Department of Information and Computer Science

Bayesian latent variable models for learning dependencies between multiple data sources

Seppo Virtanen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 110/2014

Bayesian latent variable models for learning dependencies between multiple data sources

Seppo Virtanen

Doctoral dissertation for the degree of Doctor of Science in Techology to be presented with due permission of the School of Science for public examination and debate in Auditorium T2 at the Aalto University School of Science (Espoo, Finland) on the 25th of August 2014 at 12 noon.

Aalto University
School of Science
Department of Information and Computer Science
Statistical Machine Learning and Bioinformatics Group

Supervising professor
Prof. Samuel Kaski

Thesis advisor
Dr. Arto Klami

Preliminary examiners
Asst. Prof. Teemu Roos, University of Helsinki, Finland
Dr. Guillaume Obozinski, Ecole des Ponts - ParisTech, France

Opponent
Dr. Cédric Archambeau, Amazon Berlin, Germany and University College London, United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 110/2014

ISBN 978-952-60-5784-2
ISBN 978-952-60-5785-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5785-9

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Seppo Virtanen

Name of the doctoral dissertation
Bayesian latent variable models for learning dependencies between multiple data sources

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 110/2014

Field of research Computer and Information Science

Manuscript submitted 10 April 2014 **Date of the defence** 25 August 2014

Permission to publish granted (date) 6 June 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Machine learning focuses on automated large-scale data analysis extracting useful information from data collections. The data are frequently high-dimensional and may correspond, for example, to images, text documents, or measurements of neural responses. In many applications data can be collected from multiple data sources, that is, views.

This thesis presents novel machine learning methods for analyzing multiple data sources, especially for understanding relationships between them. The analysis provides a comprehensive summary of the data generating process, which may be used for exploring the relationships and for predicting observations of one or more sources. The methods are based on two assumptions: each view provides complementary information of the data generating process, and each view is corrupted by noise. The methods aim to utilize all available information (views), accumulating partly overlapping information and reducing view-specific noise.

In particular, this thesis presents several Bayesian latent variable models that learn a decomposition of latent variables; some of the variables capture information shared by multiple sources, whereas the remaining variables explain noise in each view. The latent variables may be efficiently inferred based on the observed data by using sparsity assumptions and Bayesian inference. The models are applied for analyzing neural responses to natural stimulation as well as for jointly modeling images and text documents.

Keywords Bayesian statistics, latent variable models, machine learning, multi-view learning, sparsity

ISBN (printed) 978-952-60-5784-2 **ISBN (pdf)** 978-952-60-5785-9

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 152 **urn** http://urn.fi/URN:ISBN:978-952-60-5785-9

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Seppo Virtanen

Väitöskirjan nimi
Bayesiläisiä piilomuuttujamalleja usean tietoaineiston välisten riippuvuuksien oppimiseen

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 110/2014

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 10.04.2014 **Väitöspäivä** 25.08.2014

Julkaisuluvan myöntämispäivä 06.06.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Koneoppiminen on suurten tietoaineistojen automaattista analysointia, jossa poimitaan hyödyllistä informaatiota näistä kokoelmista. Tietoaineistojen havainnot ovat usein moniulotteisia ja voivat esimerkiksi olla kuvia, tekstidokumentteja tai neuraalisia mittauksia. Monissa sovelluksissa aineistoja voidaan kerätä useista lähteistä.

Tässä väitöskirjassa esitellään usean tietolähteen analysointiin uusia koneoppimismenetelmiä, jotka löytävät lähteiden välisiä riippuvuuksia. Menetelmillä tehtävän analyysin tavoite on tarjota kattava tiivistelmä aineistot tuottaneesta prosessista. Tätä tiivistelmää voidaan käyttää riippuvuuksien tutkimiseen ja yhden tai useamman näkymän havaintojen ennustamiseen. Kehitetyt menetelmät perustuvat kahteen oletukseen: jokainen lähde sisältää osittaista tietoa aineistot tuottaneesta prosessista ja jokaisen lähteen sisältämä tieto on kohinaista. Menetelmät pyrkivät hyödyntämään kaikkea käytettävissä olevaa tietoa (lähteitä) kokoamalla yhteen aineistoissa olevaa päällekkäistä tietoa ja vähentämällä lähteille ominaista kohinaa.

Tämän väitöskirjan keskeinen tulos on joukko Bayesiläisiä piilomuuttujamalleja, jotka löytävät piilomuuttujien hajotelman; osa piilomuuttujista selittää lähteiden välistä yhteistä tietoa, kun taas jäljellä olevat muuttujat selittävät lähteille ominaista kohinaa. Piilomuuttujat voidaan tunnistaa havaitun aineiston perusteella tehokkaasti käyttämällä harvuusoletuksia ja Bayesiläistä päättelyä. Malleja on käytetty luonnollisen ärsykkeen neuraalisten vasteiden analysointiin sekä kuvien ja tekstidokumenttien yhteismallintamiseen.

Avainsanat Bayesiläinen tilastotiede, harvuus, koneoppiminen, oppiminen useasta tietolähteestä, piilomuuttujamallit

ISBN (painettu) 978-952-60-5784-2 **ISBN (pdf)** 978-952-60-5785-9

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 152 **urn** http://urn.fi/URN:ISBN:978-952-60-5785-9

Preface

This work has been carried out at the Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science (ICS), Aalto University. The work has been supported by the Finnish Doctoral Programme in Computational Sciences (FICS), the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170) and the aivoAALTO research project of Aalto University, including support to participate in scientific conferences abroad. My research visit to the group of Prof. Trevor Darrell at the International Computer Science Institute (ICSI) and University of California at Berkeley, USA, was supported by the Future Networks, Society and Modeling (FuNeSoMo) research exchange project.

I am grateful to my thesis supervisor Prof. Samuel Kaski and advisor Dr. Arto Klami for their invaluable advice and guidance. They are co-authors in most of the publications of this thesis and this work would not exist without them. I thank Prof. Trevor Darrell and ICSI for hosting me and providing the facilities during my visit.

My compliments belong to all my co-authors: Suleiman Ali Khan, Eemeli Leppäaho, Yangqing Jia and Trevor Darrell. I would also like to thank Prof. Mikko Sams, Dr. Juha Salmitaival, Dr. Krister Wennerberg, M.Sc. Enrico Glerean and M.Sc. Jaakko Luttinen for their help and fruitful discussions during my thesis.

I would like to thank the pre-examiners of my thesis, Asst. Prof. Teemu Roos and Dr. Guillaume Obozinski for providing useful comments to further improve the thesis.

I am thankful to the research group of Prof. Kaski and the ICS department for providing the facilities and enjoyable academic working environment. Special thanks belong to my current and former colleagues, both fellow students and postdocs, at the research group for their friendship,

splendid time together and for their support.

Finally, I am indebted to my parents Aino and Veli and to my brother Harri for always being there for me.

Espoo, July 11, 2014,

Seppo Virtanen

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Seppo Virtanen, Arto Klami and Samuel Kaski. Bayesian CCA via group sparsity. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning*, pages 457–464, 2011.

II Arto Klami, Seppo Virtanen and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

III Seppo Virtanen, Arto Klami, Suleiman A. Khan and Samuel Kaski. Bayesian group factor analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of JMLR W&CP, pages 1269–1277, 2012.

IV Arto Klami, Seppo Virtanen, Eemeli Leppäaho and Samuel Kaski. Group factor analysis. *a journal*, 2014.

V Arto Klami, Seppo Virtanen and Samuel Kaski. Bayesian exponential family projections for coupled data sources. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 286–293, 2010.

VI Seppo Virtanen, Yangqing Jia, Arto Klami and Trevor Darrell. Factorized multi-modal topic model. In *Proceedings of the Twenty-Eighth Con-*

ference on Uncertainty in Artificial Intelligence, pages 843–851, 2012.

Author's Contribution

Publication I: "Bayesian CCA via group sparsity"

Publication I presents an improved algorithm for Bayesian canonical correlation analysis (BCCA).

The experiments and ideas presented in the paper were jointly developed and the manuscript was jointly written. In particular, the author proposed to use group-wise sparse priors, implemented the model and performed the experiments.

Publication II: "Bayesian canonical correlation analysis"

Publication II extends Publication I and provides an extensive literature review on BCCA.

The article was jointly designed and written. Dr. Klami had most responsibility for writing the paper and executing the experiments. The author was responsible for deriving modeling formulations and implementations of the models presented in the paper.

Publication III: "Bayesian group factor analysis"

Publication III introduces a novel problem formulation of group factor analysis and a model for solving it, for capturing shared information (statistical dependencies) between multiple views.

The modeling idea was developed jointly by the author, Dr. Klami and Prof. Kaski, whereas the author implemented the model. MSc. Khan carried out the biomedical experiment, while all the remaining experiments

were carried out by the author. All the authors took part in writing the manuscript.

Publication IV: "Group factor analysis"

Publication IV extends Publication III and proposes an improved prior for solving the GFA problem formulation.

The modeling ideas and experiments were designed jointly, whereas the author put the ideas into effect with MSc. Leppäaho. The author performed the neuroimaging application, while MSc. Leppäaho performed the biomedical application. The manuscript was written together.

Publication V: "Bayesian exponential family projections for coupled data sources"

Publication V generalizes BCCA for data domains in natural exponential family, relaxing the limiting modeling assumptions of continuous-valued data.

The author implemented the model and executed the experiments. The original modeling idea was developed by Prof. Kaski and Dr. Klami, and the paper was written jointly.

Publication VI: "Factorized multi-modal topic model"

Publication VI introduces a novel multi-view topic model based on hierarchical Dirichlet processes (HDP) for learning topics that are either shared between the views or specific for one view. The model combines the modeling framework developed in Publications I and II for BCCA with topic modeling.

This paper is the outcome of the author's research visit to ICSI. The general idea of the paper was decided together by the whole of the author team at the beginning of the visit, whereas the author proposed the specific computational model, derived the formulas and implemented it, and ran the experiments. The data collection used in the paper was prepared by Mr. Jia. The paper was written jointly, with the first three authors sharing the main responsibility.

1. Introduction

Recent advances in information technology and computer science have essentially changed almost every branch of science and engineering from data-poor to data-rich, calling for up-to-date data analysis methods to conduct the research. Machine learning plays an important role in developing such methods and is becoming increasingly important. One of the most significant current research directions in machine learning is based on Bayesian statistics, using probability theory to construct such methods in a unified and well principled manner.

The concept of a data generating process is central to the discipline of machine learning. This process induces structure in the observed data. In the history of machine learning, probabilistic generative latent variable models have been developed and used to capture such structure in the data. More recently, benefits of the Bayesian approach have been implemented in practice, providing efficient learning algorithms and principles for making inferences regarding the latent variables, as well as other unknown quantities of the model, based on the observed data.

Over the past decade, there has been increasing interest in collecting data from multiple sources or views. For example, web images co-occur with the surrounding text on the page, and both of these views are useful for analyzing the web content. The motivation of this approach is that each view is assumed to provide complementary information regarding the underlying process generating the views. The corresponding aim is to utilize all available views to provide a more complete understanding of the process. In particular, in many research fields it is essential to discover the process for understanding interactions between the views.

It is becoming increasingly important to jointly model multiple data sources. Each view may include incomplete and potentially weak information, as well as be corrupted by noise. Joint modeling of multiple views

is able to overcome both of these challenges. On the one hand, weak information from the views can be accumulated to provide a broader understanding of the process under study. On the other hand, the side-effect of noise in each view can be circumvented by emphasizing common structures shared by multiple views.

This thesis studies Bayesian latent variable models for multi-view data. The studied models explain the data collection and capture interactions between the views. This task is called dependency learning. The modeling approach is based on a probabilistic interpretation of canonical correlation analysis [CCA; Hotelling, 1936, Hardoon et al., 2004] as a latent variable model by Bach and Jordan [2005] and Browne [1979]. The CCA model assumes a shared process between the views that captures the dependencies. Although Bayesian inference for CCA has been presented by Klami and Kaski [2007] as well as Wang [2007] and theory exists for modeling dependencies [Klami and Kaski, 2006, 2008], several fundamental research issues remain to be studied. In particular, the existing theory and models for learning dependencies are suitable only for limited settings and suffer from inefficient learning algorithms, hindering real-world applications.

1.1 Contributions and organization of the thesis

This thesis presents several Bayesian latent variable models for learning dependencies between multiple views. These models remove some of the limitations of previous research, detailed in the following paragraphs, and advance the modeling theory. The models are applied to joint modeling of images and co-occurring text documents as well as to problems in computational neuroscience.

Publications I and II present a novel model formulation and inference algorithm for Bayesian CCA (BCCA) for capturing dependencies between two views. The novel solution results in considerably more efficient inference for BCCA, especially for high-dimensional data. These papers advance the state of the art by enabling real-word applications of BCCA, which were previously not possible to solve. In addition, Publication II provides an extensive review of BCCA and related extensions.

Publications III and IV introduce a new problem formulation referred to as group factor analysis (GFA) and models for solving it. GFA extends BCCA for learning dependencies between more than two views in a flex-

ible way. A key novelty is that GFA accounts for both inter-view and within-view dependencies. In particular, Publication IV suggests a novel prior for GFA, enabling applications with very large number of views.

Publication V generalizes BCCA to data domains in the natural exponential family distributions. Most applications of BCCA have been carried out only for continuous-valued data severely restricting the scope of research. The new model removes this constraint, thus enabling principled applications suitable for binary or integer valued data, for example.

Publication VI presents a novel model that combines BCCA-type modeling for learning dependencies with topic modeling. Topic models are generative models for discrete document data. Thus far, topic modeling has paid little attention to learning dependencies between multiple views: most topic models for multi-view document data make simplifying assumptions limiting their use in less controlled setups. The proposed multi-view topic model learns topics that capture dependencies both between and within the views.

Chapter 2 reviews the necessary background on probabilistic machine learning. In particular, it discusses various latent variable models for a single view that serve as a basis for the multi-view models. Chapter 3 contains the contributions of this thesis, and presents the main ideas behind the developed models, comparing them with previous related work. Finally, Chapter 4 concludes this thesis and offers directions for future work.

2. Bayesian machine learning

Machine learning focuses on automated large-scale data analysis, where the general goal is to extract useful information from data, such as images, text documents, gene expression or brain imaging measurements, and annotations or labels obtained by human expertise or from previous data analyses. Common tasks in machine learning are prediction and summarization, corresponding to two (partly over lapping) fields of machine learning called supervised and unsupervised learning, respectively.

Given a collection of paired observations, inputs and outputs, supervised learning is defined as learning a relationship between them, such that for an unseen input the output can be predicted accurately. Unsupervised learning, on the other hand, usually considers observations from a single source and the aim is to learn a useful description for the data collection.

A considerable amount of recent machine learning research uses Bayesian statistics. See Barber [2012], Bishop [2006], Bernardo and Smith [1994], Gelman et al. [2003], Kollar and Friedman [2009], MacKay [2003], and Murphy [2012] for recent textbook accounts. Bayesian machine learning is inherently modular and may be considered to consist of three separate, although related, stages:

i) defining a model (probabilistic description) for data,

ii) learning (or inferring) unknown quantities (that is, parameters and latent variables) of the model based on the observed data and,

iii) evaluating, interpreting and using the inferred quantities for various tasks.

A large and growing body of machine learning literature has investigated factor analysis and related latent variable models for a wide range of applications. This chapter focuses on Bayesian machine learning, which forms the basis of the models developed in the thesis. In particular, the

chapter discusses latent variable models, such as factor analysis, for unsupervised single view and multi-view learning.

2.1 Notation

A D-dimensional column vector $\mathbf{x}$ is denoted as $\mathbf{x} \in \mathbb{K}^D$, where $\mathbb{K}$ indicates the domain that is usually the set of real numbers, $\mathbb{R}$. A $D \times K$ matrix $\mathbf{A}$ is compactly written as $\mathbf{A} \in \mathbb{K}^{D \times K}$.

For matrices and vectors, subscripts are used to indicate the individual elements, with $\mathbf{W}_{:,j}$ denoting the whole jth column of $\mathbf{W}$ and $\mathbf{W}_{i,:}$ denoting the ith row transposed to a column vector. Finally, we use $\mathbf{0}$ and $\mathbf{I}$ to denote zero and identity matrices of sizes which make sense in the context, without cluttering the notation.

In general, probability densities are written as $p(\mathbf{x})$, whereas more explicit notation is alternatively used to specify the distribution in question. For example, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a random variable $\mathbf{x}$ drawn from a multivariate Gaussian distribution with parameters mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.2 Bayesian inference

Bayesian models define joint probability distributions for observed (data) variables, $\mathbf{X}$, and unobserved quantities, $\boldsymbol{\Theta}$, expressing a generative process for $\mathbf{X}$ that depends on $\boldsymbol{\Theta}$. The joint probability distribution may be written as,

$$p(\mathbf{X}, \boldsymbol{\Theta}) = p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}),$$

where the first term on the right hand side is a conditional (joint) probability distribution of $\mathbf{X}$ given $\boldsymbol{\Theta}$ and the second term, respectively, is a (joint) marginal distribution of $\boldsymbol{\Theta}$. The $\boldsymbol{\Theta}$ may contain parameters as well as (random) variables. The distributions of the parameters are often referred to as prior distributions, reflecting the *a priori* belief on those parameters.

The statistical inference then proceeds by calculating the conditional distribution of the unknown quantities[1] given the observations. This distribution is referred to as the posterior distribution of the unknown quan-

[1]Note, however, that before computations some of the parameters need to be assigned to known values. In the following, for notational simplicity, conditioning on the known parameters is omitted.

tities,

$$p(\boldsymbol{\Theta}|\mathbf{X}) = \frac{p(\mathbf{X}, \boldsymbol{\Theta})}{p(\mathbf{X})}. \tag{2.1}$$

The $p(\mathbf{X})$ is a normalization constant, alternatively referred to as evidence or marginal likelihood of the model. It is obtained by *marginalizing* over $\boldsymbol{\Theta}$:

$$p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}. \tag{2.2}$$

High posterior probability for $\boldsymbol{\Theta}$ indicates $\boldsymbol{\Theta}$ explains the observations well and possesses significant marginal (prior) probability. Uncertainties concerning the particular values for $\boldsymbol{\Theta}$ are represented in the posterior distribution in a natural way.

For most models the normalization constant $p(\mathbf{X})$ required for the posterior distribution (2.1) cannot be computed in closed form. Consequently, several approximate inference algorithms have been proposed to carry out computations (see Section 2.2.1). However, accuracy and computational load of these methods depend heavily on the model assumptions, preferring certain (usually simple) models. Thus, model construction needs to balance between background knowledge, mathematical convenience and computational tractability.

The posterior distribution can be used to generate new data or evaluate the probability of unseen data under the model. For new data $\mathbf{X}^*$ the prediction is written as

$$p(\mathbf{X}^*|\mathbf{X}) = \int p(\mathbf{X}^*|\boldsymbol{\Theta}, \mathbf{X}) p(\boldsymbol{\Theta}|\mathbf{X}) d\boldsymbol{\Theta}. \tag{2.3}$$

The tasks of machine learning may involve summarization and/or prediction. In the Bayesian setting, summarization is performed by inspecting the posterior distribution (2.1) for some variables, whereas prediction is based on equation (2.3).

Given a set of possible models for explaining $\mathbf{X}$, corresponding to alternative modeling assumptions, model selection is the process of choosing a single model that best represents the observations or is the most useful. Bayesian model selection proposes to choose the model that has the highest evidence (2.2), weighted by subjective model probabilities. A more practical alternative is to choose the model that maximizes some external performance measure for utility, such as predictive accuracy. These two solutions may not necessarily lead to equivalent solutions.

2.2.1 Approximate Bayesian inference

This thesis uses variational Bayes, Markov Chain Monte Carlo and empirical Bayes for approximate inference. These approaches are described in the following.

Variational Bayes

The variational Bayesian approach approximates the true posterior distribution by a simpler trial distribution $q(\boldsymbol{\Theta})$ [Bishop, 2006, Jordan et al., 1999]. The idea is to choose $q(\boldsymbol{\Theta})$ such that the marginalization becomes tractable. One frequently used procedure to achieve this is to assume that $\boldsymbol{\Theta}$ is factored into separate sets $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i\}_{i=1}^I$. Then, the trial distribution is given as $q(\boldsymbol{\Theta}) = \prod_{i=1}^I q(\boldsymbol{\Theta}_i)$. The parameters for the various distributions $q(\boldsymbol{\Theta}_i)$ in the approximation are updated alternatingly to minimize the Kullback-Leibler divergence $D_{KL}(q,p)$ between $q(\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta}|\mathbf{X})$ to obtain an approximation best matching the true posterior. Equivalently, the task is to maximize

$$\mathcal{L}(q) = \log p(\mathbf{X}) - D_{KL}(q,p) = \int q(\boldsymbol{\Theta}) \log \frac{p(\boldsymbol{\Theta}, \mathbf{X})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta}, \tag{2.4}$$

lower bounding the model evidence.

Variational Bayes is used for inference in Publications I, II, III, VI and IV.

Markov chain Monte Carlo methods

Markov Chain Monte Carlo (MCMC) methods construct a Markov chain over $\boldsymbol{\Theta}$ whose stationary distribution is the posterior distribution (2.1) [Gelfand and Smith, 1990, Geman and Geman, 1984, Hastings, 1970, Metropolis et al., 1953, Robert and Casella, 2004]. The chain proceeds iteratively by drawing a value for $\boldsymbol{\Theta}_i$ from a proposal distribution starting from some initial point. Then samples are collected approximating the posterior distribution.

Metropolis-Hastings [MH; Metropolis et al., 1953, Hastings, 1970] proposes a new value for $\boldsymbol{\Theta}_i^*$ from a proposal distribution $p(\boldsymbol{\Theta}_i^*|\boldsymbol{\Theta}_i)$, accepting the new value with probability proportional to the joint model. For a symmetric proposal distribution, $p(\boldsymbol{\Theta}_i^*|\boldsymbol{\Theta}_i) = p(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_i^*)$, the acceptance probability is

$$a = \min\left(1, \frac{p(\mathbf{X}|\boldsymbol{\Theta}^*)}{p(\mathbf{X}|\boldsymbol{\Theta})}\right).$$

One interesting special case of MH is Gibbs sampling [Gelfand and Smith, 1990, Geman and Geman, 1984] that draws samples for $\boldsymbol{\Theta}_i$ from a con-

ditional distribution $p(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{-i})$ given current values for the remaining quantities.

Gibbs sampling and MH are used in Publications II and V.

Empirical Bayes

Empirical Bayes [Maritz and Lwin, 1989] seeks a point estimate for $\boldsymbol{\Theta}_i$ that maximizes the partial evidence $p(\mathbf{X}|\boldsymbol{\Theta}_i)$, (approximately) marginalizing over the remaining variables.

Publications IV and VI use empirical Bayes.

2.3 Latent variable models

Factor analysis and related latent variable models, which are detailed in the following, are essential tools for data analysis that express a generative process for observations in terms of a smaller number of unobserved (latent) variables. Such models provide a lower-dimensional representation of higher-dimensional data and can be used both for summarization and prediction. In practice, even though the observations are high-dimensional, one frequently finds that they lie close to a lower dimensional subspace, implying that the distribution of the observations is constrained or, alternatively, that the data variables are strongly correlated.

These models are increasingly relevant and have been shown to perform well for a large number of applications, such as denoising, dimensionality reduction, collaborative filtering, missing value imputation, gene expression analysis, brain signal analysis, computer vision, text document analysis, information retrieval, source separation, matrix factorization or decomposition, data visualization, feature extraction, topic modeling, clustering, mixed membership modeling, latent feature modeling, feature allocation and multi-way analysis (MANOVA), to name a few.

As explained in the introduction, data may be collected from multiple sources. That is, data are assumed from M sources, constituting multi-view data: for the nth object there are M vectorial D_m-dimensional observations $\mathbf{x}_n^{(m)} \in \mathbb{R}^{D_m}$, where $m = 1, \ldots, M$ and $n = 1, \ldots, N$. For $D_m = 1$ the mth source contains a single data variable and for $D_m \geq 2$ a group of variables, respectively.

Latent variable models for multi-view data have been applied to supervised dimensionality reduction, image annotation, multi-label prediction, context based information retrieval, data integration or fusion, data

translation, multi-way analysis for multiple views, and modeling relationships (statistical dependencies) between the views.

In the following, various latent variable models and prior distributions, relevant for the scope and the developed models of this thesis, are reviewed in necessary detail, whereas Bayesian treatment for some of these models is discussed in the next chapter. In particular, this section first discusses factor analysis and related linear Gaussian factor models, exponential family factor models and topic models for both single and multiple data sources. Then it examines useful prior distributions for these models.

As discussed in the previous section, exact inference is infeasible for many interesting models including the aforementioned latent variable models. Hence, in recent years, an increasing number of approximate posterior inference algorithms have been used. In the following, a few relevant works are mentioned. Variational Bayesian inference (see Section 2.2.1) has been applied to linear Gaussian factor models by Attias [1999, 2000], Ghahramani and Beal [2000], Wiegerinck [2000] and Xing et al. [2003], to natural exponential family factor models by Khan et al. [2010] and Seeger and Bouchard [2012], and to topic models by Asuncion et al. [2009], Blei et al. [2003] and Teh et al. [2007]. MCMC approaches (Section2.2.1) have been applied to linear Gaussian models by Salakhutdinov and Mnih [2008], and to exponential family models by Mohamed et al. [2009], and to topic models by Griffiths and Steyvers [2004] and Teh et al. [2006].

2.3.1 Models for single-view data

Principal component analysis

Principal component analysis [PCA; Pearson, 1901, Jolliffe, 2005] is a well established technique for dimensionality reduction of single-view data. A probabilistic interpretation for PCA by Tipping and Bishop [1999b] and Roweis [1998] allows writing the model for the nth observation as

$$\mathbf{x}_n = \mathbf{A}\mathbf{z}_n + \mathbf{e}_n.$$

Here, on the right hand side, the first term is a latent representation for $\mathbf{x}_n$ and the second term denotes noise that captures the remaining unstructured variation or measurement error, respectively. The representation is a linear combination of K (latent) factors $\mathbf{z}_n \in \mathbb{R}^K$, weighted by loadings $\mathbf{A} \in \mathbb{R}^{D\times K}$, which are common for all observations. The factors

provide a less noisy and condensed lower-dimensional representation of the observations. Probabilistic PCA is a classical linear Gaussian factor model where both noise and factors are drawn from Gaussian distributions. Equivalently, the model may be written as

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{A}\mathbf{z}_n, \tau^{-1}\mathbf{I}),$$
$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Here, the factors are drawn from a Gaussian distribution with zero mean and identity covariance matrix, and τ is a precision (inverse variance) parameter to explain additive unstructured residual to the combination. For notational clarity zero-mean data is assumed, hence a separate mean parameter is not included.

Factor analysis [FA; Spearman, 1904] is a closely related model to probabilistic PCA that includes a more flexible noise model. The model for FA is

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{A}\mathbf{z}_n, \mathbf{\Lambda}),$$
$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\mathbf{\Lambda}$ is a diagonal matrix with a separate element for each data variable. FA is commonly applied to capturing dependencies between the individual univariate data variables. Since the $\mathbf{\Lambda}$ explains independent variation for each variable, the factors focus on capturing dependencies (correlations) between them.

Collecting N observations $\mathbf{x}_n$ and factors $\mathbf{z}_n$ into a $D \times N$ matrix $\mathbf{X}$ and a $K \times N$ matrix $\mathbf{Z}$, respectively, probabilistic PCA and FA can be interpreted as a matrix decomposition. Variation is decomposed into factors and noise as

$$\mathbf{X} = \mathbf{W}\mathbf{Z} + \mathbf{E}, \tag{2.5}$$

where columns of $\mathbf{E}$ follow the Gaussian noise distribution.

A canonical example application for linear Gaussian factor models is biological pathway analysis for microarray gene expression data, where it is assumed that the factors represent biologically relevant information [Carvalho et al., 2008]. Another successful application is missing value imputation, illustrating that the factors learn predictive structure from data [Ilin and Raiko, 2010, Lim and Teh, 2007, Salakhutdinov and Mnih, 2008].

Exponential family principal component analysis

Exponential family principal component analysis and closely related models [EPCA; Collins et al., 2002, Moustaki and Knott, 2000, Tipping, 1999, Wedel and Kamakura, 2001] generalize the Gaussian noise assumption of probabilistic PCA to any distribution in the exponential family for taking the data domain into account in a well principled way. EPCA assumes observations $\mathbf{x}_n \in \mathbb{K}^D$, where $\mathbb{K}$ is a suitable subset of the real-space $\mathbb{K} \subseteq \mathbb{R}$, drawn from a natural exponential family distribution (see Bernardo and Smith [1994]),

$$\mathbf{x}_n \sim \mathcal{E}(\boldsymbol{\omega}_n) = h(\mathbf{x}_n) \exp(\mathbf{x}_n^T \boldsymbol{\omega}_n + g(\boldsymbol{\theta})).$$

Here, $\boldsymbol{\omega}_n \in \mathbb{K}^D$ denotes the natural parameters of the distribution for the nth observation, $g(\cdot)$ is a log-cumulant function specifying the distribution in question and $h(\cdot)$ is a function of data. The $\boldsymbol{\omega}_n$ is decomposed into a linear combination of K factors $\mathbf{z}_n$ and factor loadings $\mathbf{A}$, as $\boldsymbol{\omega}_n = \mathbf{A}\mathbf{z}_n$. The formulation covers a variety of different data domains such as binary or integer data, corresponding to Bernoulli and Poisson distributions, respectively. The expectation of $\mathbf{x}_n$ is given by transforming the natural parameters through a link function written as

$$\mathbb{E}[\mathbf{x}_n] = g'(\boldsymbol{\omega}_n),$$

where $g'(\cdot)$ is the link function, derivative of $g(\cdot)$.

Latent Dirichlet allocation

Latent Dirichlet allocation [LDA; Blei et al., 2003, Buntine, 2002] provides a generative model for document data. Observations correspond to documents, counts of discrete words, from a certain vocabulary over D words. Such a representation is called bag-of-words data. The model assumes K topics $\boldsymbol{\eta}_k$, where $k = 1, \ldots, K$, and the corresponding topic proportions $\boldsymbol{\theta}$. Both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_k$ are constrained to be probability distributions: $\boldsymbol{\eta}_k$ is a distribution over the vocabulary and $\boldsymbol{\theta}$ is a distribution over the topics. The generative model for an observation begins by drawing a topic proportion $\boldsymbol{\theta}$ from a Dirichlet distribution

$$\boldsymbol{\theta} \sim \mathcal{D}(\gamma \mathbf{1}),$$

where γ is a concentration parameter. Then, for the ith word w_i a topic indicator z_i is drawn first and then w_i is drawn from a multinomial distri-

bution corresponding to that topic,

$$z_i \sim \mathcal{M}(\boldsymbol{\theta}, 1), \tag{2.6}$$
$$w_i \sim \mathcal{M}(\boldsymbol{\eta}_{z_i}, 1).$$

The process (2.6) may be repeated for drawing multiple words for each document. The observations can be conveniently represented as D-dimensional vectors $\mathbf{x}_n$, whose elements correspond to word counts.

Even though the topic model was originally proposed for modeling text, it has also been applied to model images in computer vision applications [Sivic et al., 2005, Sivic and Zisserman, 2003]. LDA is commonly used for organizing a large collection of observations, facilitating information retrieval.

2.3.2 Models for multi-view data

Canonical correlation analysis

Canonical correlation analysis [CCA; Hotelling, 1936, Hardoon et al., 2004] is a well established method for dimensionality reduction of multi-view data and more importantly for capturing dependencies between two sets of variables, that is, views. A probabilistic interpretation of CCA [Bach and Jordan, 2005, Browne, 1979, de Bie and de Moor, 2003] assumes a latent representation that captures common variation (statistical dependencies) between the views. The model for the nth observation in the mth view is

$$\mathbf{x}_n^{(m)} = \mathbf{A}^{(m)}\mathbf{z}_n + \mathbf{e}_n^{(m)}. \tag{2.7}$$

Here, the factors $\mathbf{z}_n$ are shared between the views accounting for common variation, $\mathbf{A}^{(m)}$ is the corresponding loadings and $\mathbf{e}_n^{(m)}$ represents noise for the mth view accounting for any non-shared variation. In other words, the model assumes $\mathbf{x}_n^{(m)}$ depends on two sources of variation. The first source is shared between the views, whereas the second is independent of the other view. Accordingly, the model decomposes the variation in data to common variation between the views and view-specific variation.

Generative CCA is closely related to FA and probabilistic PCA. The crucial difference worth pointing out is the definition of the noise. Instead of assuming independent noise over the data variables, the model allows for arbitrary correlations between them. One technique to achieve this parameterizes the noise through a covariance matrix. As a result, the

observation model for $\mathbf{x}_n^{(m)}$ is

$$\begin{aligned} \mathbf{x}_n^{(m)} &\sim \mathcal{N}(\mathbf{A}^{(m)}\mathbf{z}_n, \mathbf{\Sigma}^{(m)}), \\ \mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{2.8}$$

where the $\mathbf{\Sigma}^{(m)}$ is an unconstrained noise covariance matrix for the mth view.

CCA is a ideal candidate for supervised dimensionality reduction and it has also been applied to multi-label prediction [Breiman and Friedman, 1997, Glahn, 1968, Ji et al., 2008, Kim and Pavlovic, 2009, Rai and Daumé III, 2009, Sun et al., 2011, Waugh, 1942]. Here, the two views consist of inputs and outputs as in supervised learning. CCA can extract a shared latent representation that captures the dependencies (containing relevant information for prediction), while **i)** discarding view-specific irrelevant variation, which is not useful for prediction, and **ii)** exploiting output (or within-view) correlations.

In addition CCA is frequently used in a symmetric setting, where one view is not considered more important than the other. The goals are then to evaluate the amount of dependency between the views or to find which of the variables show the dependency.

Inter-battery factor analysis

The CCA model is closely related to a probabilistic interpretation of inter-battery factor analysis [IBFA; Tucker, 1958] by Browne [1979]. In recent years, the IBFA model has been re-invented by multiple authors [Archambeau and Bach, 2008, Ek et al., 2008, Klami and Kaski, 2006, 2008, Leen, 2008] using different terminology, denoting the model as extended CCA or shared-private decomposition. However, all these models correspond to the one given by Browne [1979]. While the CCA model (2.8) parameterizes the correlated noise via the unconstrained covariance matrix $\mathbf{\Sigma}^{(m)}$, IBFA assumes a low-rank decomposition for the $\mathbf{\Sigma}^{(m)}$. The model becomes

$$\begin{aligned} \mathbf{x}_n^{(m)} &\sim \mathcal{N}(\mathbf{A}^{(m)}\mathbf{z}_n + \mathbf{B}^{(m)}\mathbf{z}_n^{(m)}, \tau_m^{-1}\mathbf{I}), \\ \mathbf{z}_n, \mathbf{z}_n^{(m)} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{2.9}$$

where the factors $\mathbf{z}^{(m)} \in \mathbb{R}^{K_m}$ with loadings $\mathbf{B}^{(m)} \in \mathbb{R}^{D_m \times K_m}$ affect only the mth view. The residual variation is explained by the precision τ_m.

The connection between the IBFA and CCA models is illustrated by

marginalizing $\mathbf{z}^{(m)}$,

$$\int \mathcal{N}(\mathbf{x}^{(m)}|\mathbf{A}^{(m)}\mathbf{z} + \mathbf{B}^{(m)}\mathbf{z}^{(m)}, \tau_m^{-1}\mathbf{I})\mathcal{N}(\mathbf{z}^{(m)}|\mathbf{0}, \mathbf{I})d\mathbf{z}^{(m)} =$$
$$\mathcal{N}(\mathbf{x}^{(m)}|\mathbf{A}^{(m)}\mathbf{z}, \mathbf{B}^{(m)}\mathbf{B}^{(m)^T} + \tau_m^{-1}\mathbf{I}) = \mathcal{N}(\mathbf{x}^{(m)}|\mathbf{A}^{(m)}\mathbf{z}, \mathbf{\Sigma}^{(m)}),$$

re-parameterizing the noise covariance matrix using a low-rank decomposition,

$$\mathbf{\Sigma}^{(m)} = \mathbf{B}^{(m)}\mathbf{B}^{(m)^T} + \tau_m^{-1}\mathbf{I},$$

without loss of generality. When the rank of the decomposition K_m, in essence, the number of specific factors for the mth view, equals the dimensionality of $\mathbf{x}_n^{(m)}$, the IBFA model is equivalent to probabilistic CCA.

Multiple battery factor analysis

Multiple battery factor analysis [MBFA; Browne, 1980, McDonald, 1970] generalizes IBFA/CCA to more than two views. Recently, many models equivalent to MBFA have been presented [Archambeau and Bach, 2008, Deun et al., 2011, Lock et al., 2013, Qu and Chen, 2011, Ray et al., 2013, Salzmann et al., 2010]. These models include both factors that are shared across all views and specific factors (or a flexible noise model) for each view.

Also, several straightforward generalizations of (E)PCA for multi-view data have been proposed [Guo, 2008, Ma et al., 2008, Rish et al., 2008, Shen et al., 2009, Singh and Gordon, 2008, Yu and Tresp, 2004, Yu et al., 2006, West, 2003]. These models assume a single set of factors to account for all variation, corresponding to more simple single view models for concatenated data.

Multi-view topic models

Web data sources, such as Facebook, Flickr and Instagram, provide rich sources of images accompanied with textual captions, words that describe the visual content of the images. Further, newspaper and Wikipedia articles contain pictures related to the content appearing in the document text. When both text and image observations are represented by bag-of-word descriptions, multi-modal LDA [Blei and Jordan, 2003] can be used to explain such data jointly. The model is frequently applied to text-based image retrieval and image annotation [Barnard et al., 2003, Blei and Jordan, 2003, Yakhnenko and Honavar, 2009]. The task of image annotation is to predict the text description for an unseen image, whereas the goal of text-based image retrieval is to retrieve well matching images to a text query.

Multi-modal LDA [Blei and Jordan, 2003] extends LDA by including separate topics $\boldsymbol{\eta}_k^{(m)}$ for the M views, while the topic proportion $\boldsymbol{\theta}$ for a document is shared across all views. Then, the ith word $w_i^{(m)}$ for an observation in the mth view is generated as:

$$z_i^{(m)} \sim \mathcal{M}(\boldsymbol{\theta}, 1),$$
$$w_i^{(m)} \sim \mathcal{M}(\boldsymbol{\eta}_{z_i^{(m)}}^{(m)}, 1).$$

2.3.3 Prior distributions

Priors for the covariance matrix of the Gaussian distribution

For a diagonal noise covariance matrix, the inverse variances (precision) τ_d, where $d = 1, \ldots, D$, may be drawn from a gamma distribution $\mathcal{G}(\tau_d|\alpha, \beta)$ with common shape and rate parameters α and β, respectively. The unconstrained noise covariance matrix $\boldsymbol{\Sigma}$ may be drawn from an inverse-Wishart distribution

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\mathbf{S}, \kappa), \tag{2.10}$$

where $\mathbf{S}$ denotes the scale matrix and κ the degrees of freedom for the distribution. When no prior knowledge exists for τ or $\boldsymbol{\Sigma}$, one can simply specify a relatively non-informative prior by assigning α and β or κ to low values, respectively. Note, however, that $\kappa \geq D$ is required to guarantee a valid covariance matrix. Publications I-V use gamma and inverse-Wishart distributions for the covariance matrix of the Gaussian distribution.

Automatic relevance determination

Automatic relevance determination [ARD; Neal, 1996, Tipping, 2001] prior distribution may be used for the elements of a continuous-valued loading matrix. The prior controls model complexity by pushing irrelevant elements close to zero, inducing sparsity. ARD is widely adopted due to its simplicity and effectiveness [Archambeau and Bach, 2008, Ghahramani and Beal, 2000, Fevotte and Godsill, 2006, Fokoue, 2004, Guan and Dy, 2009, Ilin and Raiko, 2010, Li and Tao, 2013, Nakajima et al., 2013, Tan et al., 2009, Tipping and Bishop, 1999a]. Publications I-IV present extensions of ARD, detailed in the next chapter.

Bishop [1999] showed that ARD can be used to determine the number of factors K for probabilistic PCA. In particular, the model can be initialized with a suitably large K. Then during inference some of these factors may

be pruned out from the model. The prior is

$$\mathbf{A}_{:,k} \sim \mathcal{N}(\mathbf{0}, \alpha_k^{-1}\mathbf{I}), \tag{2.11}$$
$$\alpha_k \sim \mathcal{G}(\alpha, \beta),$$

where α_k is the precision for the kth column of $\mathbf{A}$ and the parameters of the gamma distribution are assigned to low values. When the α_k is large (that is, low variance) all the elements in the $\mathbf{A}_{:,k}$ will be close to zero, effectively switching off the kth factor. For an intuitive explanation how ARD induces sparsity, see Tipping [2001].

Hierarchical Dirichlet process

Teh et al. [2006] introduced a hierarchical Dirichlet process (HDP) based topic model for inferring the number of topics based on the observed data. The model is based on the clustering property of the Dirichlet Process, providing a nonparametric prior distribution for the number of topics. Publication VI builds on this work (see Section 3.4).

The HDP is based on a Dirichlet process [DP; Ferguson, 1973] that is briefly introduced in the following. Blackwell and MacQueen [1973] and Sethuraman [1994] define a draw G from a DP as

$$G = \sum_{k=1}^{\infty} p_k \delta_{\eta_k}, \tag{2.12}$$

where the set of values η_k, $k = 1, \ldots, \infty$, are drawn from a base probability distribution G_0, the δ_{η_k} abbreviates an indicator function defined on this set, and the p_k are non-negative and sum to one. The p_k are defined through a stick-breaking process

$$p_k = V_k \prod_{j=1}^{k-1} (1 - V_j),$$

where the V_k are drawn from the beta distribution $V_k \sim \mathcal{B}(1, \alpha)$. A draw from the DP is denoted as $G \sim DP(\alpha, G_0)$.

HDP couples multiple DPs, $(G_1, \ldots, G_N)$. The hierarchical structure ensures that each G_n is defined for the same set of variables η_k given in G. A two-level construction is given as

$$G \sim DP(\alpha, G_0),$$
$$G_n \sim DP(\beta, G),$$

where G is called the top-level DP and G_n the second-level DP with a concentration parameter β.

This paragraph explains a certain construction for a HDP-based topic model which is later referred to in Section 3.4. For this model, the top-level DP G_0 corresponds to $\mathcal{D}(\nu 1)$, the η_k correspond to the topics $\boldsymbol{\eta}_k$ over the vocabulary and the probabilities p_k are defined as in (2.12). Correspondingly, the second-level DP G_n is defined for the nth document and is based on a normalized gamma process [Ferguson, 1973],

$$G_n = \sum_{k=1}^{\infty} \frac{Z_{n,k}}{\sum_{j=1}^{\infty} Z_{n,j}} \delta_{\boldsymbol{\eta}_k}, \tag{2.13}$$
$$Z_{n,k} \sim \mathcal{G}(\beta p_k, 1),$$

where the auxiliary variables $Z_{n,k}$ are drawn from a gamma distribution. Finally, the topic proportion for the nth document is given as $\theta_{n,k} = \frac{Z_{n,k}}{\sum_{j=1}^{\infty} Z_{n,j}}$ and the process for drawing the words is similar to (2.6). Even though the construction is defined for an infinite number of topics, in practice only a finite set is actually used: a finite data collection is explained by a finite number of topics.

3. Models for learning dependencies between multiple data sources

The structure of this chapter follows roughly the contributions of the publications. The description is presented for each method, discussing the concept and related previous work.

3.1 Bayesian canonical correlation analysis via group sparsity

This section examines Bayesian models and inference methods for canonical correlation analysis (CCA; see Section 2.3.2).

Publications I and II present a novel Bayesian model for CCA. The proposed model is

$$\begin{aligned} \mathbf{x}^{(m)} &\sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{y}, \tau_m^{-1}\mathbf{I}), \\ \mathbf{y} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{W}^{(m)}_{:,k} &\sim \mathcal{N}(\mathbf{0}, \alpha_{m,k}^{-1}\mathbf{I}), \\ \alpha_{m,k} &\sim \mathcal{G}(\alpha_0, \beta_0), \\ \tau_m &\sim \mathcal{G}(\alpha_0^\tau, \beta_0^\tau). \end{aligned} \tag{3.1}$$

A key novelty of the model is that it uses a group-wise ARD prior to push unnecessary loadings (columns of $\mathbf{W}^{(m)}$) to zero for each of the views separately. When the loadings for the kth factor $\mathbf{W}^{(m)}_{:,k}$ become non-zero for both views, that factor captures dependencies between the views. Otherwise, when the $\mathbf{W}^{(m)}_{:,k}$ become non-zero only for one view and zero for the other, the factor describes view-specific structure. Finally, the prior still infers the effective number of factors by pushing irrelevant loadings to zero for both views. See Figure 3.1 for demonstration.

Publications I and II use variational Bayes for inference for the model in (3.1), marginalizing over the unknown quantities collected in

$$\boldsymbol{\Theta} = \{\mathbf{W}_{d,:}, \mathbf{z}_n, \alpha_{m,k}, \tau_m\}_{d,n,m,k},$$

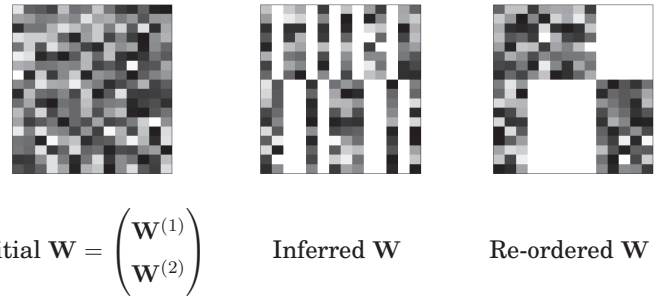

Figure 3.1. Illustration of group-wise sparsity. The developed model (3.1) assumes a group-wise sparse prior distribution for the factor loadings. Here the groups correspond to a partition of variables according to the two views. The prior results in the model converging to a solution that decomposes the variation in data into shared and specific factors. For each factor the loadings corresponding to one view are pushed jointly to zero or all are allowed to be non-zero.

and factored respectively. Open-source implementation of the model, written in the R language, is available in CRAN: `http://cran.r-project.org/package=CCAGFA`.

The developed model (3.1) differs from the existing Bayesian treatments for CCA [Klami and Kaski, 2007, Wang, 2007]. The proposed solution uses uncorrelated diagonal noise covariance matrices, instead Klami and Kaski [2007] and Wang [2007] complemented the CCA model in (2.8) by applying non-informative inverse-Wishart (2.10) priors for the unconstrained noise covariance matrices $\mathbf{\Sigma}^{(m)}$. Even though Wang [2007] provided a variational Bayesian algorithm and Klami and Kaski [2007] derived Gibbs sampling formulas, inferring the $\mathbf{\Sigma}^{(m)}$ becomes very difficult for high-dimensional data, severely limiting practical applications. In particular, Publications I and II demonstrate that the corresponding model becomes computationally inefficient for high-dimensional data and, more importantly, fails to infer the dependencies accurately.

While a low-rank decomposition for the $\mathbf{\Sigma}^{(m)}$, leading to the IBFA model (2.9), solves the problem of high-dimensional covariance estimation, it results in an arduous model selection problem [Archambeau and Bach, 2008]. Hence, this approach has not been shown to work well in real-world applications. Since the model (2.9) comes with three separate sets of factors with factor numbers K, K_1 and K_2, it becomes very difficult to correctly assign variation for each factor. Essentially, the proposed solution (3.1), as shown empirically in Publications I and II, solves this model complexity problem.

In summary, Publications I and II present a novel Bayesian solution

for the CCA/IBFA model that can be used efficiently for real-world applications with large dimensionalities and/or low amount of observations. The solution extracts the dependencies between the views, additionally decomposing the variation in data into shared and specific factors. The novel solution imposes group-wise sparsity to infer the posterior of the Bayesian CCA/IBFA model.

Publication I demonstrates applicability of the model for analyzing neural responses to natural stimulation. Conventional experimental settings and computational methods in neuroscience use block-type stimuli; the experimental setting consists of repeated blocks of the same stimuli and rest. However, such artificial setups provide limited connections to the natural environments our brains usually work in. Being able to study the brain functions in less artificial setups opens up opportunities for understanding the complex functioning of human brains [Malinen et al., 2007]. Publication I uses neural measurements (fMRI) recorded under natural musical stimulation. Given those measurements and a description of the stimuli, the variation shared by the brain activity and the stimulus can be assumed to correspond to stimulus-related activation, while variation only seen in brain activity corresponds to the back-ground processes. Similarly, variation only seen in the stimulus is not being processed by the brain.

3.2 Bayesian group factor analysis

This section presents a novel problem formulation called *group factor analysis* (GFA) for learning dependencies between more than two views and approaches for solving it, following Publications III and IV.

For a multi-view data collection with $M \geq 3$ views and N D_m-dimensional observations $\mathbf{x}_n^{(m)}$, where $n = 1, \ldots, N$ and $m = 1, \ldots, M$, the task of GFA is to find $K < \sum_{m=1}^{M} D_m$ factors that describe the collection and in particular dependencies between the views. The GFA model is

$$\begin{aligned} \mathbf{x}_n^{(m)} &\sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{y}_n, \tau_m^{-1}\mathbf{I}), \\ \mathbf{y}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned}$$

where $\mathbf{x}_n^{(m)}$ is generated as a linear combination of K factors $\mathbf{y}_n$ and the corresponding loadings $\mathbf{W}^{(m)}$ for the mth view. The loadings for all the

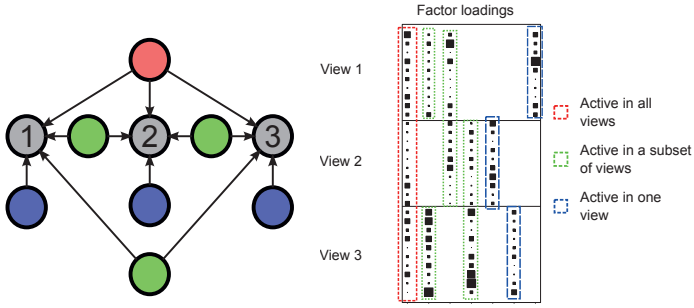

Figure 3.2. Left: An illustrated graphical plate diagram of group factor analysis for three views. The variation in the views, denoted as gray nodes, are divided into various factors (the remaining nodes). **Right**: The corresponding factor loadings $\mathbf{W}$ are group-wise sparse. Thus each factor may be active in any subset of the views. In particular, the factors capture either dependencies between subsets of the views (red nodes are active in all the views and green nodes are active for two views) or explain independent variation or structured noise for one view (blue nodes are active in one view).

factors and views are denoted by

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}^{(1)} \\ \vdots \\ \mathbf{W}^{(M)} \end{pmatrix}.$$

The key novelty of GFA is an assumption of group-wise sparse factors. Some factors capture dependencies between any subset of the views, whereas the remaining factors explain independent variation or structured noise for each view. In particular, the kth factor describes dependencies between a subset of the views if the $\mathbf{W}^{(m)}_{:,k}$ are non-zero only for those views and zero for the others. Furthermore, the factors that are non-zero only for a single view explain non-shared variation for one view. Figure 3.2 illustrates the potential factor loadings $\mathbf{W}$ for three views.

The key in solving the GFA problem is in correctly inferring the sparsity structure. Publications III and IV present group-wise sparse prior distributions for solving the GFA task. Then the factors in the GFA model (loadings $\mathbf{W}$) become group-wise sparse, pushing the unnecessary elements corresponding to some subsets of the views to zero separately for each factor. Publication III generalizes the group-wise ARD prior (3.1) for more than two views. Instead, Publication IV presents a more advanced sparsity prior to better account for inter-view dependencies.

The advanced prior is

$$\begin{aligned} \mathbf{W}_{:,k}^{(m)} &\sim \mathcal{N}(\mathbf{0}, \tilde{\alpha}_{m,k}^{-1}\mathbf{I}), \\ \tilde{\alpha}_{m,k} &= \exp(\mathbf{u}_m^T \mathbf{v}_k + \mu_m^{\mathbf{u}} + \mu_k^{\mathbf{v}}), \\ \mathbf{u}_m, \mathbf{v}_k &\sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}), \\ \mu_m^{\mathbf{u}}, \mu_k^{\mathbf{v}} &\sim \mathcal{N}(0, \lambda_\mu^{-1}). \end{aligned} \tag{3.2}$$

Here $\mathbf{u}_m \in \mathbb{R}^R$ and $\mathbf{v}_k \in \mathbb{R}^R$ are location variables for the mth view and for the kth factor, respectively. In addition, the vectors $\boldsymbol{\mu}^{\mathbf{u}} \in \mathbb{R}^M$ and $\boldsymbol{\mu}^{\mathbf{v}} \in \mathbb{R}^K$ model the mean profiles. The locations induce correlated sparsity between the views. For example, proximity for two views in the location space implies high probability for sharing the same factors. Publication IV demonstrates that the prior (3.2) is especially useful for large number (hundreds) of views.

Publications III and IV use variational Bayes for inference, whereas Publication IV uses empirical Bayes for the location variables. Corresponding implementations in R language are provided in CRAN: `http://cran.r-project.org/package=CCAGFA`.

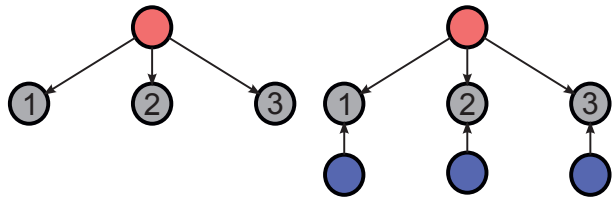

Figure 3.3. Illustrated graphical plate diagrams of the linear Gaussian factor models (see Section 2.3.2) for three views. **Left**: Probabilistic PCA and FA assume a single set of common factors for all views (red nodes) to explain the variation in data (gray nodes). **Right**: MBFA as well as probabilistic CCA and IBFA assume in addition to the common factors another set of factors (or a flexible noise model) for each view to account for view-specific variation (blue nodes).

The linear Gaussian factor models for multi-view data, discussed in Section 2.3.2 and illustrated in Figure 3.3, are not able to solve the GFA problem correctly. In particular, these models fail to capture dependencies between subsets of views, falsely identifying such dependencies either view-specific or shared between all views.

In summary, a novel problem formulation referred to as GFA is proposed for inferring factors that describe dependencies between any subset of views. For solving the GFA problem, group-wise sparse prior distributions are developed.

In essence, GFA is a basic data analysis tool for unsupervised integration of multi-view data. Importantly, the formulation enables addressing

new data analysis problems and designing novel experimental settings. Publication III demonstrates a novel kind of an analysis setup for computational neuroscience where the same subject has been exposed to several variations of the same musical piece. The brain activity measurements (fMRI) recorded under these separate variations are considered as views and the task is to reveal brain activity patterns shared by a subset of the views.

Recently, multiple authors have considered closely related problem formulations to GFA. Gupta et al. [2010] extended Bayesian probabilistic matrix factorization [BPMF; Salakhutdinov and Mnih, 2008] to multiple data matrices with co-occurring observations (views). Their model explicitly includes sets of factors for all possible combinations of views. Since the amount of unique combinations grows exponentially, their approach is not practical for increasing M. In addition, they failed to address the model selection problem. Independently, Gupta et al. [2012] extended a beta process factor analysis model [Paisley and Carin, 2011] building on the hierarchical beta process [HBP; Thibaux and Jordan, 2007] for solving the task similar to GFA. The HBP formulation infers the factors, as well as the number of them using MCMC. Damianou et al. [2012] extended Bayesian Gaussian process latent variable model by Titsias and Lawrence [2010] to multiple views, corresponding to a non-linear formulation of GFA. Even though non-linearity increases modeling flexibility, interpreting the factors may be very difficult. Moreover, non-Bayesian approaches have been proposed to solve related formulations to GFA. They are based on multi-view matrix factorizations using point estimates and structured sparsity inducing regularizers or norms [Bengio et al., 2009, Garrigues and Olshausen, 2010, Jenatton et al., 2010, Jia et al., 2010, Deun et al., 2011]. See Bach et al. [2011, 2012] for further introduction to these approaches. Welling et al. [2008] relates these approaches to the Bayesian approach in the context of latent variable modeling, showing the perils of using point estimates.

3.3 Bayesian exponential family canonical correlation analysis

This section presents a generalization of BCCA that removes the assumption of Gaussian noise, following Publication V. In particular, the noise distribution is generalized to any distribution in the natural exponential family, similarly to how EPCA generalizes probabilistic PCA (Section

2.3.1).

Bayesian exponential family CCA (BECCA) assumes observations drawn from an exponential family distribution,

$$\mathbf{x}_n^{(m)} \sim \mathcal{E}(\boldsymbol{\omega}_n^{(m)}) = h(\mathbf{x}_n^{(m)}) \exp(\mathbf{x}_n^{(m)^T} \boldsymbol{\omega}_n^{(m)} - g(\boldsymbol{\omega}_n^{(m)})).$$

Here, $g(\cdot)$ specifies the distribution for each variable[1], and $\boldsymbol{\omega}_n^{(m)}$ denotes the natural parameters corresponding to the observation $\mathbf{x}_n^{(m)}$. The $\boldsymbol{\omega}_n^{(m)}$ is decomposed as

$$\boldsymbol{\omega}_n^{(m)} = \mathbf{A}^{(m)} \mathbf{z}_n + \mathbf{e}_n^{(m)},$$
$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where the latent variables $\mathbf{z}_n$ capture shared variation between the views and $\mathbf{e}_n^{(m)}$ explains view-specific variation, following the notation and the decomposition idea of probabilistic CCA/IBFA.

Publication V provides a relatively efficient and general solution combining Gibbs and Metropolis-Hastings sampling (Section 2.2.1). In particular, a simple two-level alternating sampling strategy proposed by Hoff [2005, 2007] is adopted. The sampler utilizes Gibbs sampler derived for the fully Gaussian model that is coupled with standard Metropolis-Hastings sampling for providing a generalization to various distributions. If the domain of $\boldsymbol{\omega}$ is constrained, proposals outside this domain are rejected. Due to the difficult model selection problem of IBFA (see Section 3.1), the model parameterizes the view-specific noise as $\mathbf{e}_n^{(m)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(m)})$, following Klami and Kaski [2007] and Wang [2007].

However, this solution does not scale well for high-dimensional data. A potentially more efficient approach to solve BECCA would utilize the group-wise sparsity assumption by sampling the natural parameters for the two-level sampler from the model in (3.1). Consequently, the group-wise sparse prior could reduce the computational load in addition to solving the difficult model selection problem.

Since the exponential family factor models for multi-view data, discussed in Section 2.3.2, assume a single set of factors to account for all variation, they are incapable of solving the task of CCA.

In summary, the proposed model generalizes BCCA to the data domains in the exponential family. The solution utilizes a relatively efficient MCMC schema that could be improved further by incorporating group-wise sparsity assumption. In addition, such a formulation would generalize GFA to

[1]The distribution may vary for the variables. Common $g(\cdot)$ is used in order not to clutter the notation.

the exponential family distributions. The model may be useful in various neuroscientific settings, where the stimulus description contains binary or integer valued data.

Further research should be done to investigate variational Bayes for inference, following Khan et al. [2010], Klami et al. [2013] and Seeger and Bouchard [2012]. In particular, the recent model by Klami et al. [2013] could be modified to solve exponential family GFA.

3.4 Factorized multi-modal topic model

This section introduces a novel multi-view topic model that generates the document counts similarly to how CCA/IBFA generates continuous-valued data, following Publication VI. Given multiple bag-of-words descriptions, the proposed model learns topics that are either shared between the views or specific to each view. The presentation of the model in this introductory part differs from the original publication to better illustrate the non-parametric nature of the model. In particular, the model description provided here relies on the construction of the HDP based topic model in Section 2.3.3.

The model assumes both separate topics and topic proportions for the views, but importantly it captures correspondences between the topics both within as well as between the views and infers the number of topics for each view separately using a HDP formulation. Then the non-zero (active) topics with high correspondence (that is, correlation) across views capture dependencies, whereas the remaining active topics with low correspondence, respectively, capture view-specific variation.

The topics for each view are drawn from the distinct top-level DPs. Each topic is assigned a location variable $\boldsymbol{\ell}_k^{(m)} \in \mathbb{R}^C$ to induce correlations between any two topics either within or across views. For example, two topics that are close to each other in the topic location space tend to co-occur.

The process is written as

$$
\begin{aligned}
G^{(m)} &= \sum_{k=1}^{\infty} p_k^{(m)} \delta_{(\boldsymbol{\eta}_k^{(m)}, \boldsymbol{\ell}_k^{(m)})}, \\
p_k^{(m)} &= V_k^{(m)} \prod_{j=1}^{k-1} (1 - V_j^{(m)}) \\
V_k^{(m)} &\sim \mathcal{B}(1, \alpha^{(m)}), \\
\boldsymbol{\eta}_k^{(m)} &\sim \mathcal{D}(\nu^{(m)} \mathbf{1}), \\
\boldsymbol{\ell}_k^{(m)} &\sim \mathcal{N}(\mathbf{0}, c\mathbf{I}),
\end{aligned}
$$

where c is a non-negative constant. When the view-specific stick-breaking parameter is (close to) zero, that is, $\mathbf{p}_k^{(m)} \approx 0$, the kth topic in the mth is effectively switched off.

The document-level DPs are defined via a weighted normalized gamma process,

$$
\begin{aligned}
G_d^{(m)} &= \sum_{k=1}^{\infty} \frac{Z_{d,k}^{(m)}}{\sum_{j=1}^{\infty} Z_{d,j}^{(m)}} \delta_{(\boldsymbol{\eta}_k^{(m)}, \boldsymbol{\ell}_k^{(m)})}, \\
Z_{d,k}^{(m)} &\sim \mathcal{G}(\beta^{(m)} p_k^{(m)}, \exp(-\mathbf{h}_d^T \boldsymbol{\ell}_k^{(m)})), \\
\mathbf{h}_d &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}).
\end{aligned}
\tag{3.3}
$$

Even though the process appears similar to (2.13), the crucial difference is the second parameter of the gamma distribution. The $\mathbf{h}_d$ is a location variable for the dth document shared between all the views. This variable linearly combines the topic locations, inducing dependency between the topic proportions, $\theta_{d,k}^{(m)} = \frac{Z_{d,k}^{(m)}}{\sum_{j=1}^{\infty} Z_{d,j}^{(m)}}$. The expected value of the $\theta_{d,k}^{(m)}$ is proportional to $\mathbb{E}[Z_{d,k}^{(m)}] = \beta^{(m)} p_k^{(m)} \exp(\mathbf{h}_d^T \boldsymbol{\ell}_k^{(m)})$.

Finally, the ith word $w_{d,i}^{(m)}$ for the dth document in the mth view is drawn as follows

$$
\begin{aligned}
w_{n,i}^{(m)} &\sim \mathcal{M}(\boldsymbol{\eta}_{z_{n,i}^{(m)}}^{(m)}, 1), \\
z_{n,i}^{(m)} &\sim \mathcal{M}(\boldsymbol{\theta}_d^{(m)}, 1)
\end{aligned}
$$

This process may be repeated to draw more words.

Publication VI re-parameterizes the model and uses truncated variational Bayesian approximation for inference. The truncation is equivalent to assigning $V_T^{(m)} = 1$ for a pre-determined truncation number T. The re-parameterization replaces the $\mathbf{h}_d^T \boldsymbol{\ell}_k^{(m)}$ in (3.3) by a variable $\boldsymbol{\xi}_{d,k}^{(m)}$ drawn from a joint Gaussian distribution over the topics and views $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unconstrained covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$. Essentially, each

element of the $\mathbf{\Sigma}$ can be interpreted as a function of the location variables. Further details regarding the posterior inference algorithm can be found in Publication VI and Paisley et al. [2011].

When $M = 1$, the model reduces to the one presented by Paisley et al. [2011] that is similar to the correlated topic model [CTM; Blei and Lafferty, 2007]. When $M \geq 2$, the model is similar to the multi-field CTM [Salomatin et al., 2009]. However, the developed approach uses nonparametric prior distribution for inferring the number of topics based on the observed data.

Multi-modal LDA (see Section 2.3.2) and its non-parametric version by Yakhnenko and Honavar [2009] assume all views to share the same topic proportions for the documents. For this reason, such models fail to capture dependencies between the views in an interpretable way when modeling multi-view data with strong view-specific variation.

In summary, the proposed model combines the modeling principle of CCA/IBFA for learning dependencies with topic modeling. The developed model is able to learn topics that are shared between the views as well as topics specific to one view, using a HDP formulation for learning the number of topics based on the observed data.

In Publication VI, the model is demonstrated on a relatively large collection of web images from Wikipedia pages paired with surrounding text on the page. In particular, both image and text representations may contain strong view-specific variation. By learning dependencies between the visual and textual views, the analysis focuses towards the shared content, isolating aspects that are view-specific.

4. Conclusions

This thesis presents Bayesian latent variable models for multi-view data targeting one important task: learning dependencies between multiple views. Prior studies that have noted the importance of modeling dependencies suffer from inefficient posterior inference algorithms or are limited to constrained settings, hindering real-world applications. This thesis set out with the aim of developing new solutions, removing some of these limitations and advancing the modeling theory for the task.

The developed models of this thesis advance the state of the art for learning dependencies following two principles: decomposition of latent variables and advanced prior distributions. A unifying modeling principle for learning the dependencies decomposes the latent variables into shared and specific. The underlying motivation is that the shared latent variables capture systematic joint variation (statistical dependencies) between the views, while specific latent variables explain remaining non-shared variation. While this approach results in more complicated latent variable models including separate sets of various types of latent variables, the decomposition may be inferred efficiently from the observed data by using group-wise sparse prior distributions and Bayesian inference.

Having obtained an efficient inference solution, this thesis provides many practical data analysis tools for multi-view data that comprise a number of important implications for future practice. A more efficient and accurate method to solve Bayesian CCA (BCCA) is presented, enabling novel applications for high-dimensional data. Such application scenarios were not amenable to address with the existing methods. A novel problem formulation, group factor analysis (GFA), and models for solving it are also presented, for learning dependencies between more than two views, extending BCCA. The new formulation enables massively multi-view set-

tings with tens or hundreds of views. Novel exponential family generalizations of BCCA and GFA can be computed for various data domains, increasing modeling flexibility and scope of research. A novel multi-view topic model is introduced for multi-view document data collections, combining the modeling approach of BCCA/GFA with topic modeling.

The models developed in this thesis serve as a basis for future studies. A Bayesian approach to machine learning, as discussed in this thesis, consists of three stages. These stages involve making model assumptions, inferring unknown quantities (that is, latent variables and parameters) of the model based on observed data and interpreting, evaluating or using the inferred quantities. Future work can be continued in all of these three directions by extending the model structures to better suit particular data, developing and using more efficient posterior inference algorithms and applying the models to various application scenarios in several research fields.

Bibliography

C. Archambeau and F. Bach. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems*, pages 73–80, 2008.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, pages 27–34, 2009.

H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems*, pages 209–215, 2000.

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

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *Advances in Neural Information Processing Systems*, pages 82–89, 2009.

J. M. Bernardo and A. F. M. Smith. *Bayesian theory*. Wiley, New York, 1994.

C. Bishop. Variational principal components. In *Artificial Neural Networks*, pages 509–514, 1999.

C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

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 *International Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of Royal Statistical Society B*, 59(3), 1997.

M. Browne. Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, 33:184–199, 1980.

M. W. Browne. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32:75–86, 1979.

W. L. Buntine. Variational extensions to em and multinomial pca. In *European Conference on Machine Learning*, pages 23–34. Springer-Verlag, 2002.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 2008.

M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624, 2002.

A. Damianou, C. Ek, M. Titsias, and N. Lawrence. Manifold relevance determination. In *International Conference on Machine Learning*, pages 145–152, 2012.

T. de Bie and B. de Moor. On the regularization of canonical correlation analysis. In *International Symposium on Independent Component Analysis and Blind Source Separation*, pages 785–790, 2003.

K. V. Deun, T. F. Wilderjans, R. A. v. Berg, A. Antoniadis, and I. V. Mechelen. A flexible framework for sparse simultaneous component based data integration. *BMC Bioinformatics*, 12(448), 2011.

C. H. Ek, J. Rihan, P. H. Torr, G. Rogez, and N. D. Lawrence. Ambiquity modelling in latent spaces. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 62–73, 2008.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

C. Fevotte and S. J. Godsill. A bayesian approach for blind separation of sparse sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (6):2174–2188, 2006.

E. Fokoue. Stochastic determination of the intrinsic structure in bayesian factor analysis. *Statistical and Applied Mathematical Sciences Institute, Tech. Rep. TR-2004-17*, 2004.

P. Garrigues and B. A. Olshausen. Group sparse coding with a Laplacian scale mixture prior. In *Advances in neural information processing systems*, pages 676–684, 2010.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409, 1990.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data nalysis*. CRC press, 2003.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analyzers. In *Advances in Neural Information Processing Systems*, 2000.

H. R. Glahn. Canonical correlation and its relationship to discriminant analysis and multiple regression. *Journal of the Atmospheric Sciences*, 25(1):23–31, 1968.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004.

Y. Guan and J. G. Dy. Sparse probabilistic principal component analysis. In *Artificial Intelligence and Statistics*, pages 185–192, 2009.

Y. Guo. Supervised exponential family principal component analysis via convex optimization. In *Advances in Neural Information Processing Systems*, pages 569–576, 2008.

S. K. Gupta, D. Phung, B. Adams, T. Tran, and S. Venkatesh. Nonnegative shared subspace learning and its application to social media retrieval. In *International Conference on Knowledge Discovery and Data Mining*, pages 1169–1178, 2010.

S. K. Gupta, D. Phung, and S. Venkatesh. A Bayesian nonparametric joint factor model for learning shared and individual subspaces from multiple data sources. In *SIAM International Conference on Data Mining*, pages 200–211, 2012.

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

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

P. D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295, 2005.

P. D. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478), 2007.

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

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing data. *Journal of Machine Learning Research*, 11:1957–2000, 2010.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Artificial Intelligence and Statistics*, pages 366–373, 2010.

S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification. In *International Conference on Knowledge Discovery and Data Mining*, pages 381–389, 2008.

Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems*, pages 982–990, 2010.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

M. E. Khan, G. Bouchard, K. P. Murphy, and B. M. Marlin. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, pages 1108–1116, 2010.

M. Kim and V. Pavlovic. Covariance operator based dimensionality reduction with extension to semi-supervised settings. In *Artificial Intelligence and Statistics*, pages 280–287, 2009.

A. Klami and S. Kaski. Generative models that discover dependencies between data sets. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 123–128, 2006.

A. Klami and S. Kaski. Local dependent components. In *International Conference on Machine Learning*, pages 425–432, 2007.

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

A. Klami, G. Bouchard, and A. Tripathi. Group-sparse embeddings in collective matrix factorization. *arXiv preprint arXiv:1312.5921*, 2013.

D. Kollar and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

G. Leen. *Context assisted information extraction*. PhD thesis, University of the West of Scotland, 2008.

J. Li and D. Tao. Simple exponential family PCA. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):485–497, 2013.

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Knowledege and Data Discovery Cup and Workshop*, 2007.

E. F. Lock, K. A. Hoadley, J. Marron, and A. B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013.

H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *Conference on Information and Knowledge Management*, pages 931–940, 2008.

D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

S. Malinen, Y. Hlushchuk, and R. Hari. Towards natural stimulation in fMRI–issues of data analysis. *Neuroimage*, 35(1):131–139, 2007.

J. Maritz and T. Lwin. *Empirical Bayes methods*. Number 35 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2 edition, 1989.

R. McDonald. Three common factor models for groups of variables. *Psychometrika*, 37(1):173–178, 1970.

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.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems*, pages 1089–1096, 2009.

I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65 (3):391–411, 2000.

K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *Journal of Machine Learning Research*, 14:1–37, 2013.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning*, pages 777–784, 2011.

J. W. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*, pages 74–82, 2011.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

X. Qu and X. Chen. Sparse structured probabilistic projections for factorized latent spaces. In *International Conference on Information and Knowledge Management*, pages 1389–1394, 2011.

P. Rai and H. Daumé III. Multi-label prediction via sparse infinite CCA. In *Advances in Neural Information Processing Systems*, pages 1518–1526, 2009.

P. Ray, L. Zheng, Y. Wang, J. Lucas, D. Dunson, and L. Carin. Bayesian joint analysis of heterogeneous data. Technical report, Duke University, 2013.

I. Rish, G. Grabarnik, G. Cecchi, F. Pereira, and G. J. Gordon. Closed-form supervised dimensionality reduction with generalized linear models. In *International Conference on Machine learning*, pages 832–839, 2008.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. Citeseer, 2004.

S. Roweis. EM-algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *International Conference on Machine Learning*, pages 880–887, 2008.

K. Salomatin, Y. Yang, and A. Lad. Multi-field correlated topic modeling. In *SIAM International Conference on Data Mining*, pages 628–637, 2009.

M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *Artificial Intelligence and Statistics*, pages 701–708, 2010.

M. Seeger and G. Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Artificial Intelligence and Statistics*, pages 1012–1018, 2012.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *International Conference on Knowledge Discovery and Data Mining*, pages 650–658, 2008.

J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, pages 1470–1477, 2003.

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision*, pages 370–377, 2005.

C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extension, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):194–200, 2011.

V. Y. Tan, C. Févotte, et al. Automatic relevance determination in nonnegative matrix factorization. In *Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Advances in neural information processing systems*, pages 1481–1488, 2007.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571, 2007.

M. E. Tipping. Probabilistic visualisation of high-dimensional binary data. In *Advances in Neural Information Processing Systems*, pages 592–598, 1999.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999a.

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

M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In *Artificial Intelligence and Statistics*, pages 844–851, 2010.

L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23: 111–136, 1958.

C. Wang. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

F. V. Waugh. Regressions between sets of variables. *Journal of the Econometric Society*, pages 290–310, 1942.

M. Wedel and W. A. Kamakura. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530, 2001.

M. Welling, C. Chemudugunta, and N. Sutter. Deterministic latent variable models and their pitfalls. In *SIAM International Conference on Data Mining*, pages 196–207, 2008.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7(2003):723–732, 2003.

W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Uncertainty in Artificial Intelligence*, pages 626–633, 2000.

E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, pages 583–591, 2003.

O. Yakhnenko and V. Honavar. Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence. In *SIAM International Conference on Data Mining*, 2009.

K. Yu and V. Tresp. Heterogenous data fusion via a probabilistic latent variable model. *Lecture Notes in Computer Science*, pages 20–30, 2004.

S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *International Conference on Knowledge Discovery and Data Mining*, pages 464–473, 2006.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD70/2013 Ylipaavalniemi, Jarkko
Data-driven Analysis for Natural Studies in Functional Brain Imaging. 2013.

Aalto-DD61/2013 Kandemir, Melih
Learning Mental States from Biosignals. 2013.

Aalto-DD90/2013 Yu, Qi
Machine Learning for Corporate Bankruptcy Prediction. 2013.

Aalto-DD128/2013 Ajanki, Antti
Inference of relevance for proactive information retrieval. 2013.

Aalto-DD205/2013 Lijffijt, Jefrey
Computational methods for comparison and exploration of event sequences. 2013.

Aalto-DD21/2014 Cho, Kyunghyun
Foundations and Advances in Deep Learning. 2014.

Aalto-DD49/2014 Lindh-Knuutila, Tiina
Computational Modeling and Simulation of Language and Meaning: Similarity-Based Approaches. 2014.

Aalto-DD80/2014 Toivola, Janne
Advances in Wireless Damage Detection for Structural Health Monitoring. 2014.

Aalto-DD105/2014 Parkkinen, Juuso
Probabilistic components of molecular interactions and drug responses. 2014.

Aalto-DD108/2014 Faisal, Ali
Retrieval of Gene Expression Measurements with Probabilistic Models. 2014.

Aalto-DD 110/2014

ISBN 978-952-60-5784-2
ISBN 978-952-60-5785-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS