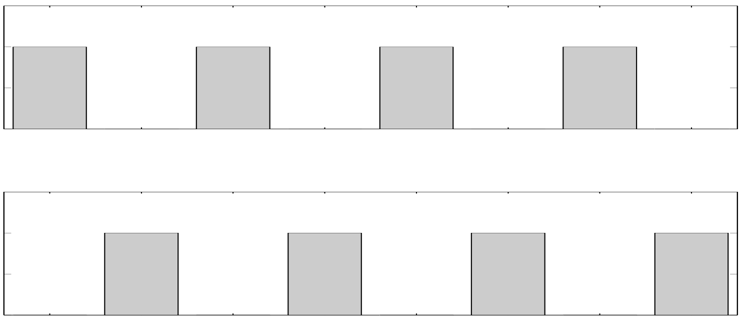


Figure 13: Simple time courses of totally exclusive stimulus sequences.

According to the definition of correlation, Eq. (16), the correlation between these two sequences is

$$\rho(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{4} \times \frac{-1-1-1-1-1-1-1-1}{\sqrt{2} \cdot \sqrt{2}} = -1 \quad . \tag{124}$$

Therefore, there exists a linear mapping from one sequence to the other between $\mathbf{s}_1$ and $\mathbf{s}_2$ and they are negatively correlated with coefficient -1, exactly because they are defined so strictly non-overlapping in order to avoid any *positive* correlations between the sequences.

In the case of only two different stimuli it is easy to construct such a balanced design, where knowing the other stimulus sequence does not convey information about the other. *E.g.*, for sequence $\mathbf{s}_3 = [1, 0, 0, 1, 0, 0, 1, 1]$, the correlation becomes

$$\rho(\mathbf{s}_1, \mathbf{s}_3) = \frac{1}{4} \times \frac{+1+1-1-1-1+1+1-1}{\sqrt{2} \cdot \sqrt{2}} = 0 \quad . \tag{125}$$

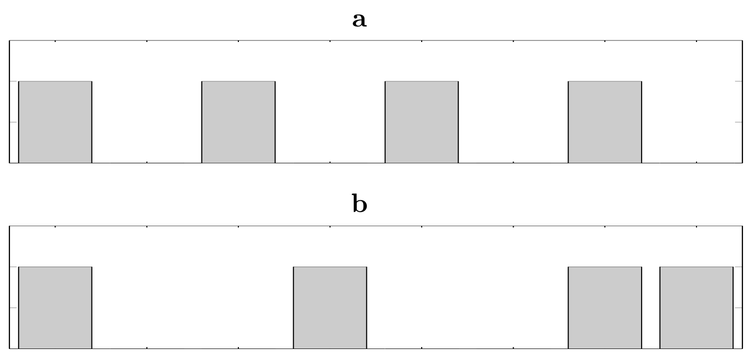


Figure 14: An illustrative example of a balanced setup where the stimulus sequences are not correlated (at all). However, they blocks do overlap one fourth of the time.

When there are k different stimuli, the balanced setup would require balancing the 2^k different combinations each to occur $1/2^k$ of the time. In this example there were only 4 different combinations, but in real experiments there would naturally be more of them.[28]

[28]With the original fMRI-study we analyzed in Publications 7 and 9, there were 7 different stimuli, resulting in $2^7 = 128$ different combinations to be balanced.

References

G. K. Aguirre, E. Zarahn, and M. D'Esposito. The Inferential Impact of Global Signal Covariates in Functional Neuroimaging Analyses. *NeuroImage*, 8(3):302–306, 1998. doi: 10.1006/nimg.1998.0367.

A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. *Information Processing & Management*, 39(1):45–65, 2003. doi: 10.1016/S0306-4573(02)00021-3.

A. Ajanki, D. R. Hardoon, S. Kaski, K. Puolamäki, and J. Shawe-Taylor. Can Eyes Reveal Interest? Implicit Queries from Gaze Patterns. *User Modeling and User-Adapted Interaction*, 19:307–339, 2009. doi: 10.1007/s11257-009-9066-4.

A. Anagnostopoulos, A. Dasgupta, and R. Kumar. Approximation Algorithms for Co-Clustering. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 201–210, New York, NY, USA, 2008. ACM.

Arabica. Arabica: Reliable ICA toolbox, 2008. `http://launchpad.net/arabica`.

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

C. Archambeau and F. Bach. Sparse Probabilistic Projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, 2009.

F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

G. H. Ball and D. J. Hall. A Clustering Technique for Summarizing Multivariate Data. *Behavioral Science*, 12:153–155, 1967. doi: 10.1002/bs.3830120210.

J. Basilico and T. Hofmann. Unifying Collaborative and Content-Based Filtering. In R. Greiner and D. Schuurmans, editors, *Proceedings of The Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 65–72. Omnipress, Madison, WI, 2004.

A. J. Bell and T. J. Sejnowski. An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129.

R. M. Bell and Y. Koren. Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights. In *ICDM'07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 43–52, Washington, DC, USA, 2007. IEEE Computer Society. doi: 10.1109/ICDM.2007.90.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons Ltd, Chichester, England, 2000.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. M. Bishop. Bayesian PCA. In M. S. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 382–388. MIT Press, 1999.

D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. M. Blei and M. I. Jordan. Modeling Annotated Data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134. ACM Press, 2003.

F. Bohnert, D. F. Schmidt, and I. Zukerman. Spatial Processes for Recommender Systems. Technical Report 2009/238, Faculty of Information Technology, Monash University, VIC 3800, Australia, 2009.

P. A. Bromiley, M. Pokric, and N. A. Thacker. Empirical Evaluation of Covariance Estimates for Mutual Information Coregistration. In C. Barillot, D. R. Haynor, and P. Hellier, editors, *Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI, Part I*, Lecture Notes in Computer Science, pages 607–614. Springer, 2004.

W. Buntine. Variational Extensions to EM and Multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the Thirteenth European Conference on Machine Learning, ECML'02*, volume 2430 of *Lecture Notes in Artificial Intelligence*, pages 23–34. Springer-Verlag, 2002.

R. S. Calsaverini and R. Vicente. An Information Theoretic Approach to Statistical Dependence: Copula Information. arXiv:0911.4207v1. To appear in Europhysics Letters, 2009. http://arxiv.org/abs/0911.4207v1.

M. G. Calvo and E. Meseguer. Eye Movements and Processing Stages in Reading: Relative Contribution of Visual, Lexical and Contextual Factors. *The Spanish Journal of Psychology*, 5(1):66–77, 2002.

N. A. Campbell and J. B. Reece. *Biology*. Benjamin Cummings, San Francisco, 6th edition, 2001.

G. Casella and E. I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46 (3):167–174, 1992. doi: 10.2307/2685208.

H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young. Remodeling of Yeast Genome Expression in Response to Environmental Changes. *Molecular Biology of the Cell*, 12:323–337, 2001.

K. Ciuffreda and B. Tannen. *Eye Movement Basics for the Clinician*. Mosby Yearbook, St. Louis, 1995.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

G. Doyle and C. Elkan. Accounting for Burstiness in Topic Models. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning (ICML'09)*, 2009.

D. Dykxhoorn, C. Novina, and P. Sharp. Killing the Messenger: Short RNAs that Silence Gene Expression. *Nature Reviews: Molecular Cell Biology*, 4:457–467, June 2003.

D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic Generation of Social Tags for Music Recommendation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 385–392, Cambridge, MA, 2008. MIT Press.

C. Eckart and G. Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1:211–218, 1936.

B. Efron. Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association*, 100(469):1–5, 2005a. doi: 10.1198/016214505000000033.

B. Efron. Modern Science and the Bayesian-Frequentist Controversy. Technical report, Department of Statistics, Stanford University, January 2005b.

C. Elkan. Deriving TF-IDF as a Fisher Kernel. In *Proceedings of the International Symposium on String Processing and Information Retrieval (SPIRE'05)*, pages 296–301, Buenos Aires, Argentina, 2005.

E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-Membership Models of Scientific Publications. *Proceedings of the National Academy of Sciences*, 101:5220–5227, 2004.

FastICA. MATLAB$^{\mathrm{TM}}$package, 1998.
`http://www.cis.hut.fi/research/ica/fastica`.

J. W. Fisher III and T. Darrell. Speaker Association with Signal-Level Audiovisual Fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

C. Fyfe and P. Lai. ICA Using Kernel Canonical Correlation Analysis. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, 2000.

S. S. Galiani. Copula Functions and their Application in Pricing and Risk Managing Multi-name Credit Derivative Products. Master's thesis, Department of Mathematics, King's College London, September 2003.

A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

S. Ghebreab, A. W. Smeulders, and P. Adriaans. Predictive Modeling of fMRI Brain States Using Functional Canonical Correlation Analysis. In *AIME'07: Proceedings of the 11th Conference on Artificial Intelligence in Medicine*, pages 393–397, Berlin, Heidelberg, 2007. Springer-Verlag. doi: 10.1007/978-3-540-73599-1_53.

S. Ghebreab, A. W. M. Smeulders, and P. W. Adriaans. Predicting Brain States from fMRI Data: Incremental Functional Principal Component Regression. In *Advances in Neural Information Processing Systems 20*, 2008.

G. Golub and C. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

A. Gretton, A. Belitski, Y. Murayama, B. Schölkopf, and N. Logothetis. The Effect of Artifacts on Dependence Measurement in fMRI. *Magnetic Resonance Imaging*, 24(4):401–409, May 2006. doi: 10.1016/j.mri.2005.12.036.

D. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information Retrieval by Inferring Implicit Queries from Eye Movements. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS'07*. Society for Artificial Intelligence and Statistics, 2007a.

D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor. Canonical Correlation Analysis; An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, 2004.

D. R. Hardoon, J. Mourão-Miranda, M. Brammer, and J. Shawe-Taylor. Unsupervised fMRI Analysis. In *NIPS Workshop on New Directions on Decoding Mental States from fMRI Data*, Whistler, Canada, 2007b.

D. R. Hardoon, J. Mourão-Miranda, M. Brammer, and J. Shawe-Taylor. Unsupervised Analysis of fMRI Data Using Kernel Canonical Correlation. *NeuroImage*, 37(4):1250–1259, 2007c. doi: 10.1016/j.neuroimage.2007.06.017.

R. Hari and M. V. Kujala. Brain Basis of Human Social Interaction: From Concepts to Brain Imaging. *Physiological Reviews*, 89(2):453–479, April 2009.
doi: 10.1152/physrev.00041.2007.

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

G. He, H.-G. Müller, and J.-L. Wang. Functional Canonical Analysis for Square Integrable Stochastic Processes. *Journal of Multivariate Analysis*, 85(1):54–77, 2003.
doi: 10.1016/S0047-259X(02)00056-8.

K. Heller, , S. Williamson, and Z. Ghahramani. Statistical Models for Partial Membership. In *Proceedings of the ICML'08*, 2008.

G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

T. Hofmann. Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems*, 22(1):89–115, 2004.

T. Hofmann, J. Puzicha, and M. I. Jordan. Learning from Dyadic Data. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 466–472. Morgan Kaufmann Publishers, San Mateo, CA, 1999.

H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *The Journal of Educational Psychology*, 24:417–441, 1933.

H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28:321–377, 1936.

S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc., 2nd edition, 2008.

R. A. Hutchinson, R. S. Niculescu, T. A. Keller, I. Rustandi, and T. M. Mitchell. Modeling fMRI Data Generated by Overlapping Cognitive Processes with Unknown Onsets Using Hidden Process Models. *NeuroImage*, 46:87–104, 2009.
doi: 10.1016/j.neuroimage.2009.01.025.

M. Hutter and M. Zaffalon. Distribution of Mutual Information from Complete and Incomplete Data. *Computational Statistics and Data Analysis*, 48(3):633–657, 2004. doi: 10.1016/j.csda.2004.03.010.

A. Hyrskykari, P. Majaranta, and K.-J. Räihä. From Gaze Control to Attentive Interfaces. In *Proceedings of HCII 2005*, Las Vegas, NV, July 2005. Lawrence Erlbaum Associates, Inc.

A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, October 1997. doi: 10.1162/neco.1997.9.7.1483.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, New York, NY, 1st edition, 2001.

R. Jin and L. Si. A Bayesian Approach toward Active Learning for Collaborative Filtering. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 278–285. AUAI Press, 2004.

S. Jokela. *Metadata Enhanced Content Management in Media Companies*. PhD thesis, Department of Computer Science and Engineering, Helsinki University of Technology, November 2001. `http://lib.tkk.fi/Diss/2001/isbn9512256932/`.

S. Jokela, M. Turpeinen, T. Kurki, E. Savia, and R. Sulonen. The Role of Structured Content in a Personalized News Service. In *HICSS '01: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, volume 7, page 7044, Washington, DC, USA, 2001. IEEE Computer Society.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 1986.

M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An Introduction to Variational Methods for Graphical Models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. MIT Press, Cambridge, 1999.

M. A. Just and P. A. Carpenter. Eye Fixations and Cognitive Processes. *Cognitive Psychology*, 8:441–480, 1976.

S. Kaski, J. Nikkilä, E. Savia, and C. Roos. Discriminative Clustering of Yeast Stress Response. In U. Seiffert, L. Jain, and P. Schweizer, editors, *Bioinformatics Using Computational Intelligence Paradigms*, pages 75–92, Berlin, 2005a. Springer.

S. Kaski, J. Nikkilä, J. Sinkkonen, L. Lahti, J. Knuuttila, and C. Roos. Associative Clustering for Exploring Dependencies Between Functional Genomics Data Sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics: Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, July–September 2005b.

S. Kaski, E. Savia, and K. Puolamäki. Predicting Binding of Transcriptional Regulators with a Two-Way Latent Grouping Model. Poster in *ISMB 2005, the 13th Annual International Conference on Intelligent Systems for Molecular Biology*, Detroit, Michigan, June 25–29, 2005c.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying Natural Images from Human Brain Activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

M. Keller and S. Bengio. Theme Topic Mixture Model: A Graphical Model for Document Representation. In *PASCAL Workshop on Text Mining and Understanding*, 2004.

D. Kelly and J. Teevan. Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum*, 37(2):18–28, 2003. doi: 10.1145/959258.959260.

M. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30:81–89, 1938.

J. Kettenring. Canonical Analysis of Several Sets of Variables. *Biometrika*, 58(3):433–451, 1971.

W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann. Center-Surround Patterns Emerge as Optimal Predictors for Human Saccade Targets. *Journal of Vision*, 9(5):1–15, 2009. doi: 10.1167/9.5.7.

A. Klami. *Modeling of Mutual Dependencies*. PhD thesis, Helsinki University of Technology, Faculty of Information and Natural Sciences. TKK Dissertations in Information and Computer Science, TKK-ICS-D6, Espoo, 2008. `http://lib.tkk.fi/Diss/2008/isbn9789512295203/`.

A. Klami and S. Kaski. Non-Parametric Dependent Components. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 209–212. IEEE, 2005.

A. Klami and S. Kaski. Local Dependent Components. In *ICML '07: Proceedings of the 24th International Conference on Machine learning*, pages 425–432, New York, NY, USA, 2007. ACM. doi: 10.1145/1273496.1273550.

M. Koivisto. *Sum-Product Algorithms for the Analysis of Genetic Risks*. PhD thesis, Department of Computer Science, University of Helsinki, January 2004. `http://ethesis.helsinki.fi/julkaisut/mat/tieto/vk/koivisto/sumprodu.pdf`.

J. Konstan, B. Miller, D. Maltz, and J. Herlocker. GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87, 1997.

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

X. N. Lam, T. Vu, T. D. Le, and A. D. Duong. Addressing Cold-Start Problem in Recommendation Systems. In *ICUIMC'08: Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, pages 208–211, New York, NY, USA, 2008. ACM. doi: 10.1145/1352793.1352837.

P. Lamere and D. Eck. Using 3D Visualizations to Explore and Discover Music. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 173–174, Vienna, Austria, 2007. Österreichische Computer Gesellschaft.

Y. Lashkari, M. Metral, and P. Maes. Collaborative Interface Agents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 444–449. AAAI Press, 1994.

D. Lee and H. Seung. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562, 2001.

S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller. Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks. In *ICML'07: Proceedings of the 24th International Conference on Machine Learning*, pages 489–496, New York, NY, USA, 2007. ACM. doi: 10.1145/1273496.1273558.

T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Tomphson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.

F. Lindskog. Modelling Dependence with Copulas and Applications to Risk Management. Master's thesis, ETH, Zürich, 2000.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, New Jersey, 2nd edition, 2002.

D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown. Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. *Nature Biotechnology*, 14:1675–80, 1996.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

S. C. Madeira and A. L. Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24–45, 2004.

R. Madsen, D. Kauchak, and C. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML'05)*, pages 545–552, 2005.

W. H. Mager and A. D. Kruijiff. Stress-Induced Transcriptional Activation. *Microbiological Reviews*, 59:506–531, 1995.

P. Majaranta and K.-J. Räihä. Twenty Years of Eye Typing: Systems and Design Issues. In *Proceedings of ETRA 2002, Eye Tracking Research and Applications Symposium*, pages 15–22, New Orleans, LA, USA, 2002. ACM Press.

P. Majaranta and K.-J. Räihä. Chapter 9: Text Entry by Gaze: Utilizing Eye-Tracking. In I. MacKenzie and K. Tanaka-Ishii, editors, *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann, San Francisco, 2007.

S. Malinen, Y. Hlushchuk, and R. Hari. Towards Natural Stimulation in fMRI – Issues of Data Analysis. *NeuroImage*, 35(1):131–139, 2007. doi: 10.1016/j.neuroimage.2006.11.015.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

B. Marlin. Modeling User Rating Profiles for Collaborative Filtering. In *Advances in Neural Information Processing Systems 16*, pages 627–634, Cambridge, MA, 2004a. MIT Press.

B. Marlin. Collaborative Filtering: A Machine Learning Perspective. Master's thesis, Department of Computer Science, University of Toronto, 2004b.

B. Marlin and R. S. Zemel. The Multiple Multiplicative Factor Model for Collaborative Filtering. In *ICML'04: Proceedings of the 21th International Conference on Machine Learning*, page 73. ACM Press, 2004.

B. Marlin, S. T. Roweis, and R. S. Zemel. Unsupervised Learning with Non-Ignorable Missing Data. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS'05*, pages 222–229. Society for Artificial Intelligence and Statistics, 2005.

B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney. Collaborative Filtering and the Missing at Random Assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI'07*, 2007.

A. McCallum, A. Corrada-Emmanuel, and X. Wang. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical report, University of Massachusetts, December 2004.

M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fMRI Data by Blind Separation Into Independent Spatial Components. *Human Brain Mapping*, 6(3):160–188, 1998.

G. J. McLachlan and K. E. Basford. *Mixture Models. Inference and Applications to Clustering*. Marcel Dekker, New York, NY, 1988.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, NY, 2000.

E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis. Modeling Dyadic Data with Binary Latent Factors. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, USA, 2007. MIT Press.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195, 2008. doi: 10.1126/science.1152876.

C. H. Morimoto and M. R. M. Mimica. Eye Gaze Tracking Techniques for Interactive Applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005. doi: 10.1016/j.cviu.2004.07.010.

K. Murphy, R. M. Birn, D. A. Handwerker, T. B. Jones, and P. A. Bandettini. The Impact of Global Signal Regression on Resting State Correlations: Are Anti-Correlated Networks Introduced? *NeuroImage*, 44:893–905, 2008. doi: 10.1016/j.neuroimage.2008.09.036.

I. Nabney. Efficient Training of RBF Networks for Classication. In *ICANN'99 Artificial Neural Networks*, 1999.

J. Nikkilä. *Exploratory Cluster Analysis of Genomic High-Throughput Data Sets and Their Dependencies*. PhD thesis, Helsinki University of Technology, Department of Computer Science. Dissertations in Computer and Information Science, Report D11, Espoo, 2005. `http://lib.tkk.fi/Diss/2005/isbn9512279096/`.

S. Ogawa, D. Tank, R. M. R.S., J. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil. Intrinsic Signal Changes Accompanying Sensory Stimulation: Functional Brain Mapping with Magnetic Resonance Imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5951–5955, 1992.

K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 2(6):559–572, 1901.

J. Peltonen, J. Goldberger, and S. Kaski. Fast Semi-Supervised Discriminative Component Analysis. In K. Diamantaras, T. Adali, I. Pitas, J. Larsen, T. Papadimitriou, and S. Douglas, editors, *Machine Learning for Signal Processing XVII*, pages 312–317. IEEE, 2007.

A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data environments. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI'01*, pages 437–444. Morgan Kaufmann, 2001.

M. Porta. Human–Computer Input and Output Techniques: An Analysis of Current Research and Promising Applications. *Artificial Intelligence Review*, 28(3):197–226, 2007. doi: 10.1007/s10462-009-9098-5.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155:945–959, 2000.

K. Puolamäki, S. Hanhijärvi, and G. C. Garriga. An Approximation Ratio for Biclustering. *Information Processing Letters*, 108:45–49, 2008.
doi: 10.1016/j.ipl.2008.03.013.

K. Rayner. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422, 1998.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309, December 2000.

A. Rényi. On Measures of Dependence. *Acta Mathematica Hungarica*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 487–494. AUAI Press, 2004.

S. Roweis. EM Algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems 10*, pages 626–632. MIT Press, 1998.

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.
doi: 10.1093/biomet/63.3.581.

H. Ruis and C. Schuller. Stress Signaling in Yeast. *Bioessays*, 17:959–965, 1995.

I. Rustandi, M. A. Just, and T. M. Mitchell. Integrating Multiple-Study Multiple-Subject fMRI Datasets Using Canonical Correlation Analysis. In *Proceedings of the MICCAI 2009 Workshop: Statistical Modeling and Detection Issues in Intra- and Inter-Subject Functional MRI Data Analysis*, 2009.

H. Sagan. *Introduction to the Calculus of Variations*. McGraw-Hill Publications, New York, NY, 1969.

R. Salakhutdinov and A. Mnih. Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *Proceedings of the Twenty-Fifth International Conference (ICML)*, pages 880–887. ACM, 2008.

F. Salmon. Recipe for Disaster: The Formula That Killed Wall Street. *Wired Magazine*, 17.03, 2009.

J. Salojärvi. *Inferring Relevance from Eye Movements with Wrong Models*. PhD thesis, Helsinki University of Technology, Faculty of Information and Natural Sciences. TKK Dissertations in Information and Computer Science, TKK-ICS-D8, Espoo, 2008.
`http://lib.tkk.fi/Diss/2008/isbn9789512296132/`.

J. Salojärvi, I. Kojo, J. Simola, and S. Kaski. Can Relevance Be Inferred from Eye Movements in Information Retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Kyushu Institute of Technology, Kitakyushu, Japan, 2003.

J. Salojärvi, K. Puolamäki, and S. Kaski. Relevance Feedback from Eye Movements for Proactive Information Retrieval. In J. Heikkilä, M. Pietikäinen, and O. Silvén, editors, *Proceedings of PSIPS 2004, Workshop on Processing Sensory Information for Proactive Systems*, pages 37–42. Infotech Oulu, Oulu, Finland, 2004.

J. Salojärvi, K. Puolamäki, and S. Kaski. Implicit Relevance Feedback from Eye Movements. In Duch, Kacprzyk, Oja, and Zadrozny, editors, *Artificial Neural Networks: Biological Inspirations – ICANN 2005: 15th International Conference. Proceedings, Part I*, volume Lecture Notes in Computer Science 3696, pages 513–518, Berlin, Germany, 2005. Springer-Verlag. doi: 10.1007/11550822.

J. Salojärvi, K. Puolamäki, J. Simola, L. Kovanen, I. Kojo, and S. Kaski. Inferring Relevance from Eye Movements: Feature Extraction. Technical Report A82, Helsinki University of Technology, Publications in Computer and Information Science, March 2005. `http://www.cis.hut.fi/eyechallenge2005/`.

E. Savia. Mathematical Methods for a Personalized Information Service. Master's thesis, Helsinki University of Technology, Department of Engineering Physics and Mathematics, November 1999.
`http://www.soberit.hut.fi/publications/SmartPush/sp_publications.html`.

M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470, 1995.

E. Segal, A. Battle, and D. Koller. Decomposing Gene Expression into Cellular Processes. In *Pacific Symposium on Biocomputing*, pages 89–100, 2003.

U. Shardanand and P. Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth". In *Proceeding of Computer Human Interaction*, pages 210–217, 1995.

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

M. B. Shelton, J. and A. Bartels. Semi-Supervised Subspace Analysis of Human Functional Magnetic Resonance Imaging Data. Technical Report TR-185, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, May 2009.

L. Si and R. Jin. Flexible Mixture Model for Collaborative Filtering. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning, ICML'03*, pages 704–711. AAAI Press, 2003.

C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann. Nonnegative CCA for Audiovisual Source Separation. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 253–258, 2007. doi: 10.1109/MLSP.2007.4414315.

J. Sinkkonen, S. Kaski, J. Nikkilä, and L. Lahti. Associative Clustering (AC): Technical Details. Technical Report A84, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2005.

C. E. Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 1904.

SPM2. MATLAB$^{\text{TM}}$package, 2002. `http://www.fil.ion.ucl.ac.uk/spm`.

M. Steyvers and T. Griffiths. Probabilistic Topic Models. In D. McNamara, T. Landauer, S. Dennis, and W. Kintsch, editors, *LSA: A Road to Meaning*. Erlbaum, Mahwah, NJ, 2005.

A. Tanay, R. Sharan, and R. Shamir. *Biclustering Algorithms: A Survey*, chapter Handbook of Computational Molecular Biology. Chapman and Hall/CRC Press, 2006.

D. Tennenhouse. Proactive Computing. *Communications of the ACM*, 43(5):43–50, 2000. doi: 10.1145/332833.332837.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

N. H. Timm. *Applied Multivariate Analysis*. Springer, New York, 2002.

N. Tintarev and J. Masthoff. Similarity for News Recommender Systems. In G. Uchyigit, editor, *Proceedings of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces*, 2006.

M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

A. Tripathi, A. Klami, and S. Kaski. Simple Integrative Preprocessing Preserves What is Shared in Data Sources. *BMC Bioinformatics*, 9:111, 2008.
doi: 10.1186/1471-2105-9-111.

A. Tsai, J. W. Fisher III, C. Wible, W. M. Weiss III, J. Kim, and A. S. Willsky. Analysis of Functional MRI Data Using Mutual Information. In C. Taylor and A. Colchester, editors, *Proceedings of the Second International Conference on Medical Image Computing and Computer-Assited Intervention*, volume 1679 of *Lecture Notes in Computer Science*, pages 473–480. Springer, 1999.

M. Turpeinen. *Customizing News Content for Individuals and Communities*. PhD thesis, Department of Computer Science and Engineering, Helsinki University of Technology, 2000. In series: Acta polytechnica Scandinavica.

R. Vertegaal. Designing Attentive Interfaces. In *ETRA '02: Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pages 23–30, New York, NY, USA, 2002. ACM. doi: 10.1145/507072.507077.

J. Vía, I. Santamaría, and J. Pérez. A Robust RLS Algorithm for Adaptive Canonical Correlation Analysis. In *Proceedingsof the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages iv/365 – iv/368, 2005.

R. Vigário and E. Oja. BSS and ICA in Neuroinformatics: From Current Practices to Open Challenges. *IEEE Reviews in Biomedical Engineering*, 1:50–61, 2008.

Votings of the British Parliament in 1997–2001, 2001.
`http://www.publicwhip.org.uk/project/data.php`.

C. Wang. Variational Bayesian Approach to Canonical Correlation Analysis. *IEEE Transactions on Neural Networks*, 18:905–910, 2007. doi: 10.1109/TNN.2007.891186.

J. Wang, A. P. de Vries, and M. J. T. Reinders. Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion. In *SIGIR'06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–508, New York, NY, USA, 2006. ACM.
doi: 10.1145/1148170.1148257.

K. West, S. Cox, and P. Lamere. Incorporating Machine-Learning into Music Similarity Estimation. In *AMCMM '06: Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, pages 89–96, New York, NY, USA, 2006. ACM. doi: 10.1145/1178723.1178737.

H. Wettig, J. Lahtinen, T. Lepola, P. Myllymäki, and H. Tirri. Bayesian Analysis of Online Newspaper Log Data. In *Proceedings of the 2003 Symposium on Applications and the Internet Workshops*, pages 282–278, Los Alamitos, California, 2003. IEEE Computer Society. doi: 10.1109/SAINTW.2003.1210173.

K. J. Worsley and K. J. Friston. Analysis of fMRI Time-Series Revisited – Again. *NeuroImage*, 2(3):173–235, 1995. doi: 10.1006/nimg.1995.1023.

X. Yin. Canonical Correlation Analysis Based on Information Theory. *Journal of Multivariate Analysis*, 91:161–176, 2004.

J. Ylipaavalniemi and J. Soppela. Arabica: Robust ICA in a Pipeline. In *8th International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, pages 379–386, Paraty, Brazil, March 2009. doi: 10.1007/978-3-642-00599-2_48.

J. Ylipaavalniemi and R. Vigário. Analysis of Auditory fMRI Recordings via ICA: A Study on Consistency. In *Proceedings of the IJCNN*, pages 249–254, Budapest, Hungary, 2004.

J. Ylipaavalniemi and R. Vigário. Analyzing Consistency of Independent Components: An fMRI Illustration. *NeuroImage*, 39(1):169–180, 2008. doi: 10.1016/j.neuroimage.2007.08.027.

K. Yu, S. Yu, and V. Tresp. Dirichlet Enhanced Latent Semantic Analysis. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS'05*, pages 437–444. Society for Artificial Intelligence and Statistics, 2005a.

S. Yu, K. Yu, V. Tresp, and H.-P. Kriegel. A Probabilistic Clustering-Projection Model for Discrete Data. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, editors, *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'05*, volume 3721 of *Lecture Notes in Computer Science*, pages 417–428. Springer, 2005b.

C. Zitnick and T. Kanade. Maximum Entropy for Collaborative Filtering. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 636–643. AUAI Press, 2004.